A modelling framework for estimation of comparative effectiveness in pharmaceuticals using uncontrolled clinical trials

Anthony James Hatswell

UCL

Statistical Science

I, Anthony James Hatswell confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Anthony James Hatswell Date

ABSTRACT

Pharmaceuticals are most commonly studied in randomised controlled trials (RCTs) against a control arm (either active, or placebo). On occasions however treatments are licensed exclusively on the basis of uncontrolled study data - this thesis investigates how comparative effectiveness can be estimated under such circumstances.

The role of RCTs in the approval and estimation of comparative effectiveness in pharmaceuticals is discussed, as well as potential methods for analysis where RCT data are not available. A review of all drug approvals from 1999-2014 by the European Medicines Agency and the US Food and Drug Administration is then presented. Performing literature searches in the majority of cases (80%), historical controls have been the primary source of estimates of comparative effectiveness, frequently without attempts to adjust for differences between studies.

Given the high usage of historical controls, I focussed on the role of adjustment. This included a simulation study to understand where the method of Matching Adjusted Indirect Comparison (MAIC) is likely to be of use, looking specifically at the effect of model misspecification. Three novel methods (with practical examples) for the creation of historical controls are then presented; using extrapolation from the previous line of therapy, using non-responders to therapy as a surrogate, and comparing to a patient's own prior data.

The conclusion of the work is that there are clearly situations where RCTs cannot, or will not be used – regardless of the statistical issues this raises. In such cases by proactively identifying appropriate historical data, and using appropriate analysis methods – the downsides can be ameliorated, at least in part. A flowchart presenting the available methods (split by data access) is presented.

Further research is required on the appropriateness of different sources of historical control data (e.g. registries versus RCT arms), and how to synthesize multiple estimates of effectiveness (e.g. multiple MAICs).

TABLE OF CONTENTS

ABSTRACT **3**

TABLE OF CONTENTS **4**

TABLE OF TABLES **7**

TABLE OF FIGURES **7**

TABLE OF EQUATIONS **9**

ACKNOWLEDGEMENTS **10**

PUBLICATIONS **11**

IMPACT STATEMENT **14**

ABBREVIATIONS **15**

1 INTRODUCTION **17**

1.1 THE HISTORY AND THEORETICAL UNDERPINNING OF RANDOMISED TRIALS 19
1.2 'CONVENTIONAL' RANDOMISED DOUBLE BLIND TRIALS 21
1.2.1 Randomisation and exchangeability *21*
1.2.2 Blinding of patients and physicians *22*
1.2.3 Multi-centre trials *23*
1.3 MATHEMATICAL NOTATION FOR RANDOMISED CLINICAL TRIALS 23
1.4 PREVIOUS WORK IN ASSESSING EVIDENCE WITHOUT RANDOMISED TRIALS 26
1.4.1 Assessing the comparative effectiveness of treatments studied without a randomised control *26*
1.4.2 Effect sizes seen in observational data, compared to RCTs *27*
1.4.3 Historical controls *29*
1.4.4 Methodologies for the use of observational data in estimation of efficacy *32*
1.5 THE ROLE OF UNCONTROLLED STUDIES IN DRUG APPROVALS 33
1.5.1 'Obviousness' *34*
1.5.2 Clinical equipoise *35*
1.5.3 The benefit-risk of trial participants and patients *36*
1.6 THE USE OF UNCONTROLLED STUDIES IN MODELLING AND HEALTH TECHNOLOGY APPRAISAL – GUIDANCE FROM AGENCIES 37
1.7 SUMMARY OF INTRODUCTION AND RESEARCH QUESTION 41

2 EXISTING METHODOLOGIES THAT COULD BE USED TO ESTIMATE EFFECTIVNESS FROM UNCONTROLLED STUDIES **44**

2.1 METHODOLOGIES FOR THE ANALYSIS OF HISTORICAL CONTROLS 44
2.1.1 Methodologies for use with published historical controls *44*
2.1.2 Where individual level data (ILD) are available for either the intervention or the historical data *51*
2.1.3 Where individual level data (ILD) are not available to a researcher for the intervention or the historical data *55*
2.2 METHODOLOGIES FOR THE SYNTHESIS OF MULTIPLE HISTORICAL CONTROLS 55
2.2.1 Meta-regression *55*
2.2.2 Meta-analysis of historical controls *56*
2.2.3 The Bayesian 'power prior' *57*
2.2.4 Comparison between approaches for combing historical controls, and applicability to uncontrolled studies *59*
2.3 WHERE NO HISTORICAL DATA ARE AVAILABLE 59

2.3.1 The use of expert opinion *59*
2.3.2 Threshold analysis and the 'E-value' *60*
2.4 EMERGING METHODOLOGIES 61
2.4.1 The use of 'real world data' to establish control arms *61*
2.4.2 Machine learning *62*
2.5 SUMMARY OF EXISTING METHODOLOGIES FOR ESTIMATING COMPARATIVE EFFECTIVENESS USING UNCONTROLLED STUDIES 63

3 IDENTIFICATION OF THE NUMBER OF TREATMENTS INVOLVED, AND METHODS USED FOR MODELLING 64

3.1 TREATMENTS APPROVED ON THE BASIS OF UNCONTROLLED CLINICAL STUDIES 64
3.1.1 Regulatory processes in the United States and the European Union *64*
3.1.2 Details of the search of the EMA and FDA drug approval databases *65*
3.1.3 Consolidated list of treatments licensed on the basis of uncontrolled studies from 1999 to 2014 ..... *70*
3.1.4 Disease areas where uncontrolled studies have most frequently been the basis for drug approvals *73*
3.1.5 Comparison between the FDA and EMA on the number of approvals, and the dates of reviews *73*
3.1.6 Congruence of findings with the existing literature *78*
3.1.7 Subsequent work performed by others in the area *80*
3.2 METHODS USED FOR ESTIMATING EFFECTIVENESS FROM UNCONTROLLED STUDIES 81
3.2.1 Literature search for modelled estimates of the efficacy of drugs licensed on the basis of uncontrolled clinical studies *82*
3.2.2 Description and de-duplication of published estimates *90*
3.2.3 A taxonomy of modelling approaches *92*
3.3 SUMMARY OF FINDINGS FROM LITERATURE SEARCHES 94

4 SIMULATION STUDY REGARDING THE PERFORMANCE OF UNANCHORED MATCHING ADJUSTED INDIRECT COMPARISON (MAIC) 95

4.1 APPROACH AND DATA GENERATION 96
4.2 APPLICATION OF MATCHING ADJUSTED INDIRECT COMPARISON 98
4.3 OUTCOMES OF THE STUDY 99
4.4 SCENARIO ANALYSES PERFORMED 100
4.5 IMPLEMENTATION IN SOFTWARE AND MODEL CONVERGENCE 104
4.6 FINDINGS FROM THE BASE CASE 106
4.7 FINDINGS FROM SCENARIO ANALYSES 108
4.8 FINDINGS FROM VARYING THE NUMBER OF PATIENTS AVAILABLE IN EACH DATASET 116
4.9 DISCUSSION ON THE MERITS AND APPROACH TO MAIC 119
4.10 THE USE OF MATCHING ADJUSTED INDIRECT COMPARISON OUTSIDE OF TIME TO EVENT OUTCOMES 119
4.11 SUMMARY OF FINDINGS 120

5 NOVEL METHODS FOR THE CREATION OF HISTORICAL CONTROLS 121

5.1 THE USE OF NON-RESPONDERS AS A CONTROL ARM 121
5.2 THE USE OF EXTRAPOLATION TO CREATE A HISTORICAL CONTROL 123
5.2.1 Extrapolation of data from the previous line of treatment to estimate the counterfactual *123*
5.2.2 Using a patient's outcomes from a prior line of treatment to estimate counterfactual outcomes *128*
5.3 SUMMARY OF NOVEL METHODS PROPOSED 130

6 PRACTICAL EXAMPLES OF THE IMPLEMENTATION OF METHODS 132

6.1 AVELUMAB FOR THE TREATMENT OF MERKEL CELL CARCINOMA 132
6.2 BRENTUXIMAB VEDOTIN FOR THE TREATMENT OF HODGKIN'S LYMPHOMA 133

7 DISCUSSION 136

7.1 THE USE OF UNCONTROLLED CLINICAL STUDIES IN PHARMACEUTICAL LICENSING 136
7.2 TECHNIQUES FOR ESTIMATING EFFECTIVENESS BASED ON UNCONTROLLED CLINICAL STUDIES 137
7.3 IMPLICATIONS FOR THE DESIGN AND CONDUCT OF UNCONTROLLED STUDIES 138
7.4 POTENTIAL IMPROVEMENTS AND MODIFICATIONS TO TRIAL DESIGNS 139
7.5 SUGGESTED DECISION PROCESS FOR METHOD SELECTION 140
7.6 MY CONTRIBUTION TO THE LITERATURE 143
7.7 LIMITATIONS OF THE WORK PERFORMED 143
7.8 FURTHER RESEARCH 144

REFERENCES 147

APPENDIX A EXCLUSION CRITERIA FOR LITERATURE SEARCH FOR DRUGS APPROVED USING UNCONTROLLED CLINICAL TRIALS 163

APPENDIX B LITERATURE SEARCH FOR ECONOMIC EVALUATIONS OF DRUGS LICENSED ON THE BASIS OF UNCONTROLLED CLINICAL TRIALS 165

B1 MEDLINE (SEARCHED USING PUBMED) 165
B2 INTERNATIONAL SOCIETY FOR PHARMACOECONOMICS AND OUTCOMES RESEARCH (ISPOR) SCIENTIFIC PRESENTATIONS DATABASE 169
B3 NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE), SCOTTISH MEDICINES CONSORTIUM (SMC) AND ALL WALES MEDICINES STRATEGY GROUP (AWMSG) HEALTH TECHNOLOGY APPRAISALS 172

APPENDIX C CODE USED FOR THE IMPLEMENTATION OF MAIC IN THE SIMULATION STUDY 175

APPENDIX D EXAMPLE R CODE TO IMPLEMENT METHODS CONCEPTUALISED IN SECTION 5.2 179

TABLE OF TABLES

TABLE 1-1: RELATIONSHIP OF MODELLING GUIDELINES FOR ECONOMIC EVALUATION TO TREATMENTS WITH ONLY UNCONTROLLED STUDY DATA AVAILABLE 40
TABLE 3-1: NUMBER OF DRUGS AND INDICATIONS APPROVED IN THE EU AND US ON THE BASIS OF UNCONTROLLED CLINICAL STUDY DATA 70
TABLE 3-2: DRUGS SUBMITTED TO THE EMA AND FDA CONTAINING ONLY UNCONTROLLED CLINICAL STUDIES 71
TABLE 3-3: NUMBER AND SOURCE OF MODELLED ESTIMATES OF EFFICACY IDENTIFIED AS BEING BASED ON UNCONTROLLED CLINICAL STUDY DATA, WITH REASONS FOR EXCLUSION SHOWN IN ITALICS 90
TABLE 4-1: BASE CASE AND SENSITIVITY ANALYSIS PARAMETERS FOR THE SIMULATION STUDY OF THE PERFORMANCE OF MAIC 102
TABLE 4-2: TABULATED RESULTS OF MAIC SIMULATION STUDY BASE CASE, 5,000 RUNS 107
TABLE 4-3: TABULATED RESULTS OF MAIC SCENARIO ANALYSIS VARYING THE SETUP OF THE SIMULATION STUDY, 5,000 RUNS 109
TABLE 4-4: TABULATED RESULTS OF MAIC SCENARIO ANALYSIS VARYING THE CONDITIONS OF THE STUDY, 5,000 RUNS 113
TABLE 4-5: TABULATED RESULTS OF MAIC SCENARIO ANALYSIS WITH SCENARIOS THAT VIOLATE ASSUMPTIONS IMPLICIT OR EXPLICIT IN MAIC, 5,000 RUNS 115
TABLE 4-6: TABULATED RESULTS OF MAIC SCENARIO ANALYSIS VARYING THE NUMBER OF PATIENTS AVAILABLE IN POPULATION A AND POPULATION B, 5,000 RUNS 117
TABLE 6-1: ESTIMATED LIFE YEARS IN EACH HEALTH STATE FOR BRENTUXIMAB VEDOTIN AND SINGLE AGENT CHEMOTHERAPY IN HODGKIN'S LYMPHOMA 135
TABLE B-1: RESULTS OF PUBMED SEARCHES FOR ECONOMIC EVALUATIONS OF DRUGS LICENSED ON THE BASIS OF UNCONTROLLED STUDY DATA 166
TABLE B-2: RESULTS OF SEARCHES FOR ECONOMIC EVALUATIONS OF DRUGS LICENSED ON THE BASIS OF UNCONTROLLED STUDY DATA IN THE ISPOR SCIENTIFIC PRESENTATIONS DATABASE 169
TABLE B-3: RESULTS OF SEARCHES FOR HEALTH TECHNOLOGY ASSESSMENTS OF DRUGS LICENSED ON THE BASIS OF UNCONTROLLED STUDY DATA 172

TABLE OF FIGURES

FIGURE 1-1: A COMPARISON OF RANDOMISED CLINICAL TRIALS AND DIFFERENT FORMS OF UNCONTROLLED CLINICAL STUDIES 17
FIGURE 1-2: HIERARCHY OF EVIDENCE, DIRECTLY REPRODUCED FROM ANNEX B OF 'SIGN 50: A GUIDELINE DEVELOPER'S HANDBOOK' (SIGN, 2011) 18
FIGURE 3-1: DRUGS@FDA DATABASE STRUCTURE, TAKEN FROM HTTP://WWW.FDA.GOV/DRUGS/INFORMATIONONDRUGS/UCM079750.HTM ON 8 MAY 2014 .. 66
FIGURE 3-2: PRISMA DIAGRAM OF DRUG INDICATIONS APPROVED ON THE BASIS OF SINGLE ARM TRIALS BY THE FOOD AND DRUG ADMINISTRATION 67
FIGURE 3-3: PRISMA DIAGRAM OF DRUG INDICATIONS APPROVED ON THE BASIS OF SINCE ARM TRIALS BY THE EUROPEAN MEDICINES AGENCY 68

FIGURE 3-4: TIMELINE FROM SUBMISSION TO APPROVAL OF PHARMACEUTICALS LICENSED ON THE BASIS OF UNCONTROLLED STUDY DATA BY BOTH THE FDA AND EMA 77
FIGURE 3-5: PRISMA DIAGRAM OF MODELLED COMPARISONS RETRIEVED FROM PUBMED 83
FIGURE 3-6: PRISMA DIAGRAM OF MODELLED ESTIMATES IDENTIFIED IN THE ISPOR SCIENTIFIC PRESENTATIONS DATABASE 85
FIGURE 3-7: PRISMA DIAGRAM OF NICE APPRAISALS INVOLVING ECONOMIC MODELS 87
FIGURE 3-8: PRISMA DIAGRAM OF SCOTTISH MEDICINES CONSORTIUM SUBMISSIONS 88
FIGURE 3-9: PRISMA DIAGRAM OF ALL WALES MEDICINES STRATEGY GROUP SUBMISSIONS 89
FIGURE 3-10: TAXONOMY OF ECONOMIC MODELLING APPROACHES USED FOR ESTIMATING INCREMENTAL BENEFIT FROM UNCONTROLLED CLINICAL STUDIES 93
FIGURE 4-1: CONVERGENCE PLOT OF THE BASE CASE MAIC SIMULATION STUDY; UP TO 100,000 SIMULATIONS FOR MEAN PERCENTAGE ERROR OF NAÏVE COMPARISON $MAIC_{FM}$, $MAIC_{HM}$ AND PSW WITH VERTICAL LINES AT 1,000 AND 5,000 SIMULATIONS 105
FIGURE 4-2: CONVERGENCE PLOT OF 30 RUNS OF 1000 SIMULATIONS FOR THE BASE CASE MEAN ABSOLUTE PERCENTAGE ERROR OF $MAIC_{FM}$ IN THE SCENARIO WITH N=30 IN POPULATION A AND N=30 IN POPULATION B 106
FIGURE 4-3: VIOLIN PLOT (WITH OVERLAID BOX PLOTS) OF THE PERCENT MEAN ERROR IN 5,000 RUNS OF THE MAIC SIMULATION STUDY BASE CASE 108
FIGURE 4-4: VIOLIN PLOT (WITH OVERLAID BOX PLOTS) OF THE PERCENT MEAN ERROR IN 5,000 RUNS OF THE MAIC SIMULATION STUDY BASE CASE 111
FIGURE 4-5: VIOLIN PLOT (WITH OVERLAID BOX PLOTS) OF THE PERCENT MEAN ERROR IN 5,000 RUNS OF THE MAIC VARYING THE CONDITIONS OF THE STUDY 114
FIGURE 4-6: VIOLIN PLOT (WITH OVERLAID BOX PLOTS) OF THE PERCENT MEAN ERROR IN 5,000 RUNS OF THE MAIC SIMULATION STUDY WITH SCENARIOS THAT VIOLATE ASSUMPTIONS IMPLICIT OR EXPLICIT IN MAIC 116
FIGURE 4-7: VIOLIN PLOT (WITH OVERLAID BOX PLOTS) OF THE PERCENT MEAN ERROR IN 5,000 RUNS OF THE MAIC SIMULATION STUDY VARYING THE NUMBER OF PATIENTS INCLUDED IN POPULATION A AND POPULATION B 118
FIGURE 5-1: ESTIMATED OVERALL SURVIVAL IN ALL PATIENTS AND NON-RESPONDERS FOR OFATUMUMAB IN DOUBLE REFRACTORY CHRONIC LYMPHOCYTIC LEUKAEMIA 122
FIGURE 5-2: PROGRESSION FREE SURVIVAL AND OVERALL SURVIVAL FROM THE CAM211 STUDY IN REFRACTORY CHRONIC LYMPHOCYTIC LEUKAEMIA EXTRACTED FROM THE FDA REVIEW (FOOD AND DRUG ADMINISTRATION, 2001) 124
FIGURE 5-3: RECREATED DISEASE-FREE SURVIVAL AND OVERALL SURVIVAL FROM THE CAM211 STUDY, WITH FITTED (LOGNORMAL) PARAMETRIC CURVE FITS OVERLAID 125
FIGURE 5-4: DIGITIZED OFATUMUMAB KAPLAN-MEIER SURVIVAL DATA FROM HX-CD20-406 AND WEIBULL CURVE FIT, PLOTTED AGAINST SURVIVAL POST DISEASE PROGRESSION FROM THE CAM211 STUDY 126
FIGURE 5-5: DENSITY PLOT OF ESTIMATED MEAN SURVIVAL GAIN OF OFATUMUMAB OVER HISTORICAL CONTROL, USING ASSUMPTIONS OF IMMEDIATE TREATMENT WITH OFATUMUMAB, 28-DAY DELAY, AND 3.6-MONTH DELAY 127
FIGURE 5-6: DIGITIZED KAPLAN-MEIER DATA FROM IDELALISIB STUDY 101-09, INCLUDING FITTED (WEIBULL) PARAMETRIC SURVIVAL CURVES 129
FIGURE 5-7: DENSITY PLOT OF ESTIMATED TIME TO PROGRESSION GAIN OF IDELALISIB OVER THE PREVIOUS LINE OF TREATMENT 130

FIGURE 6-1: MARKOV TRACE OF MODELLED PATIENT SURVIVAL WITH BRENTUXIMAB VEDOTIN IN HODGKIN'S LYMPHOMA 134
FIGURE 6-2: MARKOV TRACE OF MODELLED PATIENT SURVIVAL WITH SINGLE AGENT CHEMOTHERAPY IN HODGKIN'S LYMPHOMA 135
FIGURE 7-1: SUGGESTED ALGORITHM FOR ESTIMATION OF EFFECTIVENESS BASED ON UNCONTROLLED CLINICAL STUDIES 142
FIGURE B-1: PUBMED SEARCH STRATEGY FOR ECONOMIC MODELS OF DRUGS BASED ON UNCONTROLLED STUDY DATA 165

TABLE OF EQUATIONS

EQUATION 1: POPULATION AVERAGE OUTCOMES FROM TREATMENT $\boldsymbol{t}$ 23
EQUATION 2: POPULATION AVERAGE OUTCOMES FROM CONTROL $\boldsymbol{c}$ 24
EQUATION 3: DIFFERENCE IN RESPONSE OF A TREATMENT, GIVEN MATCHING BETWEEN STUDIES 49
EQUATION 4: DIFFERENCE IN RESPONSE OF A TREATMENT, GIVEN WEIGHTING BETWEEN STUDIES 49
EQUATION 5: EXAMPLE OF LINEAR REGRESSION 51
EQUATION 6: CALCULATION OF WEIGHTS FOR MATCHING ADJUSTED INDIRECT COMPARISON 53
EQUATION 7: REWEIGHTING OF OUTCOMES TO MATCH AND INDEX STUDY AFTER CALCULATION OF WEIGHTS FROM MATCHING ADJUSTED INDIRECT COMPARISON 53

ACKNOWLEDGEMENTS

This thesis is the work of a long period of time; around 6 years formally enrolled as a ‘student’, but really the culmination of my career - it simply would not exist without input, support, and encouragement from a vast number of people. I’ve tried to list some of the main influences, but to list all would be an unmanageable task – I owe a great deal to a great many people. I can never hope to repay them for the opportunities I’ve been given, so have tried (and will continue) to instead give opportunities to those who come after me.

During the process of the PhD I have learnt phenomenal amounts from my supervisors (Professor Gianluca Baio and Professor Nick Freemantle) – far beyond the work presented in this thesis. Their input and guidance has been exemplary, giving me perfectly timed direction and nudges. More importantly they’ve had a huge impact on the way I approach my work and live my life. Neither I, nor this thesis, would have been the same without them.

The support of people in my field has been extraordinary in the truest sense, from my previous employers (Ron Akehurst and Nic Brereton) to start, to others who provided encouragement and discussions along the way. I would highlight Ash Bullement, Dan Gladwell, Zoe Philips, Simon McNamara, Nick Latimer, James Signorovitch, Will Sullivan, Diarmuid Coughlan, Rob Smith, and Chris Sampson as going above and beyond.

Outside of my studies and work, the patience, support and belief of my wife Laura has been amazing – we took this on as a family, and literally grew as a family in the process. Here I have to acknowledge the ‘input’ of my children Alistair (5) and Lillian (3) who have been amazing (and patient) so often, with baby Abigail giving the final push to get the work completed. I’m sure the thesis would have been completed faster without them, but it wouldn’t have been as rewarding, or I as motivated to complete it.

I must also thank my friends and family who have displayed incredible understanding when I’ve needed to skip events / meetups / runs / fun to focus on analysis, writing, coding, or some combination of the above. Despite my occasional grumpiness and absence, they have been brilliant in offering moral support and distractions when I’ve needed them. I didn’t take this project on lightly, but hope the end result justifies the choices made.

The final thankyou would be to those who have gone before me. I’ve read thousands of articles (and cited hundreds) from the greats in the field (Sargent, Byar, Meier, Pocock), the diligent work of regulators at the EMA and FDA, right through to early career researchers communicating their work in papers. Without their work, mine would not have been possible.

PUBLICATIONS

The work that constitutes this thesis has been communicated in multiple articles published in peer reviewed journals and two UCL research reports. Where appropriate these publications have been highlighted at the beginning of relevant sections of the thesis. The publications are listed below in to order in which they appear in this thesis, alongside my role in each of the papers:

1. Hatswell, A.J., Bardou, M., Gallagher, M. & Beckerman, R. (2014) Modeling Alchemy: The Impact of Unorthodox Trial Design on Health Technology Appraisal Strategy. *ISPOR Connections*. 20 (4), 6–9.

   Having (collectively) presented an ISPOR Workshop discussing some of the issues around uncontrolled studies and other similarly complex trial designs, such as crossover studies, we [the authors] communicated the issues as an industry magazine article

2. Hatswell, A.J., Baio, G., Berlin, J.A., Irs, A., et al. (2016) Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. *BMJ Open*. [Online] 6 (6), e011666. Available from: doi:10.1136/bmjopen-2016-011666.

   Here I performed the reviews of drug approvals, performed further evidence searching, and derived the narrative presented

3. Hatswell, A.J., Baio, G. & Freemantle, N. (2017) *Research Report number 327: A description of the circumstances surrounding pharmaceutical approvals by the FDA and EMA from 1999 to 2014 made without randomised control trial data*.

   This work explains the full context behind each of the drug approvals listed in paper 2. Having performed the review, I wrote up the findings for each of the treatments

4. Hatswell, A.J., Freemantle, N. & Baio, G. (2017a) Economic Evaluations of Pharmaceuticals Granted a Marketing Authorisation Without the Results of

Randomised Trials: A Systematic Review and Taxonomy. *PharmacoEconomics*. [Online] 35 (2), 163–176. Available from: doi:10.1007/s40273-016-0460-6.

I performed the systematic reviews to identify the models, extracted the data, summarised the results and crafted the narrative of the paper

5. Hatswell, A.J., Freemantle, N. & Baio, G. (2017b) *Research Report number 326: A description of economic models constructed for pharmaceuticals granted a marketing authorisation without a randomised controlled trial by the FDA and EMA from 1999 to 2014*.

   This work explains the details of the approaches in paper 4. Having performed the literature reviews, I summarised the evidence, and deduplication process (as some models were presented in multiple formats)

6. Hatswell, A., Freemantle, N. & Baio, G. (2020) The effects of model misspecification in unanchored Matching Adjusted Indirect Comparison (MAIC); Results of a simulation study. *Value in Health*. Accepted for publication doi:10.1016/j.jval.2020.02.008.

   For this project I conceptualised the study, coded the simulation, designed the scenarios to test, and analysed the results

7. Bullement, A., Nathan, P., Willis, A., Amin, A., et al. (2019) Cost Effectiveness of Avelumab for Metastatic Merkel Cell Carcinoma. *PharmacoEconomics - Open*. [Online] Available from: doi:10.1007/s41669-018-0115-y [Accessed: 25 January 2019].

   My role in this project was to design (and implement) the approach to the estimation of comparative effectiveness, as well analyse quality of life data. The results of my work were then used in an economic model constructed by Ash Bullement & Anna Willis

8. Hatswell, A.J., Thompson, G.J., Maroudas, P.A., Sofrygin, O., et al. (2017) Estimating outcomes and cost effectiveness using a single-arm clinical trial: ofatumumab for double-refractory chronic lymphocytic leukemia. Cost Effectiveness

and Resource Allocation. [Online] 15, 8. Available from: doi:10.1186/s12962-017-0071-x.

For this paper I conceptualised the approach, which was implemented as a collective effort by all co-authors

9. Hatswell, A.J. & Sullivan, W.G. (2019) Creating historical controls using data from a previous line of treatment – Two non-standard approaches. *Statistical Methods in Medical Research*. [Online] 0962280219826609. Available from: doi:10.1177/0962280219826609.

   I conceptualised the two methods, and performed the statistical coding to implement them. One method (use of a patient's prior therapy data) had input from my co-author, who built an economic model using the approach for a NICE submission

IMPACT STATEMENT

The experience I have gained in the research process has had an impact in many domains of my working life. Firstly the work performed has been published, cited by others, and used in health technology assessment (HTA) submissions. I have then disseminated the work beyond publications having presented to academic groups, consultancies, the National Institute for Health and Care Excellence (NICE), and pharmaceutical companies.

In the world of work, the skills I have demonstrated have increased my credibility, allowing me to be involved in areas where I would not otherwise have had access. As a result in 2017 I left my senior role at a company of 150 people to set up my own company. *Delta Hat* has six full time employees in our office near Nottingham and is named after my research interest (the estimation of comparative effectiveness). My visibility and reputation have meant I have been invited to serve on the editorial board of a journal (*Pharmacoeconomics Open*), advise on methodology for companies & institutions, sit on a NICE committee, and mentor students. Without starting the PhD, my life would look quite different.

Beyond the work contained in this thesis, the knowledge and skills I have gained have enabled me to have influence in areas beyond my niche. This includes journal articles not included in the thesis such as an editorial on the impact of 'Brexit' (Hatswell, 2017) which was picked up by the press (Johnston, 2017; Clark, 2017). The statistical knowledge (particularly Bayesian statistics) led to work on synthesis of utility values for health technology assessment (Hatswell *et al.*, 2019) which has been widely used in decision making. Finally learning to code in the statistical software R has let me achieve things I simply could not have done otherwise – examples include research on reducing wastage in pharmaceuticals by rationally selecting vial sizes (Hatswell & Porter, 2018), and assessing the convergence of models in probabilistic sensitivity analysis (Hatswell *et al.*, 2018). I now help organise and lecture on an annual 'R in HTA' workshop.

As a return on investment, this would have been enough for me to be satisfied, however I think the main impact has been on the quality of my output as a whole. Whilst much of this thesis consists of (published) academic work, the reality is most of my work is consultancy which is proprietary and ephemeral. Little trace (save the decisions influenced) remains after even a few months, and most goes unattributed; HTA submissions do not list authors. The skills I have gained at UCL have had a clear impact on what I deliver, and how I assess / guide the work others produce – in this thesis three (public domain) examples are given. The main impact will therefore be better analyses for decision making – although unlikely to be directly attributable, it is no less real.

ABBREVIATIONS

AHRQ	Agency for Healthcare Research and Quality
AIDS	Acquired Immune Deficiency Syndrome
ALD	Aggregate Level Data
ALL	Acute lymphoblastic leukaemia
AML	Acute myeloid leukaemia
APL	Acute promyelocytic leukaemia
AWMSG	All Wales Medicines Strategy Group
CAR T	Chimeric Antigen Receptor T-cell
CHMP	Committee for Medicinal Products for Human Use
CLL	Chronic Lymphocytic Leukaemia
CML	Chronic Myeloid Leukaemia
CPH	Cox Proportional Hazards
CTCL	Cutaneous T-cell lymphoma
EC	European Commission
ECOG	Eastern Cooperative Oncology Group
EMA	European Medicines Agency
EMR	Electronic Medical Records
EPAR	European Public Assessment Report
EPO	Erythropoietin
ERT	Enzyme Replacement Therapy
EU	European Union
FDA	Food and Drug Administration
FM	First Moments
GIST	Gastrointestinal stromal tumours
GLM	Generalised Linear Model
HIT	Heparin-induced thrombocytopenia
HIV	Human Immunodeficiency Virus
HL	Hodgkin's Lymphoma
HM	Higher Moments
HPCT	Haematopoietic progenitor cell transplantation
HTA	Health Technology Assessment
ICH	International Conference on Harmonisation
ILD	Individual Level Data
IPTW	Inverse Probability of Treatment Weighting
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
IV	Intravenous
KS	Kaposi's Sarcoma
LPLD	Familial lipoprotein lipase deficiency
MAIC	Match Adjusted Indirect Comparison
MM	Multiple Myeloma
MRC	Medical Research Council
MRI	Magnetic Resonance Imaging
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NSCLC	Non-small cell lung cancer
Ph+ ALL	Philadelphia Chromosome positive Acute Lymphoblastic Leukaemia
PSM	Propensity Score Matching
PSW	Propensity Score Weighting

PTCL	Peripheral T-cell lymphoma
RCC	Renal cell carcinoma
RCT	Randomised Controlled Trial
RDD	Regression Discontinuity Design
RWD	Real World Data
sALCL	Systemic anaplastic large cell lymphoma
SCT	Stem cell transplant
SMC	Scottish Medicines Consortium
SPC	Summary of Product Characteristics
STC	Simulated Treatment Comparison
STS	Soft Tissue Sarcoma
TKI	Tyrosine Kinase Inhibitor
UK	United Kingdom
US	United States of America

1 INTRODUCTION

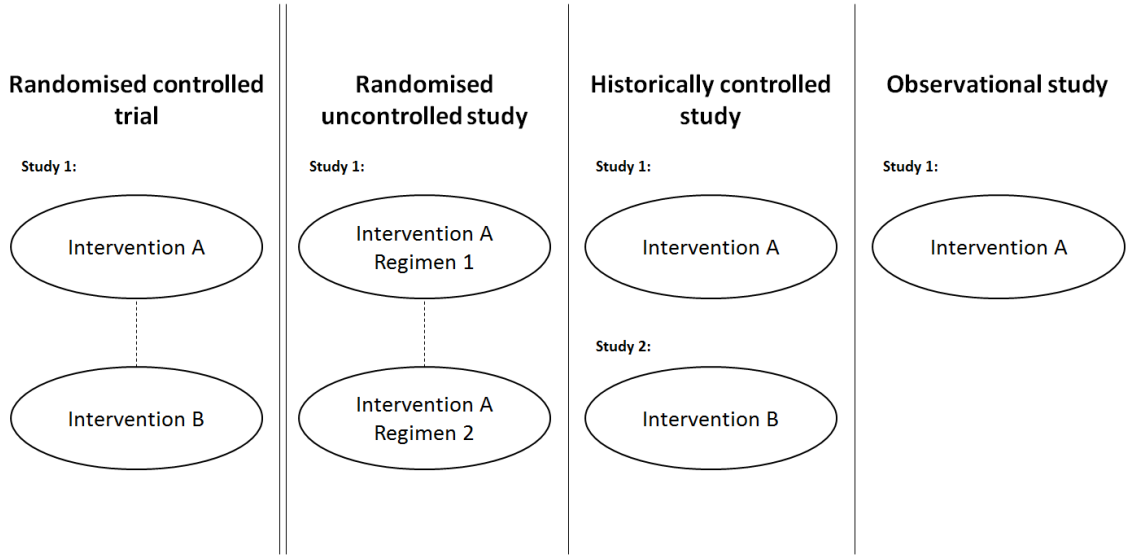
Aspects of this introduction were published in an article in ISPOR Connections July 2014 (Hatswell *et al.*, 2014)

The topic of my research has been pharmaceuticals without a controlled clinical trial – treatments used without 'conventional' evidence of their effectiveness compared to either placebo or active treatment; how frequently has the approach of non-randomised studies for licensing been taken? How have regulators approached the data? And how have statisticians and economists modelled outcomes compared to alternative treatments?

In this context, uncontrolled clinical studies refer to clinical study programmes, where a trial versus control (either placebo, or active) has not been conducted. This can involve a 'single arm trial', or programmes where all of the treatment arms in a clinical trial involve the study drug - although patients may have been randomised, this is merely to different doses or dosing schedules of the active drug with all groups receiving the intervention.

For this reason the terminology of uncontrolled studies or uncontrolled trials has therefore been adopted for this work as although a number of types of study are relevant (shown in Figure 1-1), the common feature is the lack of an internal control when compared to a RCT.

Figure 1-1: A comparison of randomised clinical trials and different forms of uncontrolled clinical studies



Circles denote arms of clinical studies; dotted lines denote comparisons within a single study

In my literature search, I demonstrate that the vast majority of treatments licensed for use in the European Union and United States (the two largest single markets for pharmaceuticals)

are licensed on the basis of randomised controlled trials (RCTs), which on the whole, are conducted on a double blind basis, i.e. neither the physician, nor the patient is aware of treatment assignment.

The use of RCTs is a key feature of most drug development programmes; although a number of alternative views exist on the 'hierarchy of evidence', there is a large degree of concordance between the different systems. In general, a meta-analysis of well conducted RCTs is placed at the top, with individual RCTs then ahead of all other forms of evidence (Evans, 2003). Shown below is an example of a hierarchy of evidence, taken from the Scottish Intercollegiate Guidelines Network (SIGN, 2011).

Figure 1-2: Hierarchy of evidence, directly reproduced from Annex B of 'SIGN 50: A guideline developer's handbook' (SIGN, 2011)

LEVELS OF EVIDENCE

1++	High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias
1-	Meta-analyses, systematic reviews, or RCTs with a high risk of bias
2++	High quality systematic reviews of case control or cohort studies High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
2+	Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2-	Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3	Non-analytic studies, eg case reports, case series
4	Expert opinion

In this hierarchy of evidence, without a RCT, a treatment cannot be said to have 'Level 1' evidence. Whilst in recent years there has been a move away from strict hierarchies of evidence (Rawlins, 2008), with increased recognition of the use of observational studies for safety data, and non-randomised studies to confirm efficacy estimates (and demonstrate efficacy in different risk groups), the RCT still plays a key part in the demonstration of efficacy for new interventions (Pearce, Raman & Turner, 2015; Tugwell & Knottnerus, 2015).

To understand the limitations of single arm trials, it is therefore necessary to understand why the RCT is paramount, the advantages the technique provides, and on which aspects other designs cannot offer the same level of confidence in the effectiveness of treatments.

1.1 THE HISTORY AND THEORETICAL UNDERPINNING OF RANDOMISED TRIALS

In considering why RCTs are generally acknowledged to be the ‘gold standard’ of clinical development (Goldacre, 2009; Kaptchuk, 2001), we must first consider the history of the scientific process surrounding clinical trials.

Pharmaceuticals, as defined in the Oxford English dictionary, are ‘compounds manufactured for use as a medicinal drug’ (Oxford Dictionaries, 2010). As such, many compounds could be classed as pharmaceuticals, and each would have an effect on any given disease, from the minute to the dramatic. In the Platonic world of forms (Penrose, 2006) each pharmaceutical will have an effect on a medical condition (potentially with different effects for subgroups of patients) that is fully known and understood. However in practice we are not able to observe this true efficacy directly, and must instead perform studies to attempt to characterise these effects. Through the use of investigational studies, effects can be estimated and hypotheses can be tested, to understand whether the effect of a given substance on a condition is positive, neutral, or negative (though absolute certainty is never possible).

The route that has been taken in much of science, which has been used in seeking to understand the effect of drugs, is Karl Popper’s scientific epistemology of ‘falsification’ (Kuhn, 1996). A hypothesis is formulated, and then observations are taken to attempt to disprove what has been stated. This idea was originally integrated in the language of formal statistics by Ronald Fisher, who laid out in his book ‘The Design of Experiments’ concepts such as a null hypothesis and factorial designs (Fisher, 1935). This work had a large and lasting impact on research methodology, importantly Fisher recognised that the importance of randomisation in ensuring the difference between groups was due to the treatment effect and not to the selection of subjects (Stanley, 1966). This built on previous work by those such as Galton (who first conceptualised correlation and association), (Karl) Pearson who provided much of the mathematical framework, and along with Neyman operationalised formal hypothesis statistical testing (Goodman, 1999).

Although various (randomised) trials had been conducted in history, the first arguably being the testing of citrus for the treatment of scurvy by James Lind in 1747 (Milne, 2012), the concept was first put into practice in the Medical Research Council (MRC) trial of streptomycin for the treatment of tuberculosis in 1948, shortly after the end of the second World War. At the time the UK was still under rationing, and suffering with both the human and financial cost of the war. This had led to shortages of money and material, and with only a limited amount of the potentially effective drug streptomycin available, Austin Bradford Hill

at the MRC argued that, given the limited supply of the drug, it would be unethical *not* to find out what efficacy level the drug offered. A trial was designed with patients between the ages of 15 and 30 enrolled and given either streptomycin, or standard care (bed rest). After six months, the death rates were 4/55 in the streptomycin group compared with 15/52 for bed rest, thus establishing the efficacy of streptomycin (Crofton, 2006).

This growth in the use of clinical trials was also encouraged by regulators. In 1938 the US Food and Drug Administration (FDA) mandated safety trials for new treatments, and after the widespread use of thalidomide caused catastrophic harm to many children, from 1962 the FDA required 'substantial evidence' of a drug's efficacy, as well as its safety (White Junod, 2015). The official guidance from the FDA (and other regulatory agencies) stresses RCTs as the most acceptable form of evidence for new treatments, comparing the experimental therapy against a control to which superiority is to be expected (French *et al.*, 2010). Beyond the desire of regulators, the role of randomised trials grew with one of the strongest proponents being Archie Cochrane, a director of the MRC and a strong advocate of the RCT. His book 'Effectiveness and Efficiency: random reflections on health services' advocated the use of RCTs to provide reliable evidence (Cochrane, 1972), and ultimately led to the Cochrane Collaboration for systematic reviews.

Another notable contribution of RCTs has been to stop harmful practice, where theoretical arguments have been in favour of a treatment working but RCT results have shown that the treatment actually has no benefit or causes more harm than good. One notable example was the CRASH trial, which demonstrated that in traumatic brain injury, steroids increased the mortality rate compared to no treatment, challenging what had been standard medical practice for 30 years on the basis of small studies and theoretical arguments (CRASH Trial Collaborators, 2004).

In addition to RCTs demonstrating the efficacy treatments, a further factor is the harm that can be caused by pharmaceuticals *not* being studied in controlled trials. This may result in patients being given ineffective (or even harmful treatments), potentially in place of effective therapies. The uncertainties associated with unstudied treatments is a reason for the regulation of 'off label' promotion of pharmaceuticals, for which pharmaceutical companies have been fined large sums of money in recent years (Fugh-Berman & Melnick, 2008).

Such is the acceptance of RCTs within medicine (as seen with the place in the hierarchy of evidence and more), even extremely rare conditions have been able to enrol for large RCTs within short periods of time (Gaddipati *et al.*, 2012; Prasad & Oseran, 2015). In Phase 2 studies randomisation to an internal control helps likely efficacy of treatment and thus reducing the number of treatments that fail (expensive) Phase 3 studies. Due to this, current

estimates are that around 28% of Phase 2 oncology trials include a placebo control (Grayling *et al.*, 2019).

1.2 'CONVENTIONAL' RANDOMISED DOUBLE BLIND TRIALS

As discussed by Fisher, Cochrane and others, RCTs have many advantages over other forms of study. These advantages are addressed in turn below, along with how each issue is relevant to the topic of uncontrolled studies

1.2.1 RANDOMISATION AND EXCHANGEABILITY

Randomisation provides an unbiased basis for the testing of an intervention. If randomisation is performed correctly any differences attributable only to the role of chance – according to Meier 'the role of randomization is to distribute the effects of baseline variables, both the measured ones and those not observed, in such a way that the statistical analysis makes due allowance for them' (Meier, 1975).

The use of randomisation therefore leads to the removal of any selection bias that could otherwise be present. A typical example is the selection of patients based on which treatment physicians believe may be more appropriate, which may lead to a difference in outcomes with even an inert intervention. Although there is no conclusive evidence, data suggests that a degree of selection bias may occur in non-randomised studies (Fellow & Director, 2008). With randomisation providing two comparable groups, given an infinite sample size, the difference between groups should be attributable to the difference in treatment effect. As the sample size decreases, there is an increased role of chance in the process; however this is calculable and allows the estimation of a probability that the outcome is not a 'true' difference – dating back to the tests originally conceived by Neyman and Pearson (Sterne & Smith, 2001).

When an uncontrolled trial design is employed, randomisation to a control arm is not possible and therefore any selection bias in patients cannot be accounted for or quantified – the trial results may be driven by the characteristics of the patients enrolled in the study, rather than the intervention under study. It may be that the 'correct' conclusion may be reached, but the results cannot be relied upon to be unbiased (unlike a well conducted RCT), and with no mechanism to verify the results seen. As previously discussed, it is possible for uncontrolled trials to have an element of randomisation to different dosages or administration schedules of the investigational drug (and not to a control group). This is the most obvious deficiency of uncontrolled trials: the lack of a control group, randomised, or otherwise. Thus, an uncontrolled trial will generate evidence of the outcomes seen with a

treatment, but does not answer the question relevant for regulators, physicians or economists: how well an intervention works *compared to current practice*.

1.2.2 BLINDING OF PATIENTS AND PHYSICIANS

A second factor in the 'ideal' study is the 'blinding' (or 'masking') of patients and clinicians to the intervention received. When this is performed, neither the clinician nor the patient is aware of which treatment arm the patient is assigned to, preventing any inbuilt bias for (or against) either treatment from affecting the outcome of the study. Concerns around ensuring that studies are blinded also lead to attention on other aspects of clinical study design, for example the use of independent central review committees, who verify clinical measures such as tumour size without seeing the patient, or having hints as to the assigned treatment.

Although not essential to the conduct of a good study with an objective endpoint (for example overall survival), blinding is increasingly important when endpoints are subjective or open to interpretation on the part of the reviewer (for example the reading of scans). Schultz and Grimes (2002) discuss how these results may be affected.

> If investigators are not blinded, their attitudes for or against an intervention can be directly transferred to participants. Their inclinations could also be manifested in differential use of ancillary interventions of supplemental care or treatment (co-interventions) or differential adjustments to the medication dose. Investigators might also encourage or discourage continuation in a trial on the basis of knowledge of the intervention group assignment. (Schulz & Grimes, 2002:p.696)

Empirical evidence of a bias from uncontrolled studies was shown in a study by Schulz et al., where unblinded trials had higher estimates of treatment effectiveness than studies where allocation was unclear. These in turn had higher estimates of effectiveness of studies that were adequately blinded (Schulz KF *et al.*, 1995). More recent research has seen similar results in trials with binary outcomes, with unblinded assessors (compared to blinded assessors) reporting substantially biased effect estimates, exaggerating odds ratios by 36% in randomised controlled trials due to misclassification of some patients (Hrobjartsson *et al.*, 2012).

In an uncontrolled trial, blinding is not feasible as all patients will be receiving the treatment – any concerns regarding bias would increase as the objectivity of the trial endpoint decreases. This lack of blinding increases the risk of bias (conscious or subconscious). To guard against this, open label trials often use independent review panels, which are charged with reviewing patient information to determine the effectiveness of interventions.

1.2.3 MULTI-CENTRE TRIALS

A third important factor in a 'good quality' study is the use of multiple centres to increase the generalizability of results. Without this, results of any trials could be seen to be specific to the setting in which the trial was conducted.

These specific outcomes could be linked, for example, to the types of patients at the centre (a specialist centre may attract more complex patients), the staff and protocols used at the centre (for example the approach to dose titration in the presence of adverse events), or to any number of other factors that may either be observable or unobservable.

Although not necessarily subject to this limitation, uncontrolled studies are more likely to be conducted in low numbers of patients and low numbers of centres. As such, there may also be concerns regarding the reproducibility of study results. It should also be noted that these concerns may apply to historical controls, a method used for comparison with uncontrolled trial data; if patients in a historical control are all enrolled from one centre, they may not be representative of the wider population.

1.3 MATHEMATICAL NOTATION FOR RANDOMISED CLINICAL TRIALS

The above concepts can be summarised mathematically, using notation that will be used throughout this thesis.

The goal of medical research, through the vehicle of RCTs can be said to estimate the effect of the interventions applied to the individuals/population of interest. For each individual $i = 1, \ldots, n$ included in the study, the treatment indicator is denoted as T_i. The interventions of interest will differ, and may well include a novel treatment $T_i = t$, and a control $T_i = c$. Although this latter is usually termed 'control', due to equipoise, more generally this represents 'best standard care' which may involve an active agent, or supportive care (potentially with a placebo element). Where placebo treatments are used, these are added to standard care, such that patients receive treatment on the study that is at least as good as they would have done otherwise - but are blind as to whether they are receiving the intervention.

If the outcome of the study is denoted as Y_i, the comparative effectiveness of treatment would therefore be defined as the *population average* outcomes seen with the novel treatment

Equation 1: Population average outcomes from treatment *t*

$$\overline{Y}_t = \frac{\sum_{i=1}^{n} Y_i \mathbb{I}(T_i = t)}{\sum_{i=1}^{n} \mathbb{I}(T_i = t)}$$

compared to the control

Equation 2: Population average outcomes from control *c*

$$\overline{Y}_c = \frac{\sum_{i=1}^{n} Y_i \mathbb{I}(T_i = c)}{\sum_{i=1}^{n} \mathbb{I}(T_i = c)}$$

In the above equations, $\mathbb{I}(T_i = t)$ and $\mathbb{I}(T_i = c)$ are indicator functions, taking values 1 if the argument is true (i.e. if the i-th individual is in the active treatment or control group, respectively) and 0 otherwise. Thus, the denominators simply count the number of individuals in each treatment arm, while the numerators select the individuals to whom either of the treatments are applied. Typically, we are concerned with differences in these averages to describe the effect of interest, e.g. $\Delta = \overline{Y}_t - \overline{Y}_c$. However, other estimands of interest may be defined (e.g. ratios or other non-linear functions).

The comparative effectiveness of treatments however is complex, as studies are conducted in real world populations which (even assuming the same condition) have differences in both observable and unobservable characteristics. Here the ideal would be that the patient populations exposed to each treatment are of infinite size, and independently and identically distributed (*i.i.d*). In reality this assumption can never be met – even aside from the need to limit sample size, in many cases exposure to a treatment would impact a patient's outcomes and characteristics. This means a more general assumption is required, that of 'exchangeability'.

Greenland and Robins in a seminal 1986 paper and then in a 2009 update, define populations as 'exchangeable with respect to an outcome measure if their outcomes would be the same whenever they were subjected to the identical exposure history' (Greenland & Robins, 1986, 2009). The assumption underpinning exchangeability is that treatment assignment is independent of patient characteristics (denoted by the vector $\boldsymbol{X_i}$), mathematically $T_i \perp\!\!\!\perp (\boldsymbol{X_i})$. In the context of clinical studies, the matrix $\boldsymbol{X}$ (collecting the values of the covariates for each individual) may be considered to be extremely broad, including factors such as the healthcare system in which a patient is treated (which will vary over time and space). Where the assumption of exchangeability is not met, the results (again using the terminology of Greenland and Robins) are said to be confounded, which would lead to bias in any comparison of outcomes (if not adjusted for).

An RCT seeks to achieve exchangeability by enrolling patients and having chance determine treatment assignment, with others characteristics (both observable and unobservable)

equally distributed between arms. Mathematically this would be shown as $Pr(T_i = t|\boldsymbol{X_i}) = Pr(T_i = c|\boldsymbol{X_i})$. To obtain a stable estimate of the difference in effects (which may include efficacy, and safety outcomes) between treatments, sufficient patients, n, are also needed in the study to account for variability in both patient outcomes, and treatment effect. This number of patients is one of the key parameters for a trial sample size calculation, along with the acceptable Type I (false positive) and Type II (false negative) error rates. Mathematically these are usually defined as probabilities α and β (Lachin, 1981).

The issues that may be present without RCTs would apply to many of these areas;

- The lack of blinding implied in an uncontrolled study (where there is no alternative arm) may be an issue if this affects the outcome assessment, leading to the non-independence of outcome measurement and treatment assignment
- In the absence of randomisation between groups and a need for cross study comparisons, there may be differences in the characteristics of patients included in studies ($\boldsymbol{X}$) which may mean groups are not exchangeable i.e. are confounded, and would be anticipated to have different outcomes; even with the same treatment assignment. This appears to be the major concern with uncontrolled studies
- Without sufficient patients, N, there may be substantial uncertainty in the size of any effect as the randomisation may result in a difference (due to chance) of differences between arms; the sample size is therefore a key consideration in power calculations for clinical trials. This is compounded if comparing between - and not within - studies, as there is additional (potentially unquantifiable) uncertainty introduced - as raised by (Byar *et al.*, 1990)
- In comparing across studies there may have been differences in how endpoints were measured, for instance the type of scan or definition of progression. Should this be the case it may be that outcomes (Y) are non-comparable between studies, or at the very least, may need to be re-estimated
- Beyond the outcome measurement, studies may also report (or not fully report) the methods used for determining outcomes; for instance having a primary endpoint of duration of response and therefore not providing information on non-responders – again meaning that outcome Y cannot be compared between studies
- If a trial is not performed across multiple centres (as is often the case with uncontrolled studies, in being smaller), there may also be concerns that there are structural differences which could affect a number of parameters in our model; patients may be systematically different (affecting $\boldsymbol{X}$), or outcome assessment may not be the same as in other centres (affecting Y)

1.4 PREVIOUS WORK IN ASSESSING EVIDENCE WITHOUT RANDOMISED TRIALS

There has been relatively little work in how best to assess the efficacy of interventions that do not have randomised controlled trials. The work that has been published is discussed below.

1.4.1 ASSESSING THE COMPARATIVE EFFECTIVENESS OF TREATMENTS STUDIED WITHOUT A RANDOMISED CONTROL

Much of the discussion in the literature describes criteria for the acceptance of treatment effectiveness without RCT data. To identify past work unstructured literature searches were conducted (as no specific key words are available for this issue), combined with hand searching the reference lists of relevant papers.

In a BMJ letter (Black, 1994), Nick Black argued that a RCT is not required in cases where the effect size is extremely large (his example is ventricular fibrillation), where a RCT would need to be unfeasibly large (for example rare adverse events), where long term outcomes are needed (for example hip prostheses), where clinicians would not accept a RCT (and observational data may convince them of uncertainty in their beliefs), and where practical or ethical concerns make a RCT impossible (for example reorganisation of healthcare services, or admission to intensive care).

The criteria of an extremely large effect size is similar to that proposed in the Oxford Centre for Evidence Based Medicine criteria for acceptance of uncontrolled studies (Phillips *et al.*, 2009). In their hierarchy, evidence can be judged to be of the highest grade (grade 1), if it meets the ‘all / none criteria’ where all patients experienced an outcome before the introduction of a therapy, whereas none experience the outcome with the intervention – death and ventricular fibrillation would meet this criteria.

Other work of relevance is that by Glasziou et al. (Glasziou *et al.*, 2007), and considers that interventions (not necessarily pharmaceuticals) can be deemed to be effective if the treated and untreated observations are taken from the same pool, and there is a ‘dramatic’ rate ratio for the intervention. The rate ratio was defined as the amount of time with the condition, divided by the amount of time for the intervention to take effect. For example if a patient’s heart has stopped for 60 seconds, and is restarted within one second of a defibrillator being used, the rate would be 60/1, i.e. 60.

For a dramatic effect size, a rate ratio of 10 is stated to be a rule of thumb, where the effect is unlikely due to chance or confounding variables (Glasziou *et al.*, 2007). The authors conclude that the most obvious candidates for their criteria being met are mechanical interventions (where the intervention is obvious, and has a prior expectation and theory as to why it will work), on a stable background (i.e. no varying conditions where chance may play a part), where a dramatic improvement is made. Related to this idea of a dramatic effect size Gerstein et al. (2019) in dismissing the use of observational data as a routine part of drug development state that in order to believe data from observational sources, they would wish to see a ratio of four in the effect size to minimise the risk the result is due to confounding. Historically such values are also stated to be convincing to Austin Bradford-Hill, who stated a 20-30x effect size was unlikely to be due to chance, but a 2-3x effect size might be (Hill, 1965).

Beyond this work, the majority of existing literature relates to the use of historical controls, which are discussed extensively in Section 1.4.3.

1.4.2 EFFECT SIZES SEEN IN OBSERVATIONAL DATA, COMPARED TO RCTS

Work by Colditz et al., looked at the effect sizes seen with randomised and non-randomised designs in the fields of cardiology, neurology, psychiatry and respiratory medicine. The results of the literature review showed that studies using a non-randomised design had larger effect sizes than unblinded RCTs, and that unblinded RCTs had larger effect sizes than blinded RCTs (Colditz, Miller & Mosteller, 1989). A comprehensive review of the literature found a similar pattern across 45 disease areas, with RCT results ($n = 240$) showing a smaller effect size than non-randomised studies ($n = 168$), although on the whole, the studies reached the same conclusion, i.e. the treatments remained effective (or ineffective), with the direction of effect not changing (Ioannidis *et al.*, 2001).

Two similar studies have been funded by the UK National Institute for Health Research (NIHR) and published as health technology appraisal (HTA) reports, on the effect sizes seen when comparing randomised study designs and observational study designs.

The first report is titled ‘Choosing between randomised and non-randomised studies: a systematic review’ (Britton *et al.*, 1998). The systematic review identified 18 treatments with both RCT and non-RCT evidence (for example case control studies) and found that whilst the effect sizes between the two types of study varied, there was no identifiable systematic bias in the direction of effect. They also highlight several issues that may lead to differences from a blinded RCT, most notably patient selection, patient preference in observational data,

and publication bias in observational studies (which are less likely to be published than RCTs, particularly if neutral). The authors conclude that for comparisons to be made between studies, the patient characteristics should be well matched, and that whilst baseline characteristics could be adjusted for, this should be done in a rigorous way.

The second HTA report (MacLehose *et al.*, 2000) is titled 'A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies' and focusses on comparing effect sizes in two specific interventions; mammographic screening to reduce breast cancer mortality, and folic acid supplementation to prevent neural tube defects. The results of the review were that non-randomised studies assessed as low quality (using a checklist) were more likely to differ from RCT evidence than high quality non-RCT evidence (which gave approximately equal effect sizes). The authors conclude that provided the observational evidence is deemed to be high quality (with confounding data controlled for), these comparisons may be appropriate. However they caution against the generalizability of their findings to other contexts as they found relatively few papers discussing the topic (n=38), most of which were in the same diseases (cardiovascular conditions in particular), and temper their conclusions with concerns regarding publication bias.

Although the majority of meta-studies have concluded that effect sizes are exaggerated in observational data, this is not a universal finding. Benson & Hartz (2000) found in 136 reports of 19 treatments, that observational studies produced similar effect sizes to RCTs. Of the 19, only two estimates fell outside the 95% confidence interval (which statistically is around the number that might be expected). The differences in this study may be due to chance or the type of studies included – in specifically investigating published observational studies (with well designed protocols to attempt to reduce bias), the issue may more lie in the use of unconnected data and comparisons made from outcomes between studies.

Whilst these studies provide interesting information and raise valid points, they are not directly relevant to this thesis due to the fields considered. The published work looks at either interventions with non-RCT evidence *and* RCT evidence in fields with large sample sizes and well conducted RCTs (Ioannidis et al., Colditz et al. & Britton et al.,), or at non-pharmaceutical interventions (MacLehose et al.). Colditz et al. discuss this issue in their conclusion, stating:

> Studies that use external controls or an observational design occur rarely in the evaluation of medical therapies. This may reflect, in part, the requirements of the U.S. Food and Drug Administration (FDA) that evaluations of new therapies require randomized controlled trials. The small number of studies in these two categories of design preclude any firm conclusion regarding possible biases encountered with them. (Colditz, Miller & Mosteller, 1989:p.451)

As this work is aimed at precisely those treatments approved without RCT evidence, the published literature will inform the approach, but cannot necessarily be generalised to these cases where RCT evidence is not available for a variety of reasons – it is precisely these 'small numbers of studies' in which I am interested.

1.4.3 HISTORICAL CONTROLS

The use of 'historical controls' appears to be the most widely used and referred to approach, with the different aspects of their use discussed in turn.

1.4.3.1 ORIGINAL USAGE OF HISTORICAL CONTROLS

The first use of a historical control I have been able to find (relating to medicine) was in 1884. At the time a condition named 'beriberi' was rife throughout the Japanese Army and Navy, disabling (and killing) personnel, massively reducing the effectiveness of the services. The previous year, the Japanese naval ship Ryujo was on a training voyage which resulted in 45% of sailors falling ill (and 25/376 die). In order to prove his theory that the cause of the disease was not a virus or ethnic link (as was commonly believed), but linked to the diet of crews - the true cause being a deficiency in vitamin B1 - the vice-director of the Naval Medical Bureau, Kanehiro Takaki, persuaded the Japanese emperor to let him control the diet of a ship to prove his theory (Simpson, 2014).

To prove that protein deficiency was the cause of the beriberi, Takaki took the ship Tsukuba on a training mission, using the same route and schedule as the ill fated Ryujo (but one year later) in order to minimise the potential for differences. Of the 333 sailors onboard, only 14 developed the condition (all of whom had not been eating their rations properly). The impact was profound - 1878-1883 saw a mean of 1586 cases of beriberi, which fell to 41 in 1885 (the year after Takaki's theory was demonstrated), 3 in 1886, and then to zero thereafter (Sugiyama & Seita, 2013; Takaki, 1906; Simpson, 2014). In recreating the original conditions as much as possible to minimise confounding, Takaki had effectively used the voyage of the Ryujo as a historical control to demonstrate the effectiveness of his intervention (barley added to rice).

The first formal usage of a historical control was published by Pocock (1976), suggesting that historical data from similar patients could be used to either add power to new trials (reducing the sample size), or provide a control group for a uncontrolled study. Pocock then defined similarity to be the historical control having met the following criteria:

> 1. Such a group must have received a precisely defined standard treatment which must be the same as the treatment for the randomized controls.

2. The group must have been part of a recent clinical study which contained the same requirements for patient eligibility.
3. The methods of treatment evaluation must be the same.
4. The distributions of important patient characteristics in the group should be comparable with those in the new trial.
5. The previous study must have been performed in the same organization with largely the same clinical investigators.
6. There must be no other indications leading one to expect differing results between the randomized and historical controls. For instance, more rapid accrual on the new study might lead one to suspect less enthusiastic participation of investigators in the previous study so that the process of patient selection may have been different. (Pocock, 1976:p.177)

The work then lays out several examples with trials conducted by the Eastern Cooperative Oncology Group in melanoma within a short time period. By using data from the historical controls as well as randomized controls, the sample size needed for a new study can be greatly reduced, assuming the historical controls are indeed exchangeable with new patients.

In the article, Pocock is also mindful of the potential for bias (suggesting methods to reduce the effect size in the historical control for perceived bias) and stating that due to the absence of a control arm, there is no way to be certain that the groups are comparable. His conclusion is that the results of historical comparisons should be viewed as a tool to allow the estimation of effect size from uncontrolled studies, but not being as reliable as RCT evidence.

1.4.3.2 THE APPROPRIATENESS OF COMPARISONS USING HISTORICAL CONTROLS

Historical trials have been long suspected of ecological bias and ‘stage migration’, where due to advances in diagnostic technology patients are diagnosed earlier - for example from Magnetic Resonance Imaging (MRI) detection of lesions, rather than clinical diagnoses. The effect of this artefact is a seemingly improved prognosis for all stages of disease in the absence of any actual change (Sormani, 2009). Due to issues like this, several studies have investigated disease areas where both RCT and historical control information are available.

The first study of this type identified was published by Sacks, Chalmers & Smith. For the six therapies investigated, 50 RCTs and 56 historically controlled trials were identified. In 79% of historically controlled trials the intervention was found to be effective, compared to only 20% of RCTs (Sacks, Chalmers & Smith Jr., 1982). The authors highlight that the results with the intervention were similar between studies, but that it was the control arm which underperformed in the historically controlled trials – speculating that a bias in patient selection may be the cause of the discrepancy.

Diehl & Perry investigated the same question looking at overall survival or relapse free survival in oncology, finding 43 examples in the literature of well-matched historical cohorts and RCT control groups. However when comparing the outcomes of the two groups, 18 of the 43 studies had a greater than 10% difference in effect size between the control groups – the randomised group performing better on 17/18 occasions (Diehl & Perry, 1986).

The appropriateness of historical controls was also raised indirectly in a comparison of outcomes from a study comparing the results of Phase 2 and Phase 3 studies using identical chemotherapy regimens. Of the 43 chemotherapies identified, the mean response rate was 12.9% higher in Phase 2 studies indicating that the role of chance and selection bias is notable (Zia *et al.*, 2005). In paediatric oncology, Moroz et al. found a similar result using 42 studies identified in the literature where historical data had been used to calculate the sample size and power of the study. They found that the randomised control (of the same intervention) had a median improvement in outcome of 5.0% over the time to event data from the historical control group (Moroz *et al.*, 2014). The paper does not attribute the difference in outcomes to the selection criteria, improved standard of care, or drift over time, simply remarks on the difference.

Where there may be differences in historical cohorts, other research has been conducted in the development of tests for use with historical controls (for example for futility in the case of superior treatments), with the authors suggesting stratifying patients by key characteristics so as to provide a similar patient group to the historical control (Wu & Xiong, 2016).

Given these known issues, a simulation study by Tang et al. (2010) establishes that only a small ‘drift’ in patients over time is needed for false positives to occur in the estimates of effect size. Similarly the typically smaller sample size used in uncontrolled studies is also listed as a potential source of error due to chance, though even increasing this sample size does not counter the issues should the underlying control data have changed over time.

The most directly relevant and perhaps concerning study however comes from Snyders et al. (2019), who performed a systematic review and meta-analysis of the outcomes of all docetaxel arms in advanced non-small cell lung cancer trials over a 17 year period; this amounted to over 10,000 patients from 63 studies. They found substantial heterogeneity in the outcomes, with response rate ranging from 0-26% (pooled estimate of 8%), and PFS ranging from 1.4 to 6.4 months (the ‘mean of the median’ was 3.0 months). The paper indirectly also gives an idea of the ‘drift’ seen in outcomes, with each year seeing an improvement of 0.3% in response rate, and 0.5% improvement in PFS – with similar results seen with changes in overall survival.

Although there appears to be no confirmation that historical controls are appropriate for naïve comparisons (and multiple sources to suggest they are indeed, inappropriate), recent work in the area has suggested the creation of a cross-industry historical controls database (Project Data Sphere, 2015; Desai *et al.*, 2013). The aim for this project being the use this 'big data' to emulate trials (Hernán & Robins, 2016). Other work has investigated how best to adapt estimates for perceived bias, or according to mean estimates of bias in study design (Turner *et al.*, 2009).

This litany of issues relating to historical controls is recognised in existing guidance for the use of uncontrolled Phase 2 oncology studies (Rubinstein *et al.*, 2009; Seymour *et al.*, 2010) where the suggestion is made to attempt to create a similar group based on observable characteristics using data from large meta-analyses, such as that by Korn et al. in melanoma (Korn *et al.*, 2008), to avoid the potential for differences between stages. In part due to the issues in creating unbiased comparisons, there have also been a number of papers discussing the role of randomisation in Phase 2 studies – even if not used for registrational purposes (Rubinstein *et al.*, 2009; Grayling & Mander, 2016).

1.4.4 METHODOLOGIES FOR THE USE OF OBSERVATIONAL DATA IN ESTIMATION OF EFFICACY

Whilst I am interested in the estimation of efficacy where RCT data are not available, there is a degree of overlap with methods for the estimation of efficacy using observational data. The research in this area is more developed, with methodologies such as propensity scoring used to estimate safety risks and efficacy using large datasets in areas such as cardiovascular disease (Freemantle *et al.*, 2013).

In the terminology used for this form of observational data, Mathes & Pieper (2017) draw an important distinction between the different forms of historical studies. In their work they are careful to discuss the difference between case series (where all patients receive an intervention) and cohort studies (where different patients will receive different interventions, allowing for cross group comparisons). The implication here being that the use of appropriate methods may allow for unbiased estimates to be drawn from cohort studies, whilst comparison between case series will necessarily include a further level of uncertainty from between study comparisons.

Two key review papers in this area outline available methodologies - Rovithis (2013) conducted a literature review of all methods used to estimate effectiveness based on observational data. This work was conducted as a part of a wider review to investigate interventions in neonates (where limited evidence is generally available) and limits itself to

looking for evidence of matching (in different forms), regression analysis, propensity scores, instrumental variables, as well as difference-in-differences approaches, looking in particular where these methodologies have been used in cost-effectiveness analyses and finding few applications; 43 in total. The majority of the studies identified were for medical or surgical interventions, and mostly used retrospective observational data. None were analyses of pharmaceuticals.

A more complete analysis is presented in NICE Technical Support Document 17 (Faria *et al.*, 2015), which discusses the methodologies available to be used for the estimation of treatment effects when using observational data when individual level data (ILD) is available for both datasets. Following a review of all the methodologies available, recommendations are made on when each methodology is appropriate in the form of a (3 part) flow chart; Figures 1-3 in the document.

The methods suggested are discussed below, in terms of their suitability for use with uncontrolled studies. In addition several additional methodologies or approaches not highlighted by Faria et al., but that are also relevant, are presented.

1.5 THE ROLE OF UNCONTROLLED STUDIES IN DRUG APPROVALS

Work from the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), a multi-regulator consortium to set standards for drug approval across all participating nations, sets the standards for what evidence is expected. Due to the limitations of uncontrolled data, the ICH E10 guidelines (Choice of Control Group in Clinical Trials) state that a control group should be used 'to allow discrimination of patient outcomes… caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, observer or patient expectations, or other treatment' (ICH Harmonised Tripartite, 2000:p.6). This recognises the issues raised by Pocock and others regarding the comparability of historical data, with the ICH expressing:

> serious concerns about the ability of [historically controlled] trials to ensure comparability of test and control groups and their ability to minimize important biases, making this design usable only in unusual circumstances (ICH Harmonised Tripartite, 2000:p.7)

The ICH guidelines do discuss the appropriateness of trials without a control group. In the absence of formal mechanisms for effectiveness with the conclusion of the ICH regarding control groups is that:

> In unusual cases, the course of illness is in fact predictable in a defined population and it may be possible to use a similar group of patients previously studied as a historical control (see

> Section 1.3.5). In most situations, however, a concurrent control group is needed because it is not possible to predict outcome with adequate accuracy or certainty (ICH Harmonised Tripartite, 2000:p.6)

The unusual cases it describes as being appropriate for uncontrolled studies are described below.

1.5.1 'OBVIOUSNESS'

Whilst randomised trials do have advantages, it has been noted that a degree of medical knowledge is supported by only observational data, some of which is given greater weight than data supported by well conducted randomised trials, meta-analysed together. Examples of such knowledge as cited by Glasziou et al. (2007) involve the causal role of acetylsalicylic acid in Reyes syndrome, and the role of a third copy of Chromosome 21 in causing Down syndrome.

Similarly, there is a number of interventions either approved or in widespread use due to the 'obviousness' of their efficacy. These include laser therapy for 'port wine stain' birthmarks, and ganciclovir for cytomegalovirus retinitis – all patients untreated went blind, whilst treated patients did not (Rawlins, 2013). A humorous article in the British Medical Journal by Smith & Pell (2003) describes a systematic review for the effectiveness of parachute use to prevent injury from 'gravitational challenge', underlining that not all interventions that are known to be effective are necessarily studied in randomised trials. This is reflected in the ICH guidance, which states:

> In some cases sensitivity to drug effects is clear from the consistency of results of prior placebo-controlled trials or is obvious because the outcome of treated and untreated disease is very different. For example, in many infectious diseases cure rates on effective treatment far exceed the spontaneous cure rates over the course of a short term study (ICH Harmonised Tripartite, 2000:p.13)

The obviousness of treatment efficacy is difficult to ascertain with a degree of judgement needed as to the implication of results. This also relates to the difficulty of knowing whether an intervention works and the proposed 'rate ratio' (Glasziou *et al.*, 2007), as well as the Sackett all / none criteria (Phillips *et al.*, 2009).

In this way, previously published evidence and a well understood disease pathway would be relevant – a progressive disease such as retinitis is a much better candidate for an uncontrolled trial than relapsing-remitting conditions such as Crohn's disease or multiple sclerosis – the ICH give similar examples of conditions where uncontrolled studies are not appropriate:

> such conditions tend to include those in which there is substantial improvement and variability in placebo groups, and/or in which the effects of therapy are small or variable, such as depression, anxiety, dementia, angina, symptomatic congestive heart failure, seasonal allergies, and symptomatic gastroesophageal reflux disease (ICH Harmonised Tripartite, 2000:p.13)

1.5.2 CLINICAL EQUIPOISE

One issue in the design of clinical studies is that of clinical equipoise, which relates to the ethics of conducting a controlled trial. This problem was described by Freedman et al. (1987) as a clinician having genuine uncertainty regarding which treatment is superior. His conclusion (which mirrors that of other literature) is that as soon as the consensus is reached that one treatment is superior, this is the treatment that should be offered to patients, with the inferior treatment phased out.

The issue of clinical equipoise manifests itself in different ways, depending on the condition considered

- Where a standard of care exists, it determines that this standard of care should be the comparator used in a clinical trial, until the trial confirms either the new intervention or the standard of care to be 'best' (or shows similarity of effect between the two). By entering the trail however patients would receive treatment, and not be disadvantaged.
- Where no standard of care exists to compare against, then an uncontrolled study may be the ethical choice. This can occur where a new intervention is to be trialled after existing proven therapies have failed (and re-challenge is unlikely to be effective), or the practitioner has reason to believe the new therapy has greater chance of being effective (implicitly this means outcomes without treatment must be poor).
- A further argument can be made that a placebo control would be unethical if there is an immediate threat to life, and thus if no active therapy with evidence of efficacy is available (or off label treatment), then again an uncontrolled study may be the appropriate choice (Ellenberg & Temple, 2000; Temple & Ellenberg, 2000).

The discussion of whether clinical equipoise is met is, in many ways, an extension of the discussion on whether new drugs should be compared to placebo or an active therapy. Whilst there is continuing disagreement regarding the ethical arguments for clinical equipoise (and indeed placebo controls), it would appear that the status quo of a preference for placebo controlled trials is ethical in restricted set of circumstances, with a placebo being

inappropriate if patients are put at risk of a well understood, substantial (and irreversible) harm from not being treated immediately with an effective therapy (Miller & Joffe, 2011).

The issues concerning equipoise are reflected in the ICH guidelines, which implicitly include a description of equipoise within the discussion of ethics, stating:

> For example, deferred treatment of pain or other symptoms may be unacceptable to patients or physicians and they may not want to participate in a trial that requires this. Whether a particular placebo controlled trial of a new agent will be acceptable to subjects and investigators when there is known effective therapy is a matter of investigator, patient, and institutional review board (IRB)/independent ethics committee (IEC) judgment, and acceptability may differ among ICH regions (ICH Harmonised Tripartite, 2000:p.16)

This is similar to a statement included in FDA guidance on endpoints in cancer trials for when single arm trials are appropriate, though noting the limitations of these studies:

> In settings where there is no available therapy and where major tumor regressions can be presumed to be attributed to the tested drug, the FDA has sometimes supported ORR and response duration observed in single-arm studies as substantial evidence supporting accelerated approval. Response rates have been used in settings such as acute leukemia for regular approval where complete responses have been associated with decreased transfusion requirements, decrease in infections, and increased survival. Single-arm trials do not adequately characterize time-to-event endpoints such as survival, TTP, or PFS. Because of variability in the natural history of many forms of cancer, a randomized study is necessary to evaluate time-to-event endpoints. (Food and Drug Administration, 2007:p.11)

Equipoise however is however not something that can be objectively measured, and will vary depending on the context and belief in the intervention being discussed. This difficulty is highlighted in paper in the Journal of Clinical Epidemiology which combined reported data on lion attacks with his observed frequency of umbrella carrying and lion attacks in the US (Anderson, 1991). In his (statistically significant) finding that umbrellas prevent lion attacks, he states on the possibility of conducting a randomised trial ‘The study would have been unethical for me anyway, as I would not want to subject any participants randomly to the risk of being umbrella-less in the presence of a lion’.

1.5.3 THE BENEFIT-RISK OF TRIAL PARTICIPANTS AND PATIENTS

Another issue that may lead to the use of an uncontrolled study for the basis of a new drug application is related to the risk to participants of being included in a trial. The ICH Guideline E10 describes this issue as:

> Use of a placebo control may raise problems of ethics, acceptability, and feasibility, however, when an effective treatment is available for the condition under study in a proposed trial. In cases where an available treatment is known to prevent serious harm, such as death or irreversible morbidity in the study population, it is generally inappropriate to use a placebo control. There are occasional exceptions, however, such as cases in which standard therapy

> has toxicity so severe that many patients have refused to receive it (ICH Harmonised Tripartite, 2000:p.16)

An editorial by Emmanuel and Miller (2001) in the New England Journal of Medicine encourages the sample size included in a study to be thought of as the number of people put at risk to gain information. In some cases, should there already be deemed to be sufficient information, a controlled trial may be deemed inappropriate due to the harm caused to patients of being untreated, and the value of information gained either being unable to offset this risk or being 'low value'. With a strong suspicion of effectiveness, the knowledge gain from a controlled trial may be therefore insufficient to justify the risk to patients of acquiring that knowledge.

An example of this approach was the licensing of an extended release version of lamotrigine for the treatment of epilepsy. As the disease had been well studied previously, a large amount of information was available regarding the performance of a control condition; a dosage of antiepileptic too low to prevent seizures effectively (but sufficient to prevent the most serious types of seizure). A collection of eight historical control groups, all with similar performance, allowed the FDA to consider these to be a well-established control, to which an extended release formulation of lamotrigine (a drug that had been on the market for several years) could be compared (French *et al.*, 2011, 2010). The extenuating circumstances here were that, firstly, this was a new formulation of an existing (proven) therapy, secondly, that the expected behaviour of the control arm was well understood due to the number of previous studies. Further, the risk to patients of being treated with a low dose control was unethically high as seizures can cause long-term damage as well as short-term distress.

The issue of benefit-risk for trial participants is also related to the concept of clinical equipoise, where if a patient has no viable treatment option and a bleak prognosis, the potential benefits from a treatment with an unproven mechanism of action may outweigh the risk of receiving an ineffective (or even harmful) treatment. It should be noted that this should also not be seen as an all or nothing decision, as trials are frequently conducted maintaining a placebo arm, but having unequal treatment allocation (for instance 2:1) between investigational products and placebo (Chow & Chang, 2019), or using approaches such as 'crossover' designs (Ishak *et al.*, 2014).

1.6 THE USE OF UNCONTROLLED STUDIES IN MODELLING AND HEALTH TECHNOLOGY APPRAISAL – GUIDANCE FROM AGENCIES

Once treatments have been approved by regulatory agencies (with any associated analyses performed), in many healthcare systems (particularly publicly funded ones) they are

assessed for their cost-effectiveness; resources spent on one patient cannot then be spent on another patient. The principle of cost-effectiveness is that if a new treatment for one group of patients is funded, it should provide *at least* as much benefit as the treatment(s) that are displaced from the healthcare system (Eckermann & Pekarsky, 2014). Consequently, the population as a whole has better health outcomes as a result of the adoption of a new technology.

To estimate the benefits of treatment, often extrapolation and the synthesis of different evidence sources is required – taking cost data from published sources, efficacy data from the relevant trials, and estimations of resource use, a coherent picture can be constructed of the decision problem that no single evidence source could provide (Buxton *et al.*, 1997). The costs and benefits of the new treatment are compared to the outcomes that would be achieved without the new treatment (Paulden, McCabe & Karnon, 2014).

When faced with uncontrolled clinical studies, however, there are issues relating to the fundamental concept of health economics and cost-effectiveness – how effective is the treatment *compared to the next best alternative*? It is this marginal gain that must be estimated to then calculate the incremental cost-effectiveness ratio – the cost for each additional event avoided or unit of outcome gained. This is a more complex question than in medicine in general, where the question most frequently is 'which treatment is best?'. This problem can be illustrated using the example of the effectiveness of parachute use taken from Smith & Pell in Section 1.5.1. Whereas a clinician may be satisfied that a parachute represents an effective treatment, and a regulator may decide that the benefit-risk is positive, to calculate the benefits of parachute use fully, a health economist would not only need estimates of the effectiveness of parachute use but also of the survival rate without a parachute use. That not all people die without a parachute has been established (Hasler, 2010); therefore, an estimation of the mortality rate both with and without a parachute would be needed to calculate the cost per life saved - which could then be compared with other safety interventions (Siegel *et al.*, 1997).

Regulators also do acknowledge the issue that although they may have sufficient evidence to approve a product, this may be insufficient for reimbursement (Jonsson, Martinalbo & Pignatti, 2017). Since beginning this research, a study has also been published demonstrating that reimbursement agencies struggle to interpret such evidence and worry about the risk of bias, with the German reimbursement system only approving products lacking RCTs in exceptional circumstances (Griffiths *et al.*, 2017).

A review of guidance to manufacturers regarding how to approach single arm data shows that no mention is made of this type of data, and as such, there is no best practice. Table 1-1

lists the modelling guidelines for major pharmacoeconomic organisations, and their relevance to uncontrolled studies.

Table 1-1: Relationship of modelling guidelines for economic evaluation to treatments with only uncontrolled study data available

Institution	Guidance	Relevance to uncontrolled studies	Reference
International Society for Pharmacoeconomics and Outcomes Research Society for Medical Decision Making	'While there are undoubtedly topics of interest that do not fit into these 6 papers, it was felt that these would cover the major areas and were at a stage of development appropriate for issuing guidelines.'	Not discussed within the relevant guidelines	(Caro *et al.*, 2012)
National Institute for Health and Care Excellence	'RCTs directly comparing the technology under appraisal with relevant comparators provide the most valid evidence of relative efficacy. However, such evidence may not always be available and may not be sufficient to quantify the effect of treatment over the course of the disease... Any potential bias arising from the design of the studies used in the assessment should be explored and documented.'	No specific guidance is made for the use of uncontrolled studies, only that biases and uncertainty in all types of evaluation should be explored	(NICE, 2013:p.39)
Scottish Medicines Consortium	'analyses should use the best evidence available, be explicit about data limitations and any attempts to overcome these and quantify as fully as possible how the limitations of the data are reflected in the uncertainty in the results of the analysis.'	No explicit mention is made of types of data only that the best available evidence should be used, and data limitations reflected in the submission	(Scottish Medicines Consortium, 2014:p.25)
All Wales Medicines Strategy Group	No specific guidance	No specific guidance is given to manufacturers on the types of evidence, or how this should be used.	(All Wales Medicines Strategy Group, 2013)
Pharmaceutical Benefits Advisory Committee (Australia)	'If direct randomised trials are not available, then an indirect comparison of randomised trials, each including a common reference, or nonrandomised studies could be used to assess the comparative effectiveness of the proposed medicine. The results of these studies should form a basis for translation into a decision analysis to generate an economic evaluation'	No guidance is given regarding observational or uncontrolled studies	(Australian Government Department of Health, 2013:p.35)
Pharmaceutical Management Agency (New Zealand)	'PHARMAC acknowledges that in some cases it may be necessary to use lower levels of evidence if this is all there is available (for example, pharmaceuticals for rare diseases where data may be limited to case studies).'	The use of observational data are not recommended where RCT data are available, how observational data should be used when needed is not stated	(Pharmaceutical Management Agency, 2012:p.24; New Zealand Government, 2015)
Canadian Agency for Drugs and Technologies in Health	'A sound clinical review of the intervention should form the basis of the evaluation… The review may include studies with a variety of designs, reflecting different levels of internal and external validity' 'Even valid justification does not improve the quality of data that has design limitations. A lack of "perfect information" (high-quality data that are needed to fully populate a lifetime horizon model) results in a need for alternative methods in a technology assessment and is accompanied by inherent uncertainty. The results should be interpreted with caution.'	No specific guidance is given regarding treatments with only uncontrolled study data Specific to oncology submissions - the need to interpret observational data with caution is stated; however no methods are prescribed for analysis.	(Canadian Agency for Drugs and Technologies in Health, 2006:p.20) (Canadian Agency for Drugs and Technologies in Health, 2009:p.34)

Although the US healthcare system does not take cost-effectiveness into account formally, the comparative effectiveness of treatments is often estimated. The difficulties in assessing this without controlled data are highlighted in a review by the Agency for Healthcare Research and Quality reviews (Ip *et al.*, 2013). This review found a lack of consistency in the inclusion of uncontrolled studies in guidance (only 21 of 33 reviews included uncontrolled data), with often no reason given for their inclusion, and no consistent methods used in assessing their contribution to estimates of efficacy.

1.7 SUMMARY OF INTRODUCTION AND RESEARCH QUESTION

This introduction discusses the theoretical and historical reasons for the dominance of the randomised controlled trial in drug development, as well as also how uncontrolled studies may be the appropriate study design for a new treatment. It then discusses the issues this raises for interpretation, estimates of comparative effectiveness and thus economic evaluations.

In this section I show that the limitations of uncontrolled trials in terms of the evidence base then available for estimating comparative advantage are well understood. These limitations have been the focus of discussions in both the peer reviewed literature, and by pharmaceutical regulators. The general consensus reached in the literature (and in practice) appears to be that uncontrolled clinical studies, in specific circumstances, can be justified. These circumstances are generally a combination of an immediate risk to life or irreversible harm to patients, where the patient population is small, and where an objective endpoint can be used.

Where relatively little work has been performed, is in how these trials should be interpreted for the estimation of comparative effectiveness, and thus economic modelling and health technology assessment. Here the question is not whether the benefit-risk is positive (the question a regulator faces), but of the gain from therapy *compared to the current standard of care*. This estimated gain can then be used to estimate whether compared to current practice the new treatment is clinically superior, value for money, and what the level of uncertainty is around any such estimates. This is the area I have focussed on as the thesis question:

> *How can (statistical) modelling methods be used to estimate comparative effectiveness where pharmaceuticals have been licensed on the basis of uncontrolled clinical studies?*

To this end, this thesis accomplishes the following:

- Chapter 1

    - Gives context to the issue, summarises the relevant literature, available methods, and the thesis question
    - The work presented in this section is entirely my own; although ideas have been developed in discussions with others, the writing and categorisations are my interpretation.
- Chapter 2
    - Discusses existing modelling methods and their applicability to the area of interest
    - The work summaries presented in this section are entirely my own work, referencing the original papers and applications of methods appropriately.
- Chapter 3
    - Identifies drugs licensed on the basis of uncontrolled data by
        - Individually assessing all EMA and FDA approvals since 1999 for treatments licensed on the basis of uncontrolled clinical trials
        - Understanding the context and evidence on which the approval was based
        - Analysing the type of treatments licensed with only uncontrolled studies as an evidence base
    - Identifies and assesses the methods previously used in modelling of benefit in uncontrolled clinical trials by
        - Searching for published models and health technology appraisals of drugs licensed on the basis of uncontrolled clinical trials
        - Reviewing the identified methods, and categorising them
    - The work presented in this section was designed in conjunction with my supervisors before I performed the literature reviews, and summarising the work. This was then written up with input and debate with my supervisors and co-authors of papers
- Chapter 4
    - Identifies a need for further work on the applicability of Matching Adjusted Indirect Comparison, with a simulation study conducted on its applicability
    - The simulation study on MAIC I designed, getting input from my supervisors, before I coded the study, I then analysed and interpreted the results which I then took to my supervisors for further discussion. The results and scenarios were then refined following extremely helpful comments from peer reviewers.
- Chapter 5

    - Proposes three novel ways in which historical controls can be created to estimate comparative effectiveness, alongside motivating examples
    - The novel methods proposed were all ones that I conceptualised, with two being implemented alongside others, and one (extrapolation from a previous line of therapy) being solely my work
- Chapter 6
    - Gives two examples of using the methods discussed in practice
    - My contribution to each of the analyses was to perform the analysis of uncontrolled study data, which were then incorporated in to economic models by others.
- Chapter 7
    - Summarises the results of the literature searches performed, and the findings of my research
        - A flow chart is presented of the options available to an analyst based on my research
    - This section is my work entirely and represents my summary of the issues, the areas where I believe further research is needed, and how the methods that are available I believe should be used to give the best possible estimates of comparative effectiveness

The main body of the document is 39,070 words, with 245 references and 2 Appendices consisting of 3,013 words. In total the work has directly resulted in nine publications to date, and heavily influenced a further four.

2 EXISTING METHODOLOGIES THAT COULD BE USED TO ESTIMATE EFFECTIVNESS FROM UNCONTROLLED STUDIES

In this chapter I review methodologies that can been used to estimate comparative effectiveness without RCT data, and discuss their suitability for estimating comparative effectiveness of pharmaceuticals where no RCT evidence is available. Where established methodologies are available, a description of the method is given, with published examples of their use summarised.

The approaches available can broadly be separated into categories based on the amount of historical data available, with a further section on emerging methodologies (Section 2.4).

2.1 METHODOLOGIES FOR THE ANALYSIS OF HISTORICAL CONTROLS

2.1.1 METHODOLOGIES FOR USE WITH PUBLISHED HISTORICAL CONTROLS

As discussed in Section 1.4.3, historical controls have frequently been used to estimate outcomes for patients not receiving investigational treatments. The most common approach (which does not attempt to adjust for bias), is naively to compare the outcomes for the new intervention to those seen in the historical controls. Whilst there commonly appears to be a lack of adjustment, this approach contains a variety of strong assumptions i.e. that the data are perfectly exchangeable between studies.

Where this approach is used, evidence from Vickers et al. (2007) is relevant. This study was a review of 70 papers where a historical control was used to power clinical trials. An issue they note when discussing the use of historical controls is where multiple trials (and therefore estimates) are available. Their recommendation is that:

> A single estimate should be derived from the historical data: specifying only a range should be avoided. For instance, take the case where three prior studies had been reported with sample sizes of 1,000, 100, and 20 and response rates of 33%, 22%, and 15%. This is a total of 355 responses in 1,120 patients (32%). It is preferable to give this single historical response rate of 32% than to say only that “response rates in prior studies varied from 15% to 33%”, on the grounds that the latter offers no guidance as to the appropriate null: investigators tempted to pick the middle of the range would underestimate the true response rate and inflate the risk of a false positive. (Vickers, Ballen & Scher, 2007:p.975)

Whilst this position is reasonable (the use of the entirety of the data to generate an estimate of the actual response rate), the use of pooling may also give an unreasonably narrow estimate of the uncertainty – an area where meta-analysis may be able to offer more relevant insight. A related issue is discussed by Thall & Simon (1990) - the incorporation of Phase 2 data to efficacy assessments, where they state historical data should be considered

as a distribution with a point estimate and, not using point estimates alone in decision making. They state:

> Unfortunately, many pilot studies ignore [randomness] and treat the control mean as a known constant. As shown in Tables I and II, this results in an elevated type I error rate. In addition, this approach leads to an inappropriately small sample size and thus a test with inadequately low power. Whereas it may be appropriate to carry out a pilot study with type I error rate 0.10, this should be done by design and not inadvertently. In any case, a control mean computed from historical data has an associated variance which must be taken into account, whether or not the data exhibit inter-study variability. (Thall & Simon, 1990:p.227)

Even with uncertainty in estimates accounted for, this would only include the uncertainty in statistical distributions, and not the structural uncertainty in whether the patients are truly exchangeable. To account for potential differences in patient populations, more sophisticated analyses are required - these are described in the following sections.

The existing methodologies are separated in to three categories; where individual level data (ILD) are available for both intervention and historical control, where ILD are available only for either the intervention or the historical control, and where ILD are not available for either study.

2.1.1.1 WHERE INDIVIDUAL LEVEL DATA ARE AVAILABLE FOR BOTH THE INTERVENTION AND THE HISTORICAL DATA

The methodologies available where ILD are available for both the intervention and comparator are discussed in the NICE Decision Support Unit Technical Support Document 17 (‘The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data’). In the document, each method is explained fully, and an algorithm shown to determine the appropriate method to be used – this algorithm is available in Figures 1 to 3 on pages 37 to 39 of Faria et al. (2015). For this reason, the available methodologies are only summarised in this document.

2.1.1.2 DIFFERENCE IN DIFFERENCES APPROACH

The ‘difference in differences’ approach is a form of natural experiment. The method can be used where a change is made at different time points, for example if a drug was approved in Scotland before England, the differences over the time period between the two countries could be compared.

The advantage of the method is that any background changes in outcomes over time can be controlled for through the use of the control(s), though this is based on the assumption that

there are no exogenous shocks in the time period, and the trends in data would otherwise remain parallel (Dimick & Ryan, 2014).

In the context on pharmaceuticals approved without randomised clinical trials, this approach may be useful where hospital records exist and are accessible for patients who were treated before a trial of a new intervention began. Indeed one of the economic models identified in the literature review took this approach – investigating the outcomes of patients before the trial was set up, in the same centres the trial was conducted (Annemans *et al.*, 2007). Whether sufficient detail would be available however in the patient records (for instance on disease stage, and other inclusion criteria) would be a separate question. It would also be inappropriate to compare all patients with, and without treatment, due to the likely difference between patients who meet the inclusion criteria for studies, versus those treated in general.

As such although the method may be helpful for the area in which I am interested, it would be very specific to data availability.

2.1.1.3 REGRESSION DISCONTINUITY DESIGN

The regression discontinuity design (RDD) is a quasi-experimental approach, that investigates the impact of an intervention around a cut-off on a continuous variable which determines treatment selection. An example would be should a patient be required to be 18 years of age (those under being similar, but untreated), or whether an intervention is required if a blood value falls below a given level. By looking at results either side of a margin (using regression techniques), the effect of the intervention can be observed by assuming the unobservable characteristics of patients either side of the margin are identical. The continuous variable may be deemed ‘sharp’ if it is a strict cut-off, or ‘fuzzy’ if there is a degree of overlap and ambiguity in the group allocation. The way the technique is implemented would be for a preferred regression model to be fit, with a coefficient for when the intervention is received; this is then the estimated effect of the intervention.

The approach of RDD was first proposed in 1960 in education (Thistlethwaite & Campbell, 1960) looking at the impact of merit certificates (and children who just achieved them, and just missed out). Since this time it has been used intermittently in medicine, for example to estimate the impact of having an accountable general practitioner for patients aged over 75 in England (Barker, Lloyd & Steventon, 2016) and impact of ‘smoke free’ legislation on birth outcomes (Bakolis *et al.*, 2016). Recent work has extended the approach to include a Bayesian approach, allowing the specification of prior beliefs (Geneletti *et al.*, 2019, 2015).

It is unlikely that this approach will often be relevant for pharmaceuticals licensed without RCT evidence, as exclusion criteria for clinical trials are usually due to those criteria on which patients are selected being strongly linked to outcomes – for example the exclusion of patients with metastases or heart conditions. Indeed Geneletti et al. highlight that the examples in medical science are in public health, where large datasets with such cutoffs can be used to draw inferences about the effectiveness of interventions. In contrast the diseases where RCTs are not required are comparatively rare (or are at the end stage of common diseases, where few patients remain), so few patients would be either side of a continuous cut-off e.g. 17.5 years versus 18.5 years to be eligible for inclusion in a study. Even if possible to fit, further assumptions would be required to then obtain estimates of effectiveness in the whole population and not just the population on the margin.

2.1.1.4 INSTRUMENTAL VARIABLE ANALYSIS

The instrumental variable approach looks for a factor that is associated with treatment choice, but not with the outcome (apart from its impact on treatment allocation) – this is known as the exclusion restriction. Where a sharp cut-off is applied in treatment allocation, the regression discontinuity design (above) is a form of instrumental variable.

The approach then involves looking at the differences in outcomes seen between groups – as the treatment allocation is not associated with patient characteristics, this should provide an unbiased comparison. In practice, finding a variable that meets this criteria is difficult though examples of the approach could include patients in different jurisdictions or time periods, which in turn are linked to treatment allocation.

Even given such circumstances, this approach may be difficult to operationalise in the context of uncontrolled studies - given the relatively small patient populations and substantial heterogeneity in patient populations. A major issue is also that when conducting a study treatments are not given randomly, but to patients meeting certain criteria who are enrolled for their specific criteria, again making it difficult to see where the approach could be commonly used.

2.1.1.5 PANEL DATA MODELS

Panel data models involve a patients data being tracked over time, and their own historical data used as a form of control; in this sense they would form the type of data needed for the ‘rate ratio’ to be calculated (Glasziou *et al.*, 2007). There are a number of assumptions built in to the approach such as the disease course needing to be modelled – this does appear appropriate for many uncontrolled studies, which are often terminal (Section 1.5.2). However

even beyond terminal diseases the approach has potential to be used. For instance several haemophilia products (most recently emicizumab [Hemlibra®, Roche Products Limited]) have used this form of pre- and post- treatment comparison (Pipe *et al.*, 2019) – fitting a model to the disease before and after treatment may allow for the estimation of comparative effectiveness (subject to a number of limitations).

Although the use of this method use would need to be well planned for pre-treatment data to be collected systematically, there would seem to be the potential to use this approach in some instances as supportive evidence. This is particularly if an intervention is expected to have a dramatic effect that is unlikely to occur naturally – for instance with gene therapy in conditions linked to measurable enzyme levels. As such the approach could be considered, even if there are no examples to date. It should be noted that in being models fit to patient level data, an analyst would need access to the patient data and thus this method could only be used practically by the manufacturer, or by a group working with the manufacturer.

2.1.1.6 THE USE OF PROPENSITY SCORES VIA MATCHING OR WEIGHTING

Propensity scores were proposed in the 1980s (Rosenbaum & Rubin, 1983) as a method for balancing patients between studies using their observable characteristics. By creating comparable groups, a fair comparison can be made accounting for any confounding. Since their initial publication, they have been used extensively in medicine (Shadish, 2013). More simplistic matching methods (for example matching patients according to the type of surgery they had) are also widely used (Cundy *et al.*, 2016).

The propensity score is defined as the conditional probability of receiving an intervention, given all (observed) covariates to the point of receiving treatment. These covariates may include both patient, and disease characteristics – but importantly not outcomes. This is estimated by means of a logistic regression of exposure to the intervention (as a binary variable), given the set of observed covariates. Among patients with the same propensity score, treatment is conditionally independent of the covariates, allowing replacement of the covariates with a single summary value representing the probability of treatment assignment.

To implement the method, each patient in the dataset has their propensity score estimated, after which balanced groups can be created for an unbiased comparison. Mathematically the propensity score was defined by Rosenbaum and Rubin (1983) as $e(x) = \mathrm{pr}\,(z = 1|\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector of vector of covariates, z is an indicator of the treatment received (a variable taking value 1 or 0); for consistency with the source literature the original notation has been kept i.e. pr as opposed to $\Pr$ to denote probability. Balancing on the propensity score, estimates of the counterfactual treatment effect can be made, with (assuming

response r in the notation of Rosenbaum and Rubin) the average treatment effect is then estimable.

To create unbiased comparisons using the propensity score, there is a number of approaches that can be used (and multiple ways in which each can be implemented). Fundamentally however these reduce to two general approaches; matching and weighting.

- In propensity score matching treated patients would be matched with a control patient with a similar propensity score (the allowable difference being termed the 'calliper'). This method may generate a 'fair' comparison by ensuring similar patients are matched, and is particularly useful where datasets may contain patients with a range of severities or even conditions; on the other hand it does mean that a potentially large amount of data would be discarded if matches are not achieved. When using matching, outcomes are then estimated using the matched samples (in this case assuming 1:1 matching, with N patients in each group), working with the notation of Rosenbaum and Rubin as

Equation 3: Difference in response of a treatment, given matching between studies

$$\frac{1}{N}\sum_{i=1}^{N} r_{1i} - \frac{1}{N}\sum_{i=1}^{N} r_{0i}$$

Where r_{1i} and r_{0i} denote the outcomes for the ith treated, and untreated patient

- Alternatively all the data may be weighted by the propensity score (though they may also be trimmed first to account for differences in patients included). When using propensity score weighting, inverse probability of treatment weighting is used on the propensity score, such that the mean score is matched between groups (Ho *et al.*, 2011). Although this approach uses the totality of the data (even if some patients have only a low weight), it should be noted that this may not be a positive; if datasets do include patients with different conditions, their inclusion may introduce bias ('confounding by indication'). Where weighting is used, the estimate of the average treatment effect is therefore estimable (again using the notation of Rosenbaum and Rubin) as

Equation 4: Difference in response of a treatment, given weighting between studies

$$\frac{1}{N}\sum_{i=1}^{N} \frac{r_{1i} z_i}{e_i} - \frac{1}{N}\sum_{i=1}^{N} \frac{r_{0i}(1 - z_i)}{(1 - e_i)}$$

Where z_i represents treatment assignment for patient i

To be successful the approach of propensity scoring relies on the assumption being met that treatment assignment is 'strongly ignorable', conditional on the observed baseline characteristics. This relies on two conditions being met

- Firstly that treatment assignment is independent of outcomes
    - Mathematically in the notation of Rosenbaum & Rubin, $(r_1, r_0) \perp\!\!\!\perp z|\boldsymbol{x}$ where r_1 and r_0 represent the response of treated and untreated patients, z the treatment assignment, and $\boldsymbol{x}$ a vector of all covariates that are used to assign treatments and / or are related to the response
- Secondly that all patients are not guaranteed to receive one treatment or the other i.e. there is overlap between the studies
    - Mathematically $0 < pr(z = 1|\boldsymbol{v}) < 1$ where $\boldsymbol{v}$ represents a vector of covariates

Whilst the number of studies using propensity scoring has increased dramatically in recent years (most likely due to the proliferation of data and software), there are some limitations with the method. First the assumption of treatment assignment being strongly ignorable which may not be the case (and is fundamentally unprovable). Secondly studies have shown that with 'few' patients – typically defined as under 200, propensity scoring may increase the bias seen in comparisons. Finally the groups to be matched (intervention and control) should be as closely matched as possible - for example in location and time, so as to minimise potential bias which is not always the case (Shadish, 2013).

To understand the appropriateness of propensity score based approaches, researchers have sought to recreate published RCTs using observational data. In general although the direction of effects is generally similar with propensity score approaches, the magnitude of effect sizes can often differ markedly (Dahabreh & Kent, 2014). These findings are replicated in a comprehensive simulation study which compared many approaches to the implementation of propensity score based methods. The results of the analysis demonstrated that no one method performed best under all circumstances (Zagar *et al.*, 2017).

In order to provide historical controls, Schmidli et al. (2019) discuss attempts to use propensity score matching to provide a matched control arm for a product in clinical development (though no details are provided). Similarly in the products identified in the systematic review performed in Chapter 3, several (exact number not stated) of the control arms were selected using propensity score based methods (Goring *et al.*, 2019). The methodology therefore would seem relevant for use in selecting patients to form a control

arm, even if no explicit guidance is available for uncontrolled studies where several issues (such as differences between studies, and low patient numbers) are likely to be present.

2.1.1.7 MULTIVARIABLE REGRESSION ADJUSTMENT

Regression analysis is a statistical process for estimating the relationship between a dependent variable, and one (or more) explanatory variables. An simple example of linear regression is shown below, where Y is the dependent variable, $\boldsymbol{\beta}$ a vector of the coefficients attached to the explanatory variables, $\boldsymbol{X}$. The error term is denoted by ε.

Equation 5: Example of linear regression

$$Y = \boldsymbol{\beta X} + \varepsilon$$

Many of the issues and concerns raised regarding historical controls in the literature stem from potential differences in patient populations. In this sense regression adjustment offers the chance to understand the difference each characteristic makes at the margin. Outcomes can then be re-estimated using fitted regression models to predict what may have been seen with different patient groups. For instance if a historical control population had been older with worse performance status, this could be reflected (assuming a good model fit) in predictions of outcomes.

How regression adjustment should be performed is a large area of research in both statistics and econometrics, with questions such as the approach to model selection, appropriateness of model form, and degree of extrapolation beyond the available data being areas where judgment is needed (James *et al.*, 2013). Should a reasonable model fit be possible it may be a viable approach to ameliorate concerns about differences in trial populations. Interestingly whilst propensity scoring appears frequently in the published literature, covariate adjustment appears to perform as well in simulation studies (Elze *et al.*, 2017; Zagar *et al.*, 2017).

2.1.2 WHERE INDIVIDUAL LEVEL DATA (ILD) ARE AVAILABLE FOR EITHER THE INTERVENTION OR THE HISTORICAL DATA

In pharmaceuticals, it is more common that an investigator will have access to ILD for one treatment (as the company who have developed the product will have conducted the clinical trial), but not the comparator data as this will either have been developed by another company, or will be taken from historical literature. In these cases only Aggregate Level Data (ALD) are typically available – limited data on baseline characteristics from 'Table 1' in publications (which seldom report the same characteristics), and deidentified outcomes data

– at best Kaplan-Meier or 'swimmer' plots (which plot the duration of responses by each individual), and at worst only summary statistics such as mean or medians.

The lack of access to ILD from both studies necessitates different methods for performing analyses, which are described below and can be used to produce consistent estimates between studies - avoiding the potential biases of naïve comparisons.

2.1.2.1 MATCHING ADJUSTED INDIRECT COMPARISONS

In using naïve comparisons to historical data (as discussed in the introduction) the assumption is that patients are perfectly exchangeable between studies. This issue was noted by Vickers et al. in a review of 70 Phase 2 studies using historical controls, who state

> Noteworthy was that not a single study in our analysis incorporated any statistical method to account for the possibility of sampling error or for differences in case mix between the phase II sample and the historical cohort (Vickers, Ballen & Scher, 2007:p.974)

A similar issue is raised by Mazumdar et al., (2001), who suggest reweighting patient data to account for differences between studies. They use the examples of bladder cancer and melanoma, and state

> The method also relies on having well-established risk factors for a particular disease. Although agreement on all prognostic factors for all disease systems is too optimistic, there is agreement on the most important prognostic factors for many of them. For example, risk factors for melanoma, bladder, breast, renal cancers and germ cell tumours are generally concurred upon. If there are relatively few prognostic factors, one could create risk categories based on their joint distribution. However, one must be aware that too many categories could lead to sparse data (Mazumdar, Fazzari & Panageas, 2001:p.891)

The method of Matched Adjusted Indirect Comparison (MAIC) is designed to address these caseload differences between studies, where individual patient data are only available for one study.

The approach suggested by Signorovitch et al. (2010, 2011) uses a propensity score weighting-like method to balance patient characteristics between studies. Assuming one study for which individual level data are available, and one for which they are not, Signorovitch et al. (2010) denote the patient characteristics and outcomes of the individual i patients in the study of the treatment (for which individual level data are available) as $\boldsymbol{x_i^0}$ and Y_i^0.and for the study where only aggregate characteristics are available, mean patient characteristics and outcomes as are used, denoted as $\overline{\boldsymbol{X^1}}$ and $\bar{Y}^{\boldsymbol{1}}$. It should be highlighted that the notation of Signorovitch et al. is the opposite of Rosenbaum and Rubin for treatment $(t = 0)$ and control $(t = 1)$. For consistency with the original papers, the relevant notation (in this case by Signorovitch et al) has been kept.

To implement the approach, a weight (w_i) is calculated for each patient in the individual data, such that the overall mean of the weighted individual data $\overline{\boldsymbol{X^0}}$ (which is calculable), matches that of the aggregate data ($\overline{\boldsymbol{X^1}}$). This weight for is thus the odds that a patient received treatment and not control and is defined in the notation of Signorovitch et al. as

Equation 6: Calculation of weights for Matching Adjusted Indirect Comparison

$$w_i = \frac{Pr(t_i = 1|\boldsymbol{x}_i)}{Pr(t_i = 0|\boldsymbol{x}_i)}$$

To w_i the difference between the reweighted individual level data and $\overline{\boldsymbol{X^1}}$ is minimised; a full mathematical proof is presented in the supplementary materials to Signorovitch et al. (2010) Equation A1, where this has been set to zero. Using the resulting weights it is then possible to estimate the reweighted outcomes of the study in a similar patient group that that were the outcomes $\bar{Y}^{\boldsymbol{1}}$ of the control arm were obtained:

Equation 7: Reweighting of outcomes to match and index study after calculation of weights from Matching Adjusted Indirect Comparison

$$\frac{\sum_{i=1}^{n} y_i(1 - t_i)w_i}{\sum_{i=1}^{n}(1 - t_i)w_i}$$

Implicit in the method is the assumption that the groups are similar with a reasonable level of overlap; the paper does suggest data be trimmed to ensure that patients would always have a chance of appearing in the other trial i.e. the same inclusion and exclusion criteria. This does however assume that the trial for which individual level data are available has broader (or at the very least, similar) inclusion criteria as the trial for which only aggregate data are available; this will not always be the case.

Although the method is fairly novel (<10 years since first publication), as of December 2019 there were 126 hits in PubMed, with around a half of these being applications of the method. As such it may represent a methodology which allows the comparison of trials, where patient level data are not available for both studies. At the point this research was started however, the accuracy of the method was unproven, with some conflicting reports presented and no established best practice to conduct MAIC (Shafrin *et al.*, 2017).

2.1.2.2 SIMULATED TREATMENT COMPARISONS

Simulated treatment comparisons were proposed by Caro and Ishak (Caro & Ishak, 2010), and involve setting predictive equations to estimate outcomes using the available individual level data i.e. a regression model, before re-estimating outcomes in a different population – that of the aggregate data. This differs from MAIC as whilst MAIC attempts to reweight

patients to match the aggregate data (from where outcomes can be recalculated), STC attempts perform regression modelling and thus predict outcomes.

The challenges involved in STC are similar to those with regression models (Section 2.1.1.7); what type of regression model should be used? And variables should be included? Few clinical trials are likely to be amenable to modelling using a linear regression, and thus more complex regression forms (such as survival models) are likely required. Where this is the case, not only is there a challenge in selecting the appropriate model, but there is also a potential complexity in implementation for non-linear models, with calibration required when predicting in different populations to that in which the outcome model was fitted (Ishak, Proskorovsky & Benedict, 2015).

To deal with the calibration issue, the authors propose an alternative of deriving a predictive equation, and then simulating patient profiles from the data available (taking in to account correlation between variables) such that the average of simulated patient profiles matches the aggregate of the aggregate data. This latter approach represents the novelty of STC, but adds further assumptions, (which again are unverifiable) regarding the correlation structures being similar between datasets.

As a method, STC would appear to be reasonable to apply, and (unsurprisingly) well suited to use with historical controls. The application of the method (for instance model and variable selection) however would be key. It should also be noted that only one example of STC has been identified in the literature (outside of the original concept by Caro et al.) (Phillippo *et al.*, 2019) and there exists no standardised code implementation. As a concept however the method appears promising, with how it should be implemented being an area for future research.

2.1.2.3 PREDICTION OF OUTCOMES USING SURROGATE ENDPOINTS

A potential approach for estimating comparative outcomes across studies is the use of surrogate endpoints. If an intermediate endpoint in trials (ideally one that is measured objectively), can be linked to patient relevant outcomes - for example overall survival, this may be used for the estimation of benefit from a given timepoint. Such an approach means any differences from comparator trials in characteristics which may predict achieving the outcome can more readily be assessed, without the need to extrapolate over time.

For a surrogate to have validity in such circumstances, it must have biological plausibility, and demonstrate good prognostic value in estimating the outcome of interest. The relationship between surrogate and outcome must also be causal, and be the same

regardless of treatment i.e. a given level of response should lead to the same outcome, irrespective of how that response was achieved. Equally there is a need for the measure used to be comparable across trials. A comprehensive history of the theory and use of surrogate outcomes is given by Buyse et al. (2016).

The use of a surrogate approach could also be coupled with other approaches, for instance propensity score based approaches to create unbiased comparisons of an outcome like response rate, which could then be used in a surrogate outcomes based framework.

2.1.3 WHERE INDIVIDUAL LEVEL DATA (ILD) ARE NOT AVAILABLE TO A RESEARCHER FOR THE INTERVENTION OR THE HISTORICAL DATA

Whilst not a statistical method, where no access to ILD is available, narrative conclusions may still be helpful in describing the likely impact of any differences between studies (even where the magnitude cannot be quantified). This can include highlighting the likely sources of uncertainty, the magnitude of any apparent benefit, and any potential biases in the analyses.

This wider understanding is important as studies are not conducted in a void; differences in between diseases (and understanding of diseases) may mean the same apparent effect size is interpreted differently – because of the context in which the study was conducted. Expert input to highlight likely biases can then inform decision making, even if formal methods are unable to account for such differences.

2.2 METHODOLOGIES FOR THE SYNTHESIS OF MULTIPLE HISTORICAL CONTROLS

Where multiple historical controls are available, as well as adjusting for any differences between studies, there may also be a desire to synthesize (in some form) the totality of information. This applies both to naïve estimates, but also once adjusted estimates. Where this is the case several methods are available.

2.2.1 META-REGRESSION

As discussed in Section 1.4.3, historical controls have frequently been used to estimate outcomes for patients not receiving investigational treatments, and are commonly used in the estimation of comparative effectiveness of pharmaceuticals. It is also possible that a researcher will not have access to ILD for either of the interventions, and may only have ALD for both; this would be typical of an academic researcher or a payer. Whilst this restricts the

analyses that can be performed, there are methods which allow credible estimates of effectiveness to be generated using all available data.

Where multiple studies are available, the approach of meta-regression investigates the effect of study level covariates on outcomes; one covariate of which could include treatment assignment. The Cochrane handbook suggests that it should not be attempted with fewer than 10 studies (Fellow & Director, 2008) though more recent work on regression in general has shown fewer observations may still result in accurate, unbiased estimates (Austin & Steyerberg, 2015).

The most general form of meta-regression would be a linear regression though (should the number of studies allow) more complex specifications are available including random effects models. Other factors to consider are the within trial variances of treatment effects, and the heterogeneity in studies not explained by the regression (Thompson & Higgins, 2002). Further issues may also be caused by the limited reporting of potentially important covariates, either as they were not reported in primary publications, or their importance was not understood at the time the study was conducted (for example disease markers that were not recognised, or measurable).

Whilst there are potential issues in conducting meta-regressions, they may be helpful to understand the totality of data where a large number of studies exist. For example a meta-regression of 28 studies of diabetes patient education demonstrated a positive impact on of intervention, with face-to-face delivery, cognitive reframing, and exercise content able to account for 44% of the variance in study results (Ellis *et al.*, 2004). Although likely of limited utility to uncontrolled studies given the requirement for a large number of studies, NICE DSU TSD3 discusses the topic of meta-regression (Dias *et al.*, 2011).

2.2.2 META-ANALYSIS OF HISTORICAL CONTROLS

Meta-analysis is a technique to combine multiple studies, assigning weights to each study which when aggregated produce a single summary statistic for the effect size across the included studies. If two studies were of equal size, with equal results, they would be given an equal weighting, equivalent to a simple mean. In reality studies will vary in their precision (one over the variance), with different sample sizes and standard deviations, leading to different weights being assigned – the two most common approaches being fixed effects and random effects meta-analysis.

In fixed effects meta-analysis, it is assumed there is a common effect size (μ) shared by all the studies, in addition to an error term (ε) which may vary between studies. A random

effects model allows the effect of treatment to vary between studies (for example the effect size may be larger in sicker patients). The studies included in the meta-analysis are each assumed therefore to draw from the distribution of the effect size.(Borenstein, Hedges & Rothstein, 2007).

Where data can reasonably expected to be similar to previous trials, for example with similar trials conducted in a short space of time at a single unit, it has been suggested that rather than using only one trial, or pooled results of trials, the technique of meta-analysis can be used to give a more accurate prediction of the expected response rate were a control arm included in a study. Previous uses for this approach have been in Phase 2 studies, with either a frequentist approach (where trials are simply meta-analysed) or a Bayesian form (incorporating a prior beliefs about the efficacy of treatment). When this is used as prior data, it can then be referred to as the Meta-Analytic Predictive prior (Schmidli, Wandel & Neuenschwander, 2012, 2012).

2.2.3 THE BAYESIAN 'POWER PRIOR'

A limitation with the use of a historical control, even if trials are meta-analysed together (to take in to account differences between trials), is that it is assumed that the patients in the historical study are exchangeable with the ones in the present study (Viele *et al.*, 2014). Equally when historical data are not considered in a comparison, the implicit assumption is that the historical data are of no value.

As an attempt to bridge these positions, the power prior was conceptualised by Ibrahim and Chen (Ibrahim & Chen, 2000), which '… raises the likelihood of the historical data to the power parameter a_0 which quantifies the discounting of the historical data due to heterogeneity between trials' (Neuenschwander, Branson & Spiegelhalter, 2009:p.3652). In this way historical data can be 'borrowed' to supplement a current trial and reduce the number of controls required. A modified version was developed by Duan (2005), and Neuenschwander, Branson and Spiegelhalter (2009) which can be set to apply different approaches – for instances pooling where data are similar, and discarding where dissimilar.

The use of power priors has been an active area of research, with extensive theoretical work performed, and subsequently some applied examples emerging. One key area of difficulty appears in setting the level of weight attached to the prior data, with no clear consensus. Where this weight is high, the new data has little influence, whereas with a low weight, historical studies are effectively discarded.

In terms of applied examples, one area the idea seems promising is in bridging studies, where an intervention is demonstrated to work beyond the population in which it had originally been studied. Ollier et al. (2019) use this approach to 'borrow' strength from the original trials in studies demonstrating effectiveness in Asian (compared to Western) patients to reduce sample size. A more complex example is available in the form of 'efficient platform designs' which are trials designed to enrol patients in a randomised fashion - should the controls match the historical data, borrowing from historical data is implemented, skewing randomisation to the novel intervention. This has the effect of minimising exposure to placebo, and maximising statistical power (Normington *et al.*, 2019).

To set the level of discounting of historical data, two approaches have been proposed. Firstly a constant, set for example through expert opinion by which the evidence is down-weighted. This can even be adjusted for perceived bias (Turner *et al.*, 2009) – this approach is best suited to the original power prior, as opposed to the modified form. The second approach is to calculate a parameter by looking at the level of heterogeneity between studies included in the analysis – this can be used with either form of the power prior, but allows more complexity with the modified version (for instance heavily penalising discordant studies).

Initial uses of the power prior seem to have been in early clinical development to understand whether data from a historical control can be used to augment the control patients in early clinical studies, before large confirmatory studies are commenced (Strimenopoulou & Walley, 2014; Hobbs, Sargent & Carlin, 2012; Gsteiger *et al.*, 2013; Dejardin, van Rosmalen & Lesaffre, 2014). Other applied examples are from Dron et al. (2019) who look at minimising the number of patients needed in trials by borrowing (using data from Project Data Sphere) with four applied examples.

Later work however adds cation to the approach, with work by Schoenfeld et al. (2019) demonstrating that unless effect sizes are indeed large (a greater than 30% difference in response rate), there are no sample size savings in clinical trials using borrowing due to the heterogeneity in trials. Similar finds are seen in Lewis et al. (2019) who use colorectal cancer data with simulation studies to show that the 'drift' in outcomes over time can confound results.

The approach however may offer a more sophisticated option than the simple acceptance or rejection of historical data, but giving less weight to the prior evidence where there are large differences in outcomes within the evidence base.

2.2.4 COMPARISON BETWEEN APPROACHES FOR COMBING HISTORICAL CONTROLS, AND APPLICABILITY TO UNCONTROLLED STUDIES

As the approaches of meta-analysis and the power prior are relatively new, there exists limited comparisons between the methods. Work from Isogawa et al. (2019) shows that the preferred approach can change depending on the structure of data (with no rules that can be determined). The approaches that have been attempted are also documented by Lim et al. (2018), who highlight where the 'stringent' Pocock criteria may be able to be relaxed, mainly in areas of high unmet need e.g. terminal diseases.

Although the two approaches (meta-analysis and power prior) for combining multiple studies both have relevance for uncontrolled studies, some modifications are needed as the primary objective of both has been to supplement contemporary controls – not act as a replacement. In this sense, there exists no clear path for uncontrolled studies; the concerns regarding heterogeneity of outcomes and indeed existence of the power prior to down-weight historical data (rather than simply pool) implicitly assumes that differences likely exist between studies.

The approaches may be useful however to meta-analyse (including any down-weighting brought in through the power prior) where multiple historical controls are available. How this down-weighting should be performed however is unclear, and will at this stage therefore be somewhat (unavoidably) arbitrary.

2.3 WHERE NO HISTORICAL DATA ARE AVAILABLE

Where no historical data are available for comparison, two methods have been used to estimate what the outcomes would have been were a control arm available.

2.3.1 THE USE OF EXPERT OPINION

Whilst expert opinion appears historically to have been done in an unstructured form, methods have been developed for the use of gathering such evidence robustly. The most widely cited of these is the Sheffield Elicitation Framework (SHELF), which consists of various methods implemented with a user guide and R package (Gosling, 2018). Whilst lower on the 'hierarchy of evidence' than trial data, the use of expert opinion may be required where data simply does not exist. An important distinction here is between expert evidence i.e. eliciting evidence from experts, and expert opinion, where the expert is then extrapolating based on their experience to project what may occur.

Though the SHELF frameworks acts as an ideal form, there exist several published practical demonstrations of structured elicitation from experts. To do this Sperber et al. (2013) adapted the SHELF method (specifically the quartile method) to work in the more realistic setting of geographically dispersed experts, whilst Grigore et al. (2016) tested two of the methods from SHELF (the quartile method, and the histogram method), and how they might apply to different problems. In a slightly different context Dallow et al. (2018) give examples of how formal priors can be elicited (again using the SHELF framework as a guide) in the drug development process – work that could relatively easily be adapted for later stage products.

2.3.2 THRESHOLD ANALYSIS AND THE 'E-VALUE'

Threshold analysis is used to distinguish where the point at which a model ceases to behave in a particular way – for instance how expensive a treatment would need to be before it was deemed not cost-effective. An example can be seen in the analysis by Tappenden et al. (2006), where the relative risk needed for a treatment to be cost effective was estimated, allowing decision makers to consider whether or not this threshold was likely to be reached.

Work from the GetReal collaborative (an EU funded project including members from the EMA to develop the use of observational data in medical decision making) has also proposed a similar concept; 'Threshold-crossing'. The method proposes a series of steps to be undertaken before a study is conducted, to ascertain the degree of heterogeneity in the disease area such that should this threshold for a 'clear' demonstration of efficacy (and without serious safety signals) then the product is deemed to be effective. If the threshold is not met then equipoise is still present i.e. it is not clear the intervention is effective, legitimising a RCT. The approach seems rational, and would ensure in depth analysis on the suitability of control data - and indeed whether historical control data are available - prior to the conduct of an uncontrolled study. In taking a systematic approach to evidence generation the approach seems reasonable, even if it would not directly allow a quantification of the amount of benefit. Interestingly in the paper they suggest that a Bayesian framework is more natural for such analysis (citing work on the power prior), but that they could not identify any successful uses of the approach to date.

A similar approach has been proposed for use in observational studies, termed the 'E-value' (VanderWeele & Ding, 2017). The definition proposed is 'the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and outcome, conditional on the measured covariates, to fully explain away a specific treatment–outcome association'. The higher the E-value, the less likely the

effect seen is a result of unmeasured confounders. This is mathematically defined below, with Table 2 of the paper giving formulations for other common endpoints (for instance hazard ratios).

$$E - value = Relative\ Risk + \sqrt[2]{Relative\ Risk\ x\ (Relative\ Risk - 1)}$$

Whilst proposed for observational studies, the use of E-values could logically be extended to uncontrolled studies, whether estimates are derived from historical controls, or from other methods. The authors propose the calculation of several E-values for any given analysis

- a simple E-value between the point estimates of outcomes
- one based on a covariate adjusted model
- one based on the minimum E-value needed to reach the limit of the 95% confidence interval for outcomes.

Although no firm conclusions can be drawn regarding the absolute size of the E-value, it may be helpful in informing decision making by quantifying the strength of evidence for the intervention (and providing reassurance where results are unlikely to be due to confounding). Given patient numbers are often small in uncontrolled studies this is particularly important (as a large effect size may be highly uncertain). Alongside the provision of E-values for the comparison in question, communication of their meaning also also be helped by provision of E-values from other interventions in the disease area.

2.4 EMERGING METHODOLOGIES

The above sections consider the methods which have been established for use in medicine, however there exist a number of other approaches which are in wide in other disciplines (such as computer science), and may be applied to medicine to estimate comparative advantage.

2.4.1 THE USE OF ‘REAL WORLD DATA’ TO ESTABLISH CONTROL ARMS

Although all data are from the ‘real world’, the term ‘real world data’ (RWD) is often used to describe data collected as a part of routine clinical practice, as opposed to collected for a specific purpose such as a trial, registry, or case series (Berger *et al.*, 2017). With the growth in data availability and data science as a discipline - added to the increased use of Electronic Medical Records (EMR) there is the potential for control arms to be identified using these databases, as opposed to being taken from clinical studies. Whilst attempts to replace RCTs entirely are likely to face fierce resistance (Gerstein, McMurray & Holman, 2019), they may be able to be used to quantify outcomes and create historical controls.

Although work in the area is still relatively early, publications (largely from Flatiron Health, now purchased by Roche) are beginning to be able to estimate control group outcomes (Carrigan *et al.*, 2019). An example of the approach is a paper estimating the comparative effectiveness of alectinib in lung cancer, using EMR data as a control group – this approach was required as the drug was licensed on the basis of an uncontrolled study (Davies *et al.*, 2018).

Other work is also ongoing on a similar theme from the GetReal collaborative - here 'workpackage 1' has the aim to 'develop a framework for the acceptability of real world evidence for estimating the effectiveness of new medicines' (Egger *et al.*, 2016).

Whilst the approaches used are still in their early stages (and yet to be fully validated), in the absence of other data, they may be able to assist decision makers to understand potential outcomes. As such approaches have only recently become possible, it appears policies towards the use of such data are yet to be fully defined (Makady *et al.*, 2017), and it remains to be seen how the complex models underlying the derivation of the datasets will be received by regulators and payers.

The idea that RCTs may be able to be replaced by observational data is also not a new one; in his 1980 paper 'Why Data Bases Should Not Replace Randomised Clinical Trials', Byar cites a 1972 book by Cochrane, which in turn states 'Observational evidence is clearly better than opinion, but it is thoroughly unsatisfactory. All research on the effectiveness of therapy was in this unfortunate state until the early 1950's' (Byar, 1980). It therefore appears that although the approach of using EMR data is promising, there are recognised issues which have yet to be solved, and repeated concern over the years about such an approach.

2.4.2 MACHINE LEARNING

Machine learning is a topical subject and active area of research across many fields. The concept being to allow algorithms to explore data; testing a wide array of models and finding associations that may not have been apparent to humans. These models have seen rapid uptake by internet advertising giants, where large datasets are available to obtain insight (James *et al.*, 2013).

In terms of medicine, there exist few practical examples, though this may change in the coming years. The most recent work published (McConnell & Lindner, 2019) shows the method has promise in estimating treatment effect sizes, though at present is unlikely to be practical when used in the context of uncontrolled studies; the sample size used in the McConnell and Lindner paper was 5000 patients (1000 was the smallest sample - where

performance was poor). In practical terms this therefore implies work will be needed to establish the preferred methods in related areas, before they are applied uniformly.

2.5 SUMMARY OF EXISTING METHODOLOGIES FOR ESTIMATING COMPARATIVE EFFECTIVENESS USING UNCONTROLLED STUDIES

In this section I show that there exist a variety of methods for the estimation of comparative effectiveness. In this section I classify them according to their purpose (analysis of pair of trials, versus creating estimates from a number of studies), and required access to individual level data.

Over the time period I have been performing my research (with reading beginning in 2013) the most active area appears to have been that around the use of the power prior. In this area a number of papers and approaches have appeared (Isogawa *et al.*, 2019; Banbeta *et al.*, 2019; van Rosmalen *et al.*, 2018; Ibrahim *et al.*, 2015; Nikolakopoulos, Tweel & Roes, 2018) where the method has been applied to different settings, and with different aims. The other growth that appears to have happened is in the uptake of MAIC, which was first published in the time before my research began, but by the end was in widespread use (with NICE DSU guidance). Despite these changes, over the time period there do not appear to have been substantial changes in the publicly stated willingness of regulators or payers to receive uncontrolled study data, or preferred methods of analysis.

Although a number of methods are available, there are notable limits to existing knowledge. This applies to areas such as MAIC where there is uncertainty where the method is appropriate, to areas such as the power prior, where (despite developments) is no framework to determine for the degree of down weighting to apply to each study. There is also scope for new methods for the creation of historical control data, as evidenced by attempts to use RWD and machine learning techniques to create synthetic control arms.

3 IDENTIFICATION OF THE NUMBER OF TREATMENTS INVOLVED, AND METHODS USED FOR MODELLING

In order to understand the scale of the issue, I conducted reviews to establish how many treatments were licensed on the basis of uncontrolled studies, and how their comparative effectiveness had been modelled. The aim of this was to establish the disease areas where uncontrolled studies are frequently used, and the type of data (and approaches) used in the analysis of the data – both for regulators and for payers. This chapter describes these literature searches.

3.1 TREATMENTS APPROVED ON THE BASIS OF UNCONTROLLED CLINICAL STUDIES

A summary of the identification of treatments licensed on the basis of uncontrolled studies was published in BMJ Open (Hatswell *et al.*, 2016)*, with a full details of each treatment published as a UCL research report* (Hatswell, Baio & Freemantle, 2017)

The first stage of this review was to identify treatments licensed on the basis of uncontrolled clinical studies. This was done for the EU using the EMA website, and the US using the FDA website.

As this study relates to how modelled estimates of efficacy have been constructed, the search was limited to licenses granted since 1999, which also coincides with when NICE in the UK first began to appraise the cost-effectiveness of medicines – requiring the estimation of comparative effectiveness. The end date of the literature search was 7 May 2014, when this research began.

3.1.1 REGULATORY PROCESSES IN THE UNITED STATES AND THE EUROPEAN UNION

The regulatory approval process in the US and Europe is slightly different. In the US the role of the FDA dates back to the 19th century (Food and Drug Administration, 2014), with companies engaging with the FDA before the submission of a New Drug Application, which if approved, allows the manufacturer to promote and sell the drug in the US (Lipsky & Sharp, 2001). Of particular relevance to this study is the accelerated approval process, in which there is no set process, but where the FDA are willing to approve products on the basis of surrogate outcomes pending confirmatory trials (Ciociola *et al.*, 2014; Senderowicz & Pfaff,

2014). As a whole, however, the FDA require ‘substantial’ evidence from ‘adequate and well-controlled’ studies (Chow & Chang, 2019).

In Europe the situation is more complex – until recently each country was responsible for its own decisions on drug availability. In 1995 the EMA was formed, with a key part of their role being the administration of the ‘centralised authorisation procedure’ (Jefferys & Jones, 1995). Any product approved under this programme is given a marketing authorisation valid in all EU countries, as well as Iceland, Norway and Liechtenstein – this approval route is mandatory for new biotechnologies, orphan medicines, and treatments for cancer, HIV/AIDS, diabetes, and other high profile / burden diseases (Netzer, 2006). The UK have now withdrawn from the EMA, though this did not occur until after the period the review covers.

The alternative (and precursor) to the central authorisation procedure of the EMA is to gain approval in one nation state, after which companies may apply for mutual recognition, where that approval is converted to a Europe-wide approval. Should any individual regulatory body object to this approval, the dispute can then be taken to the EMA (Powell, 2000; Miguel *et al.*, 2014). For this reason it is plausible that a product will be available within the EU but not approved by the EMA – either by pre-dating the centralised procedure or by falling outside the centralised procedure and being approved via mutual recognition. This underlines the importance of including FDA approvals in this literature review so as to identify as many possible drugs (and therefore economic models) that meet the criteria as possible.

3.1.2 DETAILS OF THE SEARCH OF THE EMA AND FDA DRUG APPROVAL DATABASES

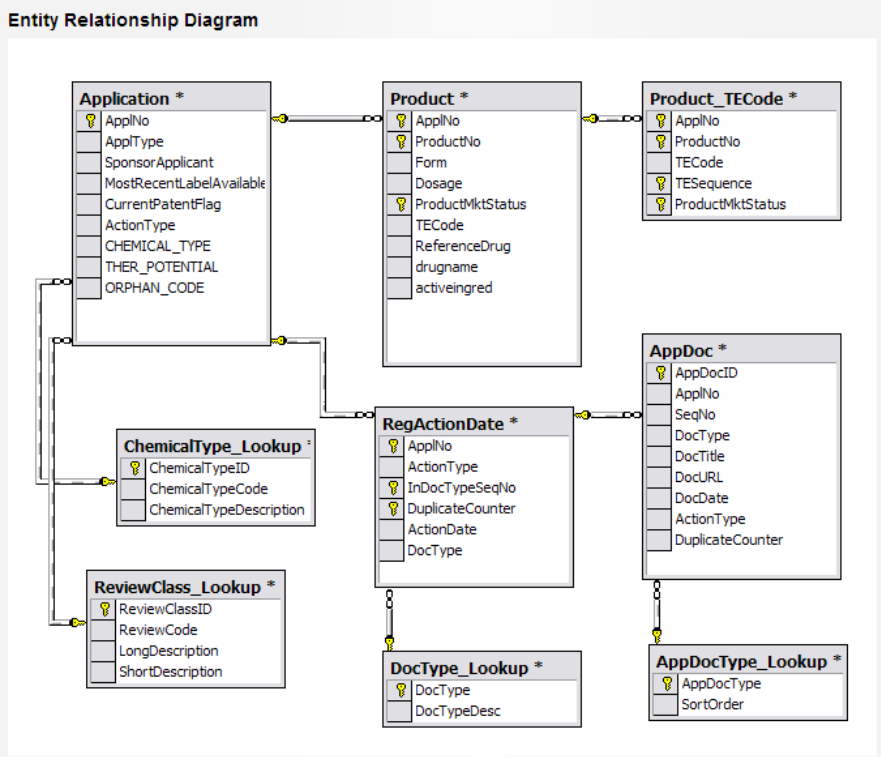
To identify treatments licensed on the basis of uncontrolled clinical studies, all treatments licensed via the EMA centralised procedure and all label approvals by the FDA since 1999 were reviewed. Full details of the inclusion and exclusion criteria (as well as search terms) are given in Appendix A, with details of the methods and search results described below.

In addition to data on approvals, also extracted was whether the treatments were approved in an existing indication on the basis of RCT data (and applying for a license extension with uncontrolled data), and whether applications were for treatments with an existing RCT based approval.

3.1.2.1 FOOD AND DRUG ADMINISTRATION

To identify drugs in the US, the database ‘‘Drugs @ FDA’ was downloaded on 8 May 2014 from the FDA website. This is a relational database, the structure of which is shown in Figure 3-1.

Figure 3-1: Drugs@FDA database structure, taken from http://www.fda.gov/Drugs/InformationOnDrugs/ucm079750.htm on 8 May 2014



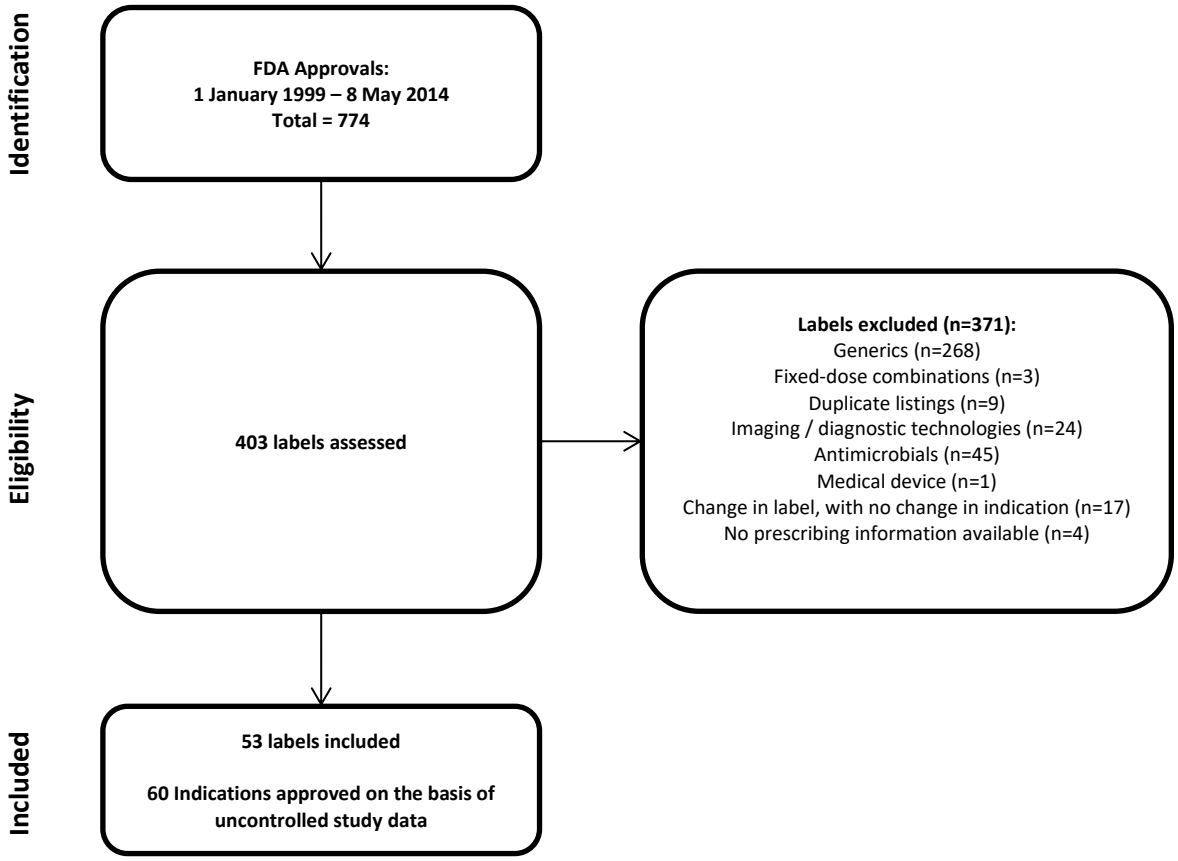
Once the database tables had been downloaded, they were imported into Microsoft Excel 2010 and linked with the use of the *vlookup* function. The list of drug approvals was then used to identify the treatments licensed on the basis of uncontrolled study data. The structure of this database is different to that of the EMA, in that it is arranged around label approvals and not around the drugs themselves. Consequently there are multiple entries for the majority of drugs.

For each approval the relevant documents were accessed and downloaded. The evidence for each indication was reviewed in turn for the evidence included in the approval. The documents most frequently containing this information were the FDA label, or the clinical review of the New Drug Application.

From the 774 listed labels approved since 1999 by the FDA, 403 were deemed relevant, with 53 including at least one indication approved on the basis of uncontrolled study data. This

left a total of 60 indications approved on this basis, as shown in the PRISMA diagram in Figure 3-2 – some approvals were for more than one indication.

Figure 3-2: PRISMA diagram of drug indications approved on the basis of single arm trials by the Food and Drug Administration



3.1.2.2 EUROPEAN MEDICINES AGENCY

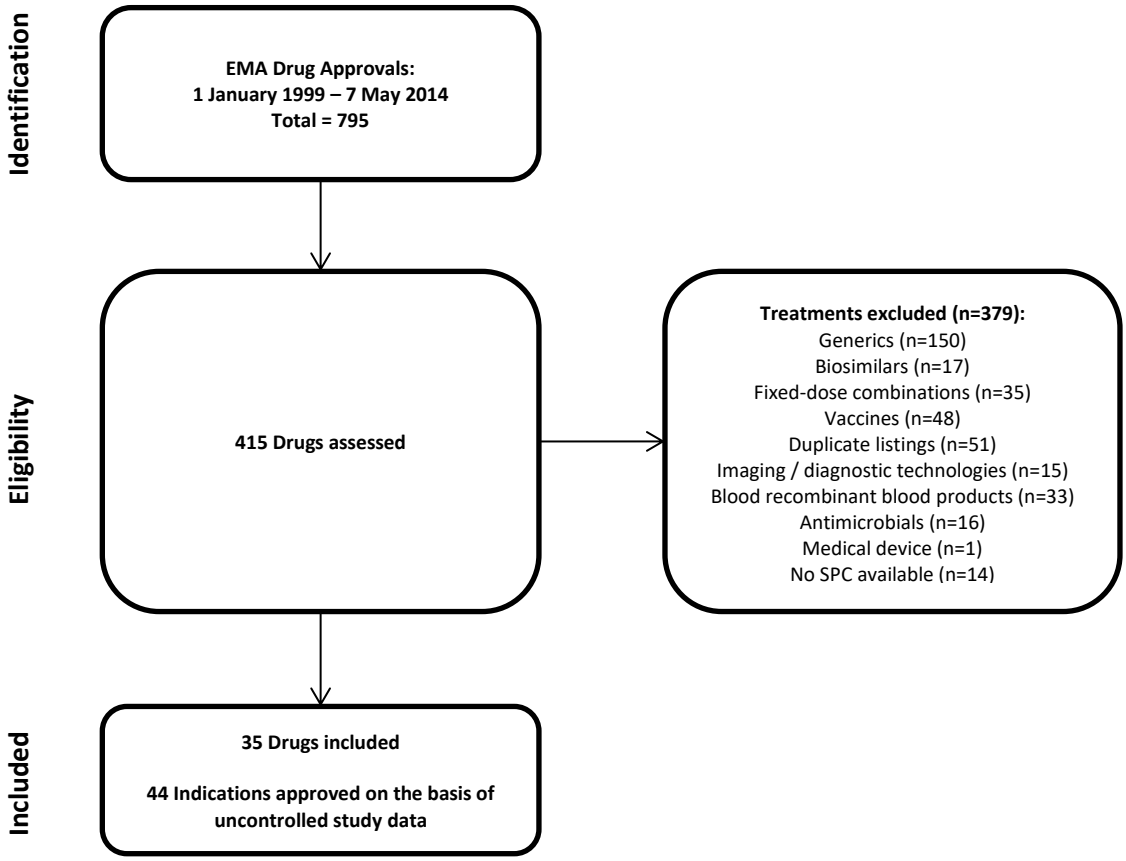
To identify drugs approved in the EU, all treatments listed as on the EMA website as being given a central marketing authorisation were downloaded, and imported into a Microsoft Excel spreadsheet. This was completed on 7 May 2014.

To identify relevant approvals, the EMA Summary of Product Characteristics (SPC) was searched for the main evidence in support of the product (generally found in Section 5.1 – Clinical efficacy and safety, or alternatively in the scientific discussion document). The documents most frequently containing this information were the EMA license (in Section 5.1), or the Scientific Discussion, a document produced by the EMA detailing the evidence for an application, and the reasons a decision was made. Only those products that met the

inclusion criteria and that were licensed exclusively on the basis of uncontrolled data were selected for extraction.

Reviewing drug approvals led to the review of 795 drug approvals, of which 415 were deemed relevant drugs. This led to 37 drugs being identified as having been licensed in 44 indications on the basis of uncontrolled studies, as shown in the PRISMA diagram in Figure 3-3.

Figure 3-3: PRISMA diagram of drug indications approved on the basis of since arm trials by the European Medicines Agency



3.1.2.2.1 ERRORS IN THE EMA DRUG DATABASE

Whilst searching the EMA drug approval databases, I noticed an error on the Summary of Medicinal Product Characteristics (SmPC) for lipefilgrastim (Lonquex®) (European Medicines Agency, 2014b). The primary endpoint of many of the studies, 'DSN', was not

defined at any point. With further research on the primary studies, I believe this was intended to be defined as ‘Duration of Severe Neutropenia’.

Based on this I contacted the EMA, who have confirmed this is an error, which was rectified on the next update of the license (European Medicines Agency, 2014a).

3.1.2.3 HOW THE EVIDENCE WAS ASSESSED

In order to establish the evidence used, the first item checked was the product label, where the evidence informing efficacy is typically stated (the relevant sections are noted under the headings for each agency).

Where no randomised data was included in the product label, the original submission was consulted. For FDA approvals (which have an exceptional level of transparency) the clinical review for the product was reviewed – this included all data reviewed for the approval. For EMA approvals after each decision (positive or negative) a public ‘Assessment Report’ is produced. These documents (approximately 100 pages long), detail the disease background, evidence provided, and justification of the decision made by the regulator. If no RCT data with an active or placebo control was available (even in early stages), the product was deemed to be approved based on uncontrolled study data.

After classification of each product, the dates of submission, decisions, and whether the approval was a first approval for the drug (or a follow on indication) were documented in a Microsoft Excel spreadsheet, along with any reasons for exclusion. After identification, this spreadsheet was then able to be used to not only count approvals, but also compare between agencies.

3.1.2.4 FURTHER LITERATURE SEARCHING

After the initial filtering of the FDA website, I searched for each of the drugs approved by the EMA but not by the FDA within the dates of my search, and vice versa. This was with the objective of identifying drugs that had been approved in both jurisdictions, but with one of the approvals having been granted outside the date range of literature search.

As a result of this searching, four further approvals were identified, all for drugs licensed by the EMA after 1999 but approved by the FDA prior to this. These were

- Paclitaxel, approved by the EMA in 1999, approved by the FDA 1997
- Sodium Phenylbutyrate, approved by the EMA in 1999, approved by the FDA in 1996

- Zinc acetate, approved by the EMA in 2004, approved by the FDA in 1997
- Anagrelide, approved by the EMA in 2004, approved by the FDA in 1997

These were therefore also included in my review, as they facilitate comparisons of the approach of FDA and EMA to uncontrolled data. For completeness, any drug not listed as approved (or rejected) by an agency was searched for using Google, in case of press releases or financial statements that detailed a withdrawn or rejected application.

3.1.3 CONSOLIDATED LIST OF TREATMENTS LICENSED ON THE BASIS OF UNCONTROLLED STUDIES FROM 1999 TO 2014

The number of approvals based on uncontrolled clinical study data for each regulatory body and the consolidated number of approvals (taking into account treatments licensed in both jurisdictions for the same indication) are shown in Table 3-1.

Table 3-1: Number of drugs and indications approved in the EU and US on the basis of uncontrolled clinical study data

	Drugs	**Indications**
FDA approvals	54	64
EMA approvals	35	44
Total number of approvals	62	74

A full list of the approved drugs, disease area, categorisation, approval status for each agency is shown in Table 3-2.

Table 3-2: Drugs submitted to the EMA and FDA containing only uncontrolled clinical studies

Generic name	Condition	Categorisation	FDA Status	EMA Status	RCT data available
Abarelix	Prostate cancer	Solid tumour oncology	A	-	No
Alemtuzumab	Chronic lymphocytic leukaemia (CLL)	Haematological oncology	A	A	No
Alglucosidase Alfa	Pompe disease	Rare metabolic condition	A	A	No
Alipogene Tiparvovec	Familial lipoprotein lipase deficiency (LPLD)	Rare metabolic condition	-	A	No
Anagrelide	Essential thrombocytopenia	Blood count	A*	A	No
Argatroban	Heparin-induced thrombocytopenia (HIT)	Blood count	A*	-	No
Arsenic Trioxide	Acute promyelocytic leukaemia (APL)	Haematological oncology	A	A	No
Asparaginase Erwinia Chrysanthemi	Acute lymphoblastic leukaemia (ALL)	Haematological oncology	A	-	No
Bendamustine Hydrochloride	Non-Hodgkin's Lymphoma	Haematological oncology	A	-	Yes
Betaine Anhydrous	Homocystinuria	Rare metabolic condition	-	A	No
Bexarotene	Cutaneous T-cell lymphoma (CTCL)	Solid tumour oncology	A	A	No
Bortezomib	Multiple myeloma (MM)	Haematological oncology	A	A	No
Bortezomib	Mantle cell lymphoma (MCL)	Haematological oncology	A	-	Yes
Bosutinib	Chronic myeloid leukaemia (CML)	Haematological oncology	A	A	Yes
Brentuximab Vedotin	Hodgkin's lymphoma (HL)	Haematological oncology	A	A	No
Brentuximab Vedotin	Systemic anaplastic large cell lymphoma (sALCL)	Haematological oncology	A	A	No
Busulfan	Haematopoietic progenitor cell transplantation (HPCT)	Haematological oncology	A	A	Yes
Carfilzomib	Multiple myeloma (MM)	Haematological oncology	A	-	No
Carglumic Acid	Chronic hyperammonemia	Rare metabolic condition	A	A	No
Ceritinib	Non-small cell lung cancer	Solid tumour oncology	A	-	No
Cetuximab	Colorectal cancer	Solid tumour oncology	A	A	No
Cholic Acid (Kolbam)	Inborn errors in primary bile acid synthesis	Rare metabolic condition	-	A	No
Cholic Acid (Orphacol)	Inborn errors in primary bile acid synthesis	Rare metabolic condition	-	A	No
Cladribine	Hairy cell leukaemia	Haematological oncology	-	A	No
Clofarabine	Acute lymphoblastic leukaemia (ALL)	Haematological oncology	A	A	No
Crizotinib	Non-small cell lung cancer	Solid tumour oncology	A	-	No
Dasatinib	Chronic myeloid leukaemia (CML)	Haematological oncology	A	A	No
Dasatinib	Philadelphia chromosome-positive acute lymphoblastic leukaemia (ALL)	Haematological oncology	A	A	No
Defibrotide	Veno-occlusive disease	Poisoning	-	A	Yes
Dexrazoxane Hydrochloride	Anthracycline extravasation	Poisoning	-	A	No
Ferric Hexacyanoferrate(Ii)	Internal contamination with radioactive caesium or thallium	Poisoning	A	-	No
Gefitinib	Non-small cell lung cancer (NSCLC)	Solid tumour oncology	A	-	No
Gemtuzumab Ozogamicin	Acute myeloid leukaemia (AML)	Haematological oncology	S	S	No
Glucarpidase	Toxic plasma methotrexate concentrations	Poisoning	A	S	No
Hydroxocobalamin	Treatment of cyanide poisoning	Poisoning	A	A	No
Ibrutinib	Mantle cell lymphoma (MCL)	Haematological oncology	A	-	No
Ibrutinib	Chronic lymphocytic leukaemia (CLL)	Haematological oncology	A	-	No
Imatinib Mesylate	Chronic myeloid leukaemia (CML)	Haematological oncology	A	A	No
Imatinib Mesylate	Gastrointestinal stromal tumours (GIST)	Solid tumour oncology	A	A	No

Imatinib Mesylate	Myelodysplastic / myeloproliferative diseases (MDS / MPD) associated with platelet-derived growth factor receptor (PDGFR) gene re-arrangements	Haematological oncology	A	A	Yes
Imatinib Mesylate	Soft tissue sarcoma - Dermatofibrosarcoma protuberans (DFSP)	Solid tumour oncology	A	A	Yes
Imatinib Mesylate	Philadelphia chromosome-positive acute lymphoblastic leukaemia (ALL)	Haematological oncology	A	-	Yes
Imatinib Mesylate	Aggressive systemic mastocytosis (ASM)	Haematological oncology	A	S	Yes
Imatinib Mesylate	Advanced hypereosinophilic syndrome (HES) and / or chronic eosinophilic leukaemia (CEL) with FIP1L1-PDGFR rearrangement	Haematological oncology	A	A	Yes
Ixabepilone	Breast cancer	Solid tumour oncology	A	S	Yes
Lomitapide Mesylate	Familial hypercholesterolemia (HoFH)	Rare metabolic condition	A	A	No
Metreleptin	Lipodystrophy due to leptin deficiency	Rare metabolic condition	A	-	No
Nelarabine	T-cell acute lymphoblastic leukaemia / lymphoma (T-ALL / T-LBL)	Haematological oncology	A	A	No
Nilotinib Hydrochloride Monohydrate	Chronic myeloid leukaemia (CML)	Haematological oncology	A	A	No
Nitisinone	Hereditary tyrosinemia	Rare metabolic condition	A	A	No
Ofatumumab	Chronic lymphocytic leukaemia (CLL)	Haematological oncology	A	A	No
Omacetaxine Mepesuccinate	Chronic myeloid leukaemia (CML)	Haematological oncology	A	S	No
Paclitaxel	Kaposi's sarcoma	Solid tumour oncology	A	A	Yes
Pasireotide Diaspartate	Cushing's disease	Rare metabolic condition	A	A	No
Pentetate Calcium Trisodium	Internal contamination with plutonium, americium, or curium	Poisoning	A	-	No
Pentetate Zinc Trisodium	Internal contamination with plutonium, americium, or curium	Poisoning	A	-	No
Pomalidomide	Multiple myeloma (MM)	Haematological oncology	A	-	No
Ponatinib Hydrochloride	Chronic myeloid leukaemia (CML)	Haematological oncology	A	A	No
Ponatinib Hydrochloride	Philadelphia chromosome-positive acute lymphoblastic leukaemia (ALL)	Haematological oncology	A	A	No
Pralatrexate	Peripheral T-cell lymphoma (PTCL)	Haematological oncology	A	S	No
Raxibacumab	Anthrax inhalation	Poisoning	A	-	No
Romidepsin	Peripheral T-cell lymphoma (PTCL)	Haematological oncology	A	S	Yes
Sodium Ferric Gluconate Complex	Iron deficiency	Rare metabolic condition	A	-	No
Sodium Phenylbutyrate	Urea cycle disorders	Rare metabolic condition	A*	A	No
Sunitinib Malate	Renal cell carcinoma	Solid tumour oncology	A	-	No
Taliglucerase Alfa	Gaucher's disease	Rare metabolic condition	A	S	No
Temoporfin	Head and neck cancer	Solid tumour oncology	-	A	No
Temozolomide	Anaplastic astrocytoma	Solid tumour oncology	A	A	Yes
Tocofersolan	Vitamin E deficiency due to cholestasis	Rare metabolic condition	-	A	No
Tositumomab; Iodine I 131 Tositumomab	Non-Hodgkin's lymphoma	Haematological oncology	A	-	No
Trabectedin	Soft tissue sarcoma	Solid tumour oncology	-	A	No
Vismodegib	Basal cell carcinoma	Solid tumour oncology	A	A	No
Vorinostat	Cutaneous T-cell lymphoma (CTCL)	Haematological oncology	A	S	No
Zinc	Wilson's disease (hepatolenticular degeneration)	Rare metabolic condition	A*	A	No

A=Approved, A*=Approved prior to 1999, S=Submitted but not approved

A full description of each the circumstances surrounding each individual drug approval has been presented in a separate working paper (Hatswell, Baio & Freemantle, 2017). These working papers summarise the evidence for the treatment, the regulatory milestone dates, and the decisions made by the FDA and EMA, including if a submission was withdrawn or rejected.

3.1.4 DISEASE AREAS WHERE UNCONTROLLED STUDIES HAVE MOST FREQUENTLY BEEN THE BASIS FOR DRUG APPROVALS

Of the 74 indications approved without controlled trial data, the largest single group (34 treatments) can be categorised as treatments for haematological malignancies. These treatments (for example imatinib, ofatumumab, and carfilzomib) were licensed on the basis of uncontrolled trials using response rates as the primary outcome. The next most common types of approval are treatments for metabolic disorders ($n = 15$) such as taliglucerase for Gaucher's disease, and solid tumour oncology treatments ($n = 15$) which used response rate as the primary outcome in trials - for example ixabepilone for the treatment of metastatic breast cancer.

The remaining approvals were for poisonings ($n = 8$), and treatments based on haematological markers ($n = 2$), for example anagrelide for the treatment of essential thrombocytopenia.

That the majority of approvals (49/74, 66%) were either haematological or solid tumour oncology corresponds with previous work regarding drug licensing, which shows a lower barrier to oncology drug approval in the US (Light & Lexchin, 2015). This particularly seems to be the case with FDA reviews – of the nine rejected EMA applications, seven were in oncology with the EMA highlighting uncertainty regarding the benefit-risk of the treatments.

Of the total of 74 indications, 39 of the treatments would primarily be used by haematologists, whilst 60 of the 74 approvals were for treatments that were not already licensed on the basis of RCT evidence. All of the treatments that had other RCT evidence in another indication were in either haematological oncology, or solid tumour oncology.

3.1.5 COMPARISON BETWEEN THE FDA AND EMA ON THE NUMBER OF APPROVALS, AND THE DATES OF REVIEWS

In comparing the two agencies, the FDA and EMA received different numbers of applications for treatments based on uncontrolled clinical studies (counting only treatments approved in at least one jurisdiction).

3.1.5.1 CONSISTENCY OF DECISION BETWEEN THE EMA AND FDA

Of the 44 applications made to both the FDA and EMA, there also appears to be a difference in the chance of approval. The FDA approved 43 of the 44 applications made rejecting 1, whilst the EMA approved 35 of the 44 - the nine applications not approved were either rejected by the EMA, or withdrawn by the submitting manufacturer, with a provisional negative decision in place.

In addition to the nine rejections, in a further five cases the EMA approval was only given once results from a RCT were available.

3.1.5.2 DIFFERENCES IN DATES OF SUBMISSIONS TO REGULATORS

Based on the data in this review, companies appear to submit to the FDA before the EMA. Of the 44 treatments submitted to both the FDA and EMA, 35 were submitted first to the FDA. The mean delay from the FDA submission to the EMA submission was 7.1 months in treatments approved by both agencies, and 7.2 months including those rejected by one agency.

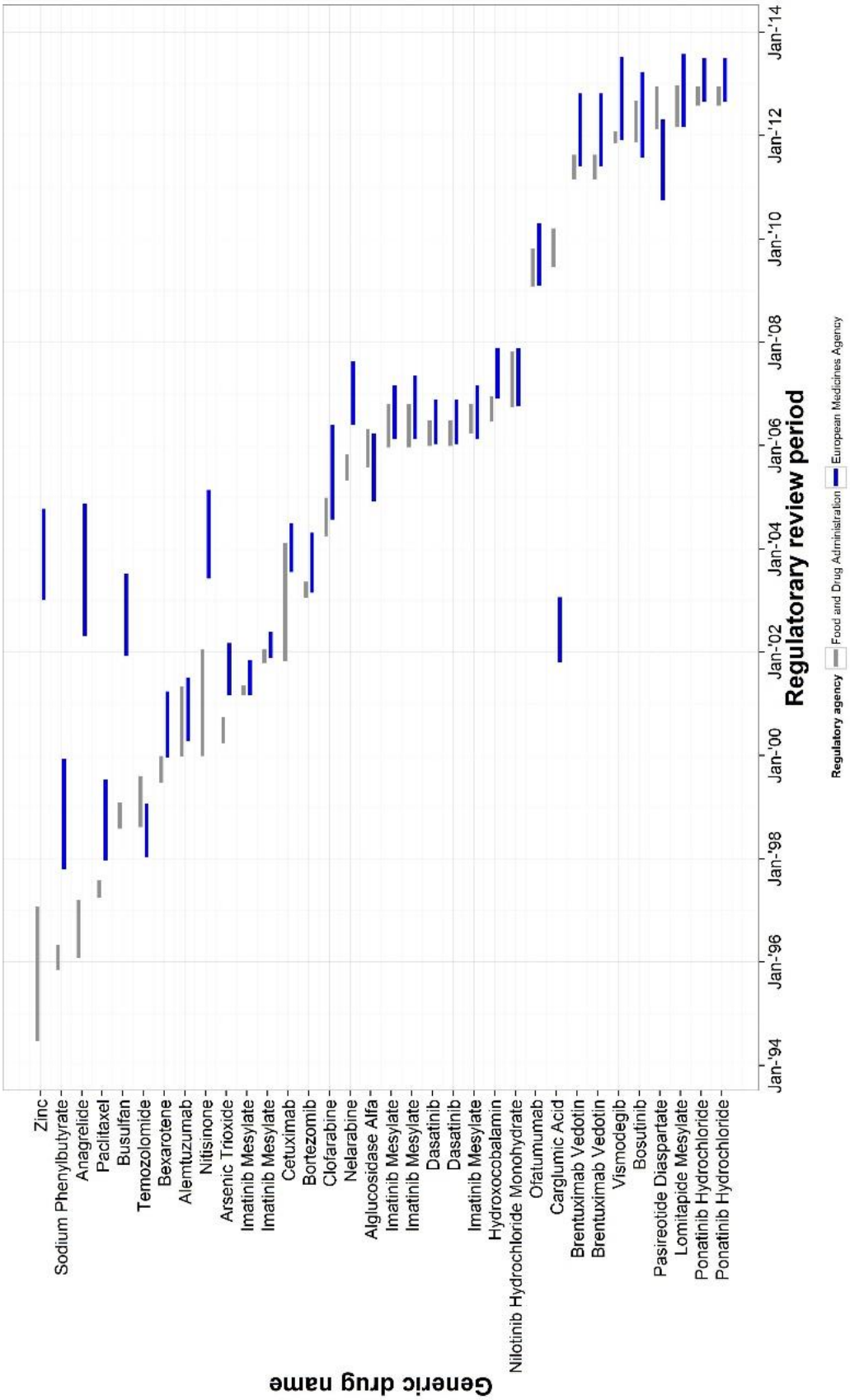
Whilst it is to be expected the submissions cannot be conducted in parallel due to the same staff working on both (for example trial statisticians and researchers), the literature search conducted suggests a strong preference for submitting to the FDA first. Whilst we can only speculate on the reasons for this apparent difference, potential reasons could include a more favourable regulatory environment (either perceived or real), differences in review time, the market size (pharmaceuticals are typically priced higher in the US), differences in uptake rates, or familiarity with the US healthcare system (many pharmaceutical companies are headquartered in the US).

3.1.5.3 COMPARISON OF FDA AND EMA REVIEW TIMES

In comparing the review times, a clear difference between the two agencies is also apparent, with the FDA taking less time to reach a decision. Of the 34 treatments approved by both agencies, the FDA had a shorter review on 31 occasions. In these instances the mean FDA review time was 8.7 months, compared to a mean of 15.5 months for the EMA – a difference of 6.8 months.

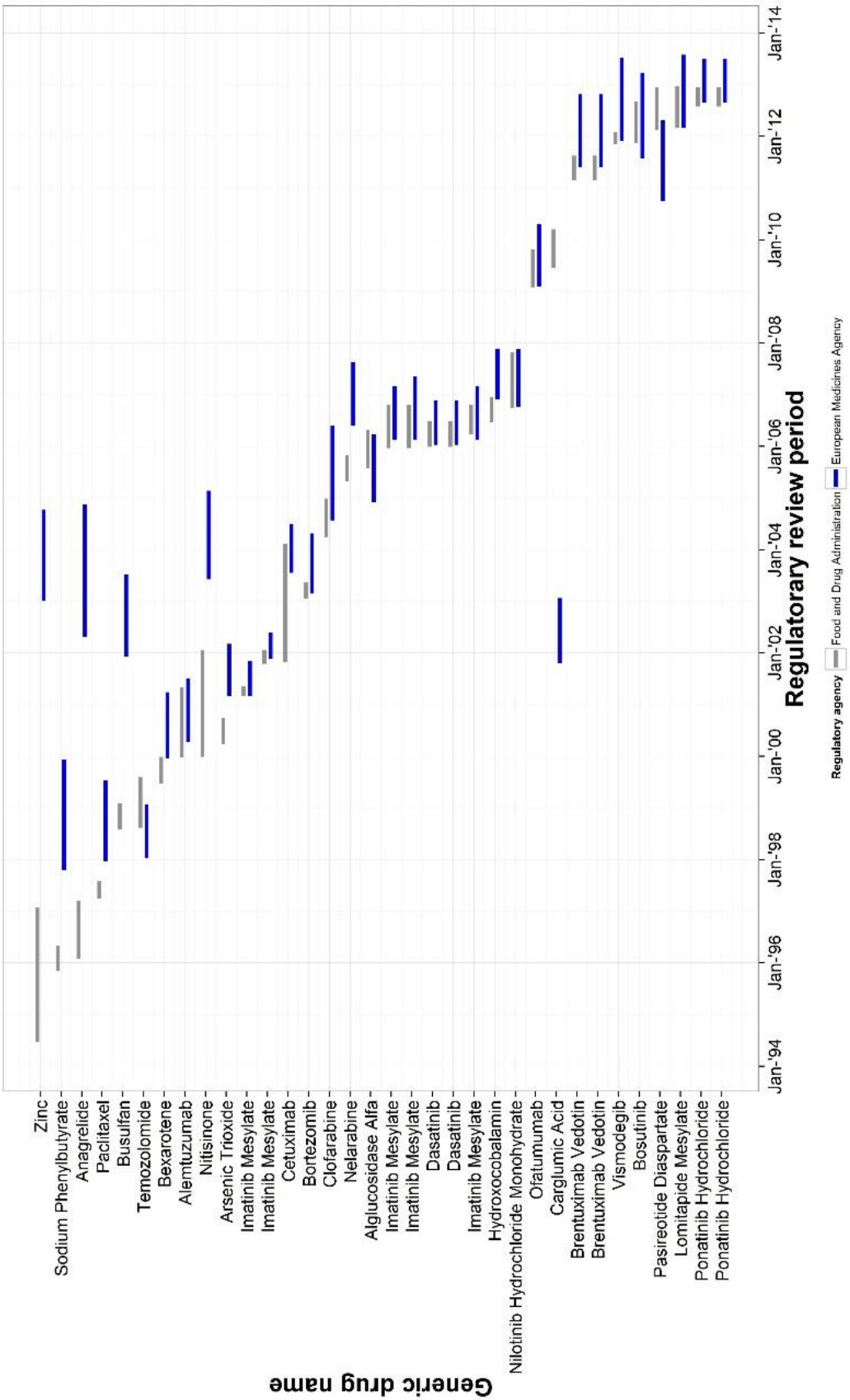
The differences in submission and review times is shown in

Figure 3-4: Timeline from submission to approval of pharmaceuticals licensed on the basis of uncontrolled study data by both the FDA and EMA



for treatments approved in both regions – the diagram clearly shown the EMA review time (blue) beginning after the FDA review time (grey), and lasting longer. Including effects of the delay in submission, the mean delay to European approval after US approval was 13.1 months. Not shown in the diagram are the five treatments approved only in Europe on the basis of comparative data, which were approved a mean of 21.5 months after US approval was granted.

Figure 3-4: Timeline from submission to approval of pharmaceuticals licensed on the basis of uncontrolled study data by both the FDA and EMA



3.1.6 CONGRUENCE OF FINDINGS WITH THE EXISTING LITERATURE

3.1.6.1 DIFFERENCES BETWEEN FDA AND EMA APPROVALS

The finding that the FDA reviewed applications faster than the EMA is not unique only to drugs licensed on the basis of uncontrolled study data. A study of tyrosine kinase inhibitors approved by the two agencies found the FDA reviewed products faster than EMA (205 vs 410 days), however found that the difference was due to 'clock stops' within the EMA process, accounting for 0 vs 184 days of review time (Shah, Roberts & Shah, 2013). A second study also noted a difference in the speed of approval of cancer drugs, a median of 182 days for the FDA compared to 350 days for the EMA (Roberts, Allen & Sigal, 2011). A further study has shown a similar trend of longer review times by the EMA in all pharmaceuticals (not specific to oncology) – 303 days for the FDA vs 366 days for the EMA (Downing *et al.*, 2012).

There are several potential reasons for the differences observed between agencies. Firstly the FDA's extensive use of 'accelerated approvals', with products allowed to show results on a surrogate endpoint, with confirmatory RCTs completed at a later point (Dagher *et al.*, 2004). This may also explain why the difference in oncology treatments is much larger than that seen with pharmaceuticals in general. A second potential reason may be differences in the attitude of regulators to benefit-risk (which may in turn reflect differences in population attitudes). Qualitative interviews with American and US regulators found those in the US to be more willing to give patients the opportunity to benefit from treatments - even if the outcomes were less certain. In contrast a more conservative approach was seen in European regulators (Tafuri *et al.*, 2014). This apparent difference in attitude however does not always lead to the FDA being more receptive to uncertainty; as a different study found that although regulatory processes between the two regions do occasionally lead to clinically meaningful differences in outcomes, the direction of these differences is not consistent (Trotta *et al.*, 2011).

Another potential reason for the differences between the two agencies in review time, and approval rate, is the level of interaction prior to a submission. When applying for approval to market a drug in the US, companies will frequently take advice from the FDA on trial design, and meet to discuss what outcomes are expected. For example FDA guidance on clinical trial endpoints states:

> Although general principles outlined in this guidance should help applicants select endpoints for marketing applications, we recommend that applicants meet with the FDA before submitting protocols intended to support NDA or BLA marketing applications. The FDA will ensure that these meetings include a multidisciplinary FDA team of oncologists, statisticians,

> clinical pharmacologists, and often external expert consultants. Applicants can submit protocols after these meetings and request a special protocol assessment that provides confirmation of the acceptability of endpoints and protocol design to support drug marketing applications (Food and Drug Administration, 2007:p.12)

This is a different process to the EU, where companies may choose to take scientific advice from the EMA, though this is not a required step. It does seem logical however that if companies are encouraged to meet regularly with the FDA to ensure a submission package will be acceptable, provided the data from trials is supportive, there should be fewer questions regarding whether an application is approvable. Equally if questions around the applicability of endpoints have been discussed prior to the submission being received, this should speed the process.

Another difference between the two agencies, is how interactions are conducted, and 'stop clock'. The process for the EMA centralised procedure states that the Committee for Medicinal Products for Human Use (CHMP) must issue an opinion within 210 days of a review beginning. As a part of this process there are two opportunities for the EMA to ask questions of the manufacturer, during which the 210 day 'clock' is not counting, until the company responds (Jefferys & Jones, 1995). This is different to the FDA process of back and forth questions and continuous review until a decision is reached (Ciociola *et al.*, 2014). Given the limited opportunities for questions from the EMA, it does seem logical that those asked will be extremely comprehensive, and may exceed what may be seen to be required as the review continues. This in turn may lead to a longer process than one where the regulator is allowed to ask questions on an ad hoc basis.

A final factor influencing the timing of decisions may be that that once a recommendation has been made by the CHMP, it is then passed to the European Commission (a body of civil servants and politicians) to give a final decision, a step that takes 67 days (although this can take substantially longer). This stage adds to the time needed for treatments to be approved in the EU (Wade, 2010).

3.1.6.2 NUMBER OF TREATMENTS APPROVED ON THE BASIS OF UNCONTROLLED STUDIES OVER TIME

Whilst my review did not identify any clear trend for approvals over time based on uncontrolled studies, anecdotal evidence is that this rate is either increasing, or expected to increase. The reason for this relates to regulatory initiatives to allow patients access to medicines at earlier stages, for example the EMA Priority Medicines (PRIME) scheme. Under this scheme drugs will be allowed earlier market access (usually conditional on further trials) for diseases with an immediate threat to life (Antoñanzas, Terkola & Postma, 2016). In

practice such early data are likely to take the form of uncontrolled studies, though how the process will evolve is unclear.

A second potential reason for not identifying a trend is that the cut-off date for my review was the date of searches (May 2014), shortly before the approval of a series of treatments for Hepatitis C – many of which were based on uncontrolled studies. Whilst these products may or may not have resulted in an increase in the number of treatments approved without RCT evidence in being for a large patient population (and with a high price), their availability may have raised awareness of such approvals (Kish, Aziz & Sorio, 2017).

3.1.6.3 DISEASE AREAS WHERE UNCONTROLLED STUDIES ARE CONDUCTED

That the majority of approvals were in cancer, and specifically haematology mirrors the finding of Saccà (2010) who found that whilst only 13% of ongoing chronic heart failure studies were uncontrolled, 66% of studies in acute myeloid leukaemia were uncontrolled. Although not directly comparable (the studies were not necessarily registrational studies), this finding does indicate that uncontrolled studies are more acceptable in some areas of medicine.

3.1.7 SUBSEQUENT WORK PERFORMED BY OTHERS IN THE AREA

The review published in 2014 as a part of my PhD has been referenced by other researchers (as of 4 August 2019 there were 42 citations listed in Google Scholar, 11 of which were listed in PubMed Central) who have used the paper in support of other work, or developed the research further.

The papers most relevant to my work look further in to drug approvals based on uncontrolled studies. In this area work by Shepshelovich et al. (2018) found that drugs licensed with supporting RCT evidence had fewer label changes for safety reasons than the drugs I identified as having being licensed without supportive RCT data. Work by Djulbegovic et al. (2018) looked at the effect sizes seen with drugs licensed using only uncontrolled study data. They found these effect sizes larger (using a variety of techniques) than the effect sizes seen in drugs licensed with RCT data – indicating (as may be expected) that the decision to pursue a license with uncontrolled study data be indeed be linked to ‘better than could be expected’ outcomes. Also of relevance is a review by Davies et al. (2017) which showed that it is uncommon that further data become available for drugs licensed on uncontrolled studies (even when no RCT evidence is available). This highlights the importance of the question addressed in the thesis (of how best to estimate comparative

effectiveness), as in most cases it appears the evidence available from uncontrolled studies at launch will not be validated in confirmatory RCTs (or potentially in studies of any kind).

The most relevant citation however would be by Goring et al. (2019) who updated the systematic review I performed (though not using as broad of a scope) to include newer products, and investigated the source of control arms further. Their findings were stated to corroborate those seen in my review – that uncontrolled studies are rarely (but consistently) used for drug approvals – with an apparent increase in the period from 2012 to 2017, compared to 2005 to 2011. This however represents an arbitrary date cutoff, and if other dates were used, the findings may have been different – it therefore remains to be seen if the rate of approvals increase. Should this be the case I suspect it will be due to the introduction of drugs with different mechanisms which have a more compelling reason for belief in a step change; for instance gene therapies and Chimeric antigen receptor therapies, as opposed to a differing willingness to accept uncontrolled data from TKIs, for instance.

Citations which are less relevant to my work though still of interest, involve work around reimbursement. The first paper suggests that the lack of a RCT does not hinder the reimbursement of treatments compared to drugs in similar indications (Anderson *et al.*, 2019). A second paper then found that the determining factor in drug approval by payers was the type of marketing authorisation, with the type of evidence available per se not being the determining feature of coverage decisions, but rather the evidence impacting the type of license granted, with lower rates of coverage from payers for drugs with conditional marketing authorisation (Vreman *et al.*, 2019).

3.2 METHODS USED FOR ESTIMATING EFFECTIVENESS FROM UNCONTROLLED STUDIES

A summary of the identification of treatments licensed on the basis of uncontrolled studies was published in Pharmacoeconomics (Hatswell, Freemantle & Baio, 2017a)*, with a full report published as a UCL research report* (Hatswell, Freemantle & Baio, 2017b)

Having identified treatments licensed on the basis of uncontrolled clinical studies, the second stage of my review was to search for modelling approaches that have been used for drugs licensed on data from uncontrolled clinical studies.

To do this, I searched published literature (via PubMed and the ISPOR Scientific Presentations Database) and HTA websites (NICE, SMC, and AWMSG) for modelling studies performed to estimate the effectiveness of the drugs identified in Chapter 3. The results of these literature searches are described in the sections below.

3.2.1 LITERATURE SEARCH FOR MODELLED ESTIMATES OF THE EFFICACY OF DRUGS LICENSED ON THE BASIS OF UNCONTROLLED CLINICAL STUDIES

To identify models based on uncontrolled study data, three sources were searched, these were

- Medline (via PubMed), for peer reviewed papers
- The ISPOR Scientific Presentations Database
- The databases of NICE, SMC and AWMSG, the UK Health Technology Assessment bodies

The searches performed in each of these databases are discussed in turn, along with the rationale for the inclusion of each paper

3.2.1.1 MEDLINE (VIA PUBMED)

MEDLINE is the US National Library of Medicine journal citation database, with over 22 million citations to biomedical and life sciences journal articles dating back to 1946. Journals are assessed for quality, before they are indexed in the database (National Institute for Health, 2014).

PubMed is a website that searches the MEDLINE database, plus an additional number of publications – those available online ahead of print (prior to full citations being available from MEDLINE), the inclusion of other general science journals, and articles from journals before that journal was included on MEDLINE. In total, this adds approximately an additional two million items (National Institute for Health, 2014). Given the research I am interested in (i.e. medicine), this will only have been published in medical and health services research journals. The comprehensiveness of the PubMed search therefore negates the need to search other indexes.

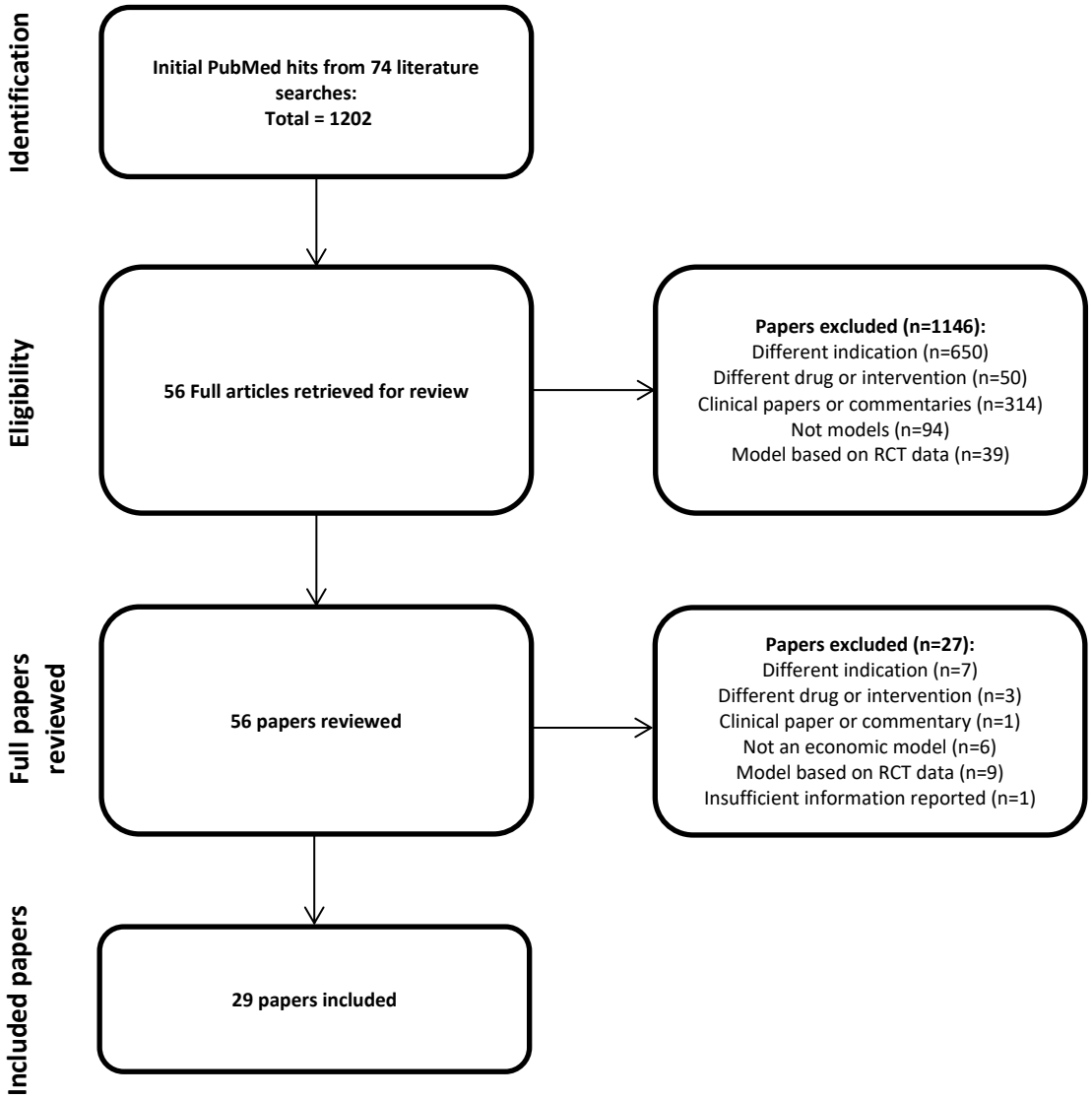
A series of search terms was used to find peer-reviewed modelled estimates of efficacy of the medicines identified in Chapter 3 – including extrapolations of trial data, economic models, and comparative effectiveness estimates. These search terms were then combined with each individual drug name, and 74 searches conducted; one for each of the drugs of interest. Full search terms are given in Appendix B.

After potentially relevant searches were completed for each drug, the results were filtered by reviewing the title and abstracts to exclude irrelevant publications, with the reason for exclusion noted. The full papers extracted were then retrieved for review. The reasons for

exclusion were evaluated in a hierarchical fashion, with the first reason for exclusion listed, and the abstract not then assessed against the other criteria.

Full text articles were then reviewed, either for inclusion as economic models based on uncontrolled study data or as irrelevant to the thesis, with the reason for exclusion noted. The result was that 29 full publications were included, from 1202 initial hits in PubMed. A full PRISMA diagram for the literature review is shown in Figure 3-5.

Figure 3-5: PRISMA diagram of modelled comparisons retrieved from PubMed



3.2.1.2 ISPOR SCIENTIFIC PRESENTATIONS DATABASE

Health economic models are not always presented in full peer reviewed publications due to the different objectives of both pharmaceutical companies and employees – although

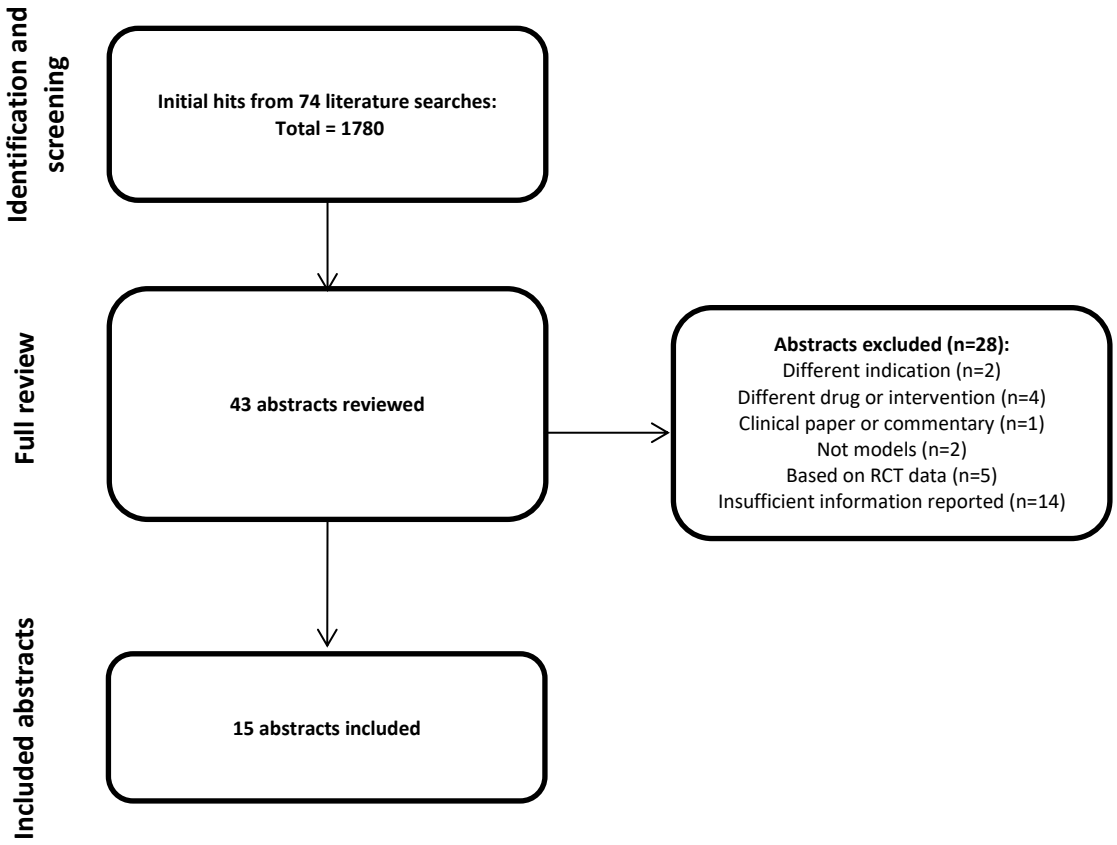
publication is desirable in some instances, it is not a priority nor a metric for career achievement. In addition, work is likely to be presented at conferences before a full (peer reviewed) publication is available. In order to capture models that fit into these categories, posters and presentations from meetings of the International Society for Pharmacoeconomics and Outcomes Research conferences were also searched. In contrast to other conferences in the field ISPOR also has a large attendance from the pharmaceutical industry, who are responsible for the production of many models – this means the archives are more likely to contain relevant abstracts.

Although the conference abstracts can be searched online (with the full poster or presentation available as a PDF if uploaded by the author), the database does not allow for complex searches. As only simple Boolean operators are possible, a search was conducted for each drug by generic OR US trade OR EU trade name.

The results of these initial searches were then screened for relevance by initial reading of the abstract and title. Potentially relevant abstracts, or where there was uncertainty regarding relevance, the abstract was included for full review. After the initial review, a full review was performed of abstracts that passed the initial screening. Where available, this included downloading and reviewing the PDF of the poster / presentation.

Based on these searches and review, 16 conference abstracts were included from 1780 initial hits, with 43 abstracts reviewed in full. A full PRISMA diagram is shown in Figure 3-6.

Figure 3-6: PRISMA diagram of modelled estimates identified in the ISPOR Scientific Presentations database



3.2.1.3 HEALTH TECHNOLOGY APPRAISAL BODIES (NICE, SMC, AND AWMSG)

In the UK NHS, decisions on whether a new pharmaceutical will be approved for use, in general, are taken by health technology assessment bodies. In England and Wales, the National Institute for Health and Care Excellence (NICE) is the main agency, with some decisions (Multiple Technology Appraisal) also applicable in Scotland. In general for a NICE submission, the company will construct an economic model, which will be critiqued by an independent academic group (the Evidence Review Group, or ERG), who may also construct their own model. These models (along with the clinical data and other analyses provided) then form a part of the deliberative process for the appraisal committee, who then issue a decision of whether the drug should be funded in the NHS.

Whilst NICE have topics referred to them by the Department of Health, on launch, the Scottish Medicines Consortium (SMC) will often ask to see a full submission for a new drug from the company (this decision is made by the SMC, following provision of outline information by the company). As a result, they conduct a lot more appraisals than NICE, with

a shorter process. Essentially for SMC submissions, the company submits a dossier with clinical and economic evidence to the SMC, who will have it reviewed by an independent economist and who will ask questions of the company to identify any issues with the submission. All the evidence is then taken into consideration at the SMC meeting, where a decision on whether the drug should be approved for use is made.

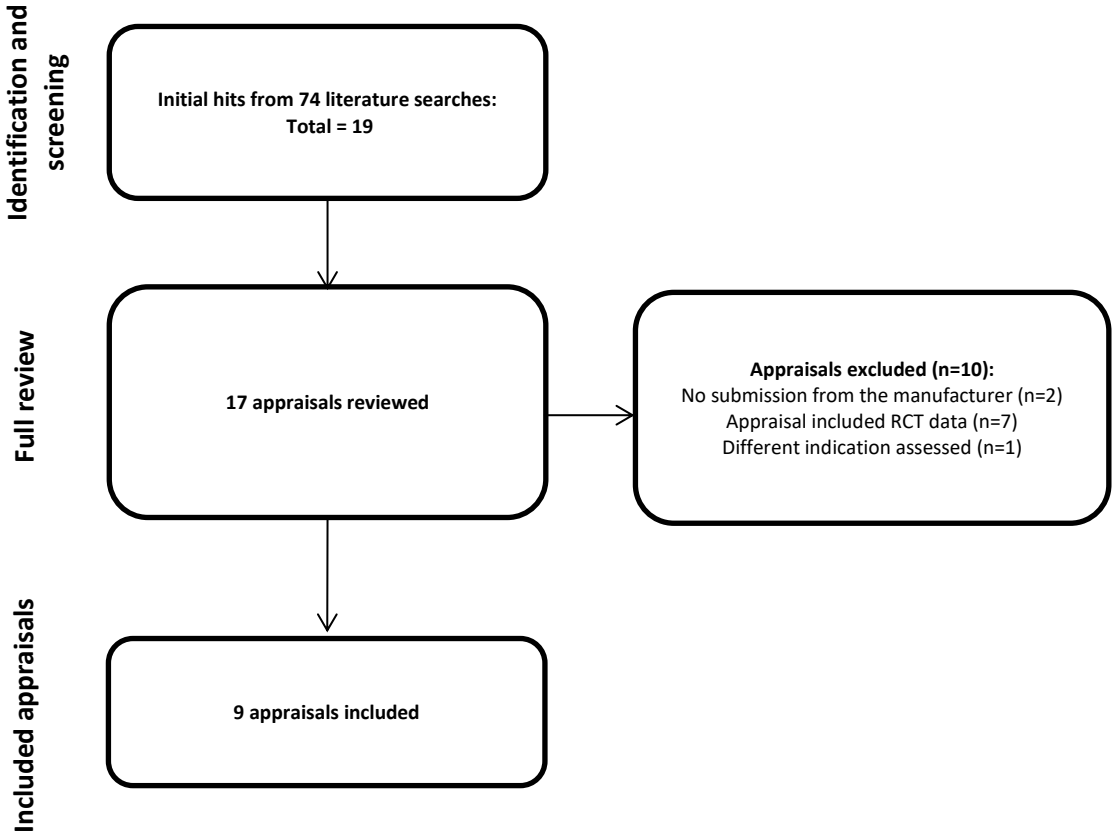
The All Wales Medicines Strategy Group (AWMSG) has a similar process to the SMC, although it does not review drugs that are due to be reviewed by NICE, unless there are exceptional circumstances (for example a long delay in NICE guidance), as NICE guidance overrules AWMSG guidance.

To identify guidance issued based on the drugs identified in Chapter 3, I searched the websites of each of the UK health technology appraisal bodies, with the results of these searches described below.

3.2.1.3.1 NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE) APPRAISALS

I searched the NICE website and downloaded the documentation surrounding each relevant appraisal, which was reviewed for economic models constructed as a part of the process, either by the manufacturer or the independent evidence review group (ERG). This search yielded 19 hits, with nine assessments having relevant models included after review. A full PRISMA diagram is shown in Figure 3-7.

Figure 3-7: PRISMA diagram of NICE appraisals involving economic models



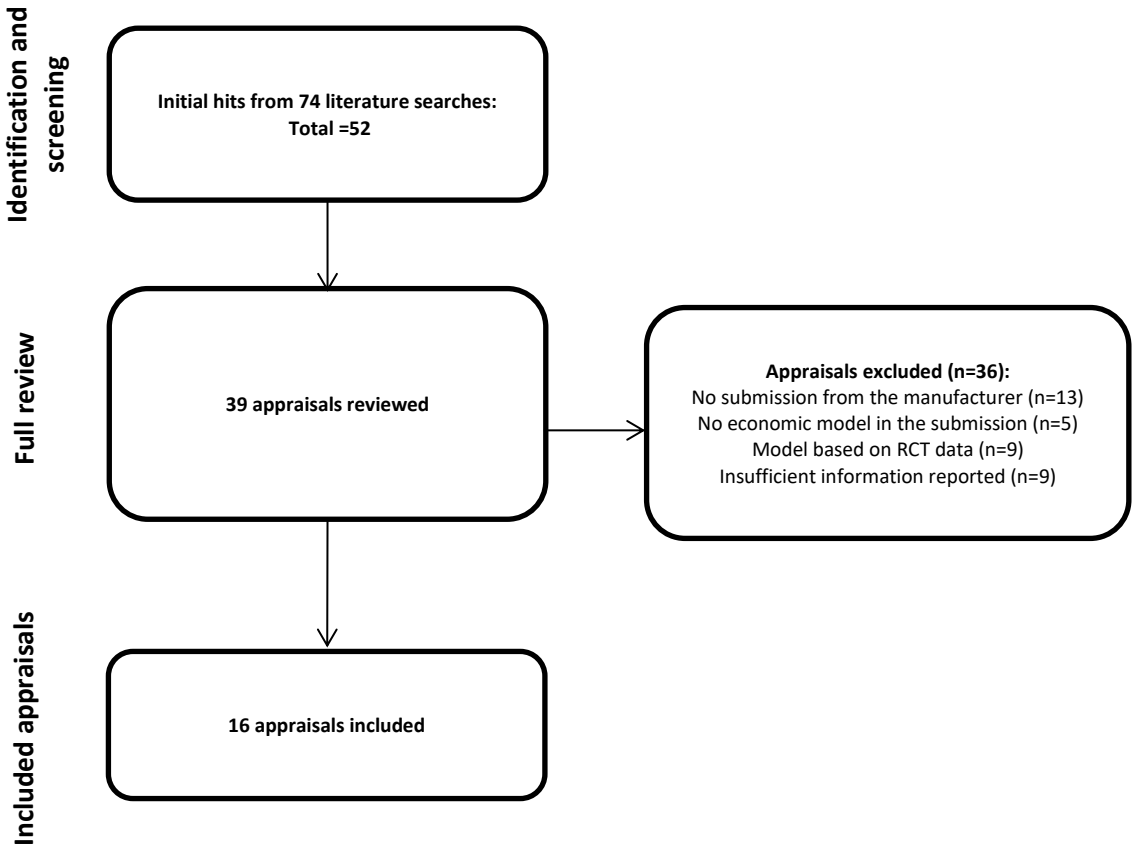
3.2.1.3.2 SCOTTISH MEDICINES CONSORTIUM SUBMISSIONS

I searched the SMC website, with all relevant appraisal documents available then reviewed.

Prior to 2005, SMC assessments offered limited publicly available information; with only a statement regarding a positive or negative decision available. Over time, the amount of information released has increased to be (at the time of the literature review) a summary of the clinical evidence provided, a summary of the health economic evidence provided, and an assessment of the strengths and weaknesses of each. This level of information has not been provided retrospectively; therefore, for some early SMC assessments, it is not possible to conclude how any economic model(s) were constructed, and even in later appraisals, there is often ambiguity in the methods.

The resulting review of SMC submissions yielded 16 submissions describing economic evaluations of interest. Interestingly, 13/52 times the manufacturer of a product chose not to submit to the SMC. A full PRISMA diagram is shown in Figure 3-8.

Figure 3-8: PRISMA diagram of Scottish Medicines Consortium submissions

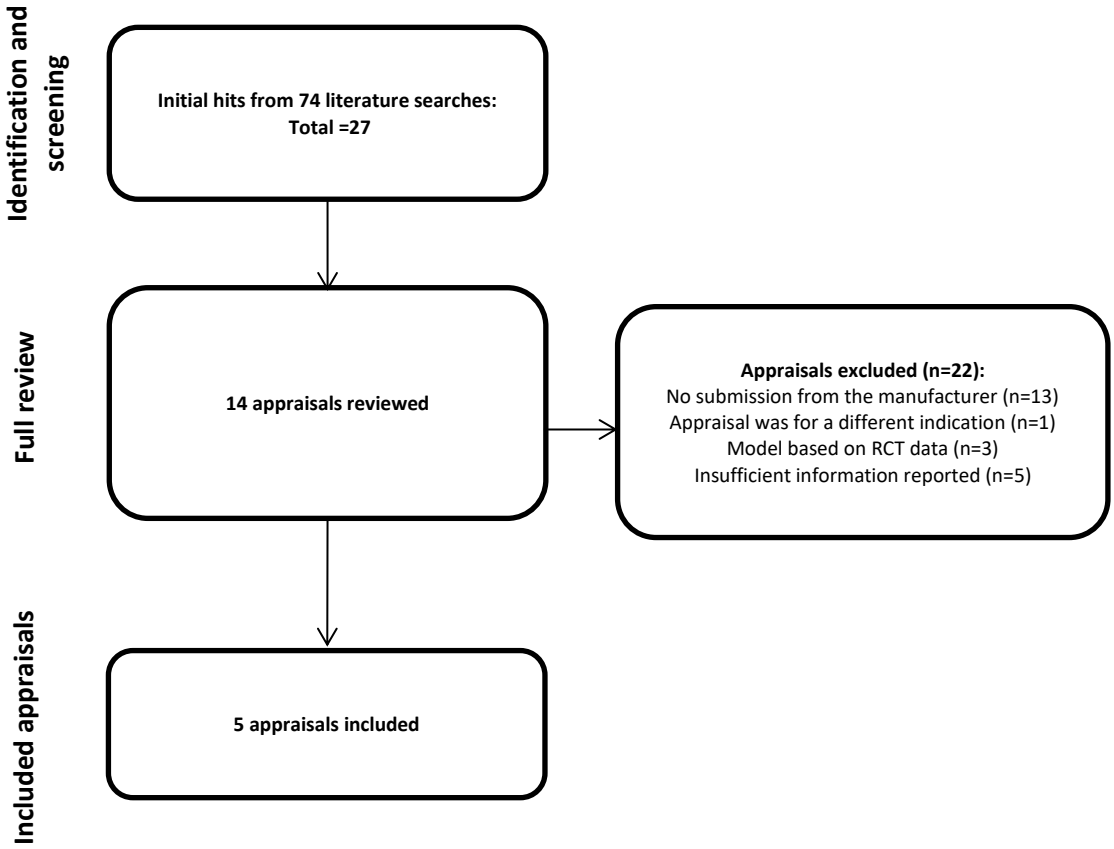


3.2.1.3.3 ALL WALES MEDICINES STRATEGY GROUP (AWMSG) SUBMISSIONS

The process for identifying AWMSG submissions of relevance was very similar to that used for identifying SMC submission, with similar issues relating to the level of reporting in the early 2000s and the number of non-submissions (13/27).

Ultimately, five appraisals were included with a full PRISMA diagram shown in Figure 3-9.

Figure 3-9: PRISMA diagram of All Wales Medicines Strategy Group submissions



3.2.1.4 CONSOLIDATED REPORTING OF IDENTIFIED MODELLED ESTIMATES OF EFFICACY

After searching the two databases and websites of the three health technology assessment bodies, a total of 76 papers was identified from the various sources. A tabulation of the number of hits, reason for exclusion (shown in italics), and number included from each source is shown in Table 3-3.

Table 3-3: Number and source of modelled estimates of efficacy identified as being based on uncontrolled clinical study data, with reasons for exclusion shown in italics

	NICE	SMC	AWMSG	ISPOR	PubMed	Totals
Number of hits	19	52	27	1780	1202	**3136**
Non-submissions	*2*	*13*	*13*	*-*	*-*	***28***
For review	17	39	14	43	56	**169**
Excluded	8	23	9	28	27	**95**
Different indication	*0*	*0*	*1*	*2*	*7*	***11***
Different drug or intervention	*0*	*0*	*0*	*4*	*3*	***7***
Clinical paper or commentary	*0*	*0*	*0*	*1*	*1*	***2***
Not an economic model	*0*	*5*	*0*	*2*	*6*	***13***
Model based on RCT data	*8*	*9*	*3*	*5*	*9*	***33***
Insufficient information	*0*	*9*	*5*	*14*	*1*	***29***
Included	**9**	**16**	**5**	**15**	**29**	**74**

3.2.2 DESCRIPTION AND DE-DUPLICATION OF PUBLISHED ESTIMATES

During full review, it was apparent that a number of the models were reported multiple times, for example used in a NICE submission, an SMC submission, and then presented at ISPOR. Equally, other papers reported several approaches to modelling the uncontrolled data, and these different approaches were therefore more relevant than describing individual papers.

De-duplication was performed by looking through the detailed descriptions of model structures, and approach to estimation of efficacy. Three examples of models without a 1:1 relationship with publications are listed below to give examples of the results of the de-duplication process. A full description of each of the models identified is given in Hatswell et al. (2017b)

- Cost-utility decision tree of argatroban compared to alternative treatments in heparin-induced thrombocytopenia
    - This model appears to have been used for both SMC and AWMSG submissions and compared argatroban to two alternatives and no treatment (AWMSG, 2012; Scottish Medicines Consortium, 2012).
    - Clinical efficacy data for argatroban was taken from the historically controlled trials used for the drug licensing (comparing argatroban to a historical case series of US patients). Comparisons were also made to other treatments (danaparoid and lepirudin), using a naïve comparison of the treatment arms.
    - The model structure, comparators, and clinical data were described as being the same in both the AWSMG and SMC documents, with the time frame also being identical

- Comparison of temoporfin photodynamic therapy with palliative care or chemotherapy in advanced head and neck cancer
    - This model was described in a publication by Hopper et al., as being a comparison of palliative care, chemotherapy and photodynamic therapy(Hopper, Niziol & Sidhu, 2004). The model used trial results from the two clinical studies (without adjustment) as inputs for the effectiveness of the three treatments in the model – one RCT of chemotherapy vs. palliative care, and the temoporfin trial. No adjustments were made for differences in patients between trials.
    - The same model was adapted to Germany, with the same approach to the estimation of effectiveness but German costs used – the authors (some of whom are authors on the publication by Hopper et al) state that "an already published model developed on the base of English data was fed with German cost−data" (Kübler *et al.*, 2005).
- Trabectedin for the treatment of soft tissue sarcoma (STS) using a historical control
    - Trabectedin was assessed by NICE for the treatment of STS in TA185, with the company creating a model for submission to estimate the cost effectiveness of the drug. For comparative data, the model used data from the trabectedin clinical study, and compared this to a set of historical controls. In this case the historical controls consisted of four pooled trials published by the European Organisation for Research and Treatment of Cancer – Soft Tissue and Bone Sarcoma Group. The data from the historical controls were then adjusted to match the control group using a regression with dummy variables for performance status, histopathology of disease, age, and gender – this improved the estimated survival of best supportive care slightly.
    - The model was used in the submission to NICE (Simpson, Rafia & Stevenson, 2009), and later published in a HTA report (Simpson *et al.*, 2010) and discussed in a review paper (Rafia *et al.*, 2013). In addition to the NICE appraisal, the model was also used in two SMC submissions, first without controlling for differences between baseline characteristics (Scottish Medicines Consortium, 2010), and then in a resubmission, where the regression model used for NICE was applied, and price of the drug also reduced (Scottish Medicines Consortium, 2011).
- Sunitinib compared to best supportive care in the treatment of second-line metastatic renal cell carcinoma

- The model compared the results observed in the uncontrolled sunitinib trial, to the outcomes reported in a published case series, and to Medicare data from the US (Scottish Medicines Consortium, 2007).
- The model is described in the SMC submission, with also a Belgian adaptation presented at ISPOR (Van Nooten *et al.*, 2007). A Spanish adaption was also performed where only the comparison with Medicare data was presented, both as an ISPOR poster (Aiello *et al.*, 2007) and in a peer reviewed journal (Paz-Ares *et al.*, 2010).

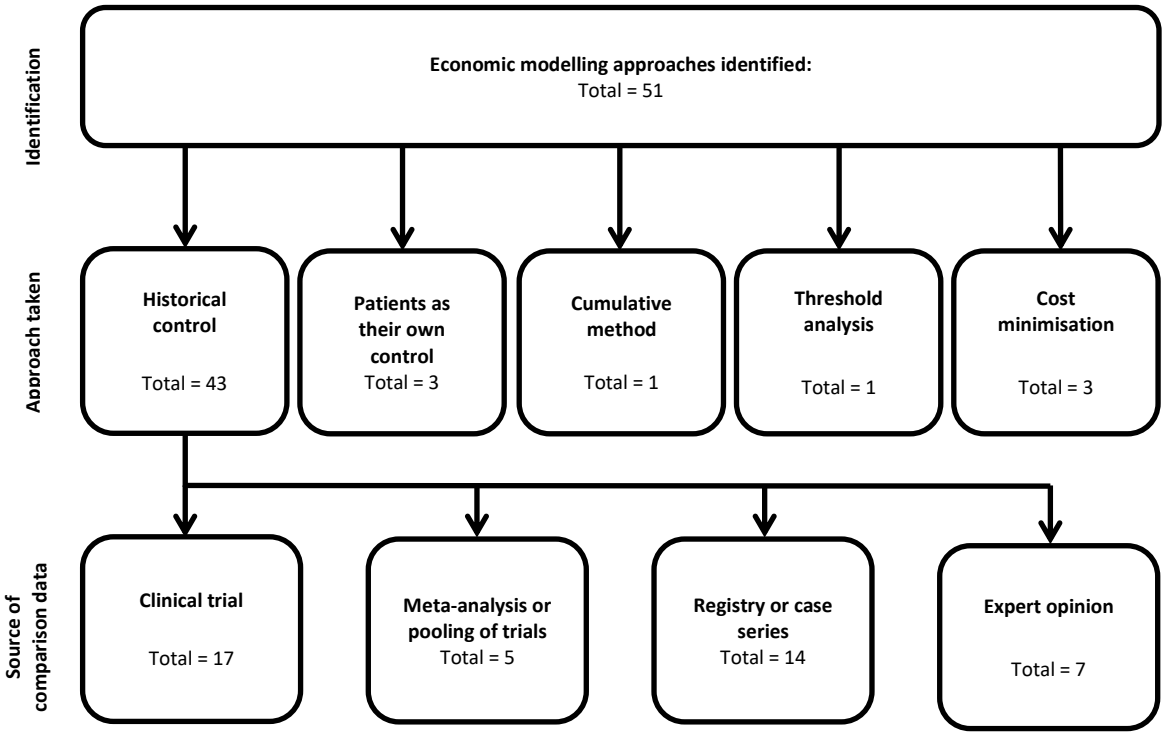
As a result of this full review and de-duplication, of the 74 publications, 91 approaches were identified (some publications contained more than one approach – for instance the argatroban and sunitinib models). The total is reduced to 51 individual models when taking duplicate reporting into account.

Each of the approaches performed has been summarised in a separate working paper (Hatswell, Freemantle & Baio, 2017b), with references to the appropriate papers, abstracts, and health technology assessment submissions.

3.2.3 A TAXONOMY OF MODELLING APPROACHES

Of the 51 modelling approaches identified, the majority (43) were based on historical controls of various forms, shown in Figure 3-10.

Figure 3-10: Taxonomy of economic modelling approaches used for estimating incremental benefit from uncontrolled clinical studies



Of the 43 historical controls, 22 used comparisons to arms from other clinical trials – 17 compared directly to the results of a single trial, and five compared to either pooled (simply combining response rates) or meta-analysed clinical trials (trials combined through the use of formal methods). Overall this form of comparison (i.e. historical control) was the most popular approach used. Despite potential differences between trials, only 7 of the 22 attempted to correct for any differences between the trials, with various types of corrections attempted.

A further 14 historically controlled models used either registry data (5) or case series data (9). The concern with these studies would be that patients enrolled in case series or registries may be systematically different to those enrolled in trials. Several studies did attempt to account for this by only including patients who met the entry criteria for the study used for comparison, although the level of selection was highly variable.

Finally, 7 historically controlled models were compared only to expert opinion due to a lack of available data. The obvious concern here is that the results of any economic analysis are completely dependent on the input given by the relevant experts.

Of the approaches not using historical controls, three models could loosely be described as using the patient as their own control, taking the change from baseline as being due to the drug (with obvious concerns regarding regression to the mean). Three models were cost minimisations, which has been extensively criticised in the literature as being an inappropriate approach by simply sidestepping the question of comparative effectiveness and assuming equality of treatments (Briggs & O'Brien, 2001; Dakin & Wordsworth, 2013). One model assumed a 'cumulative approach' (Hoyle *et al.*, 2013), where patients who received the drug had benefit only for the time on treatment (and subsequent health states were unaffected), and one model was a threshold analysis of how effective the drug would need to be to show cost-effectiveness.

3.3 SUMMARY OF FINDINGS FROM LITERATURE SEARCHES

In reviewing drug approvals, it can be seen that although uncommon, approvals granted without RCT evidence are not rare, occurring several times per year. Although the majority are in oncology, many other disease areas are included. It is also not simply the case that follow on indications were granted for a drug known to be effective in RCTs, but in rarer subgroups – the vast majority (80%) of drugs had their first approval in an uncontrolled study.

In terms of modelling it can be seen that there is no consensus on how best to model uncontrolled studies. Whilst the historical control does seem to be the main vehicle used, there was a large degree of variation in what comparisons were made and how data were adjusted (or not) to account for differences between trials. Trial data and registry data were also used interchangeably, with only a few studies accounting for patient selection. There was also no discernible pattern in the approaches by source (HTA, publication or conference presentation).

Given the frequency of uncontrolled approvals and the use of historical controls in assessing comparative effectiveness, the next chapter focuses on the role of MAIC in addressing differences between study populations, and how historical controls may be created where none are currently available.

4 SIMULATION STUDY REGARDING THE PERFORMANCE OF UNANCHORED MATCHING ADJUSTED INDIRECT COMPARISON (MAIC)

The work described in this chapter has been published in Value in Health (Hatswell, Freemantle & Baio, 2020)

As highlighted in the NICE DSU covering the adjustment methods of MAIC and STC (Phillippo *et al.*, 2016), as well as the subsequent publication (Phillippo *et al.*, 2017), a need for research on the performance of MAIC (ideally through the use of simulation studies) was noted. This was a need I had separately identified following my literature review of models as there were clear differences between studies, but uncertainty as to which circumstances MAIC may have been able to assist with. As a result this was a study I designed and conducted.

Phillippo et al. in the NICE DSU guidance assess the two applications of MAIC separately; *unanchored* MAIC, where comparisons are made across individual study arms, and *anchored* MAIC, where controlled studies are reweighted (including a common comparator arm), for inclusion in Network Meta Analysis). Due to the types of problems frequently seen with uncontrolled studies, e.g. potential differences in patient characteristics between trials, the primary interest I had was in the robustness of unanchored MAIC under model misspecification. This would appear to be the area most relevant to uncontrolled studies due to uncertainty in which variables should be included, and how their structure would affect results.

The reason MAIC as a method is relevant in this setting is that data availability when comparing to historical data is frequently an issue - with many patient and disease characteristics either not measured or not reported in historical publications. This can be seen in my literature review, where of historical controls the majority were of published studies, with few companies appearing to have access to the ILD of both studies of interest. As in such circumstances the studies included in analysis are conducted independently there is also a high chance of non-overlapping populations in at least some variables, due to difference in inclusion / exclusion criteria. The question of relevance to my research therefore is how MAIC performs in such situations, where there are incomplete data for matching, potential differences between patient populations, and whether MAIC is able to appropriately adjust for different relationships between explanatory covariates.

A secondary question for me was around the implementation of MAIC; it is possible to match on mean values of characteristics, or also on higher moments (such as the variance) – as a measure of variability is also sometimes reported alongside mean values for patient

characteristics in studies. Whilst both approaches are described in the paper proposing MAIC as a method, no preference is stated (Signorovitch *et al.*, 2010). Indeed reviewing previous publications using the approach, both approaches appear to have been used, as well as different approaches to the number of variables included – from matching on all available characteristics, to some studies which selected preferred variables.

4.1 APPROACH AND DATA GENERATION

In order to test to the robustness of the method, a simulation study was conducted mimicking data from end stage cancer – the most common area for MAIC to be used (23 of the 58 published examples). Survival data were simulated based on each simulated patients' underlying health, but also their baseline characteristics – the sum of these values was then used as a linear predictor of outcomes. In the case of treated patients this was then modified by a hazard ratio to create outcomes (Y) with presenting proportional hazards. Outcomes were generated for two arms – a contemporary study (Population A), whose units were assumed to have more favourable baseline characteristics, as well as receiving treatment (T^T) and a historical control arm (Population B) made by individuals who received the control arm (T^C). For all patients, outcomes were sampled and outcomes calculated including and excluding the effect of treatment. This allows to allow the performance of MAIC in estimating the 'true' difference to be calculated.

To ensure that the data would be applicable to the type of problems seen, patients were assumed to have six uncorrelated characteristics ($\mathrm{X}_1, \dots, \mathrm{X}_6$) that influenced outcomes. Four assumed to be fully observed and available for weighting ($\mathrm{X}_1, \dots, \mathrm{X}_4$) whilst two ($\mathrm{X}_5$, X_6) were assumed to be unobserved. Patient characteristics were sampled in a way such that there was a bias of half a standard deviation in favour of the intervention arm in observed characteristics, with unobserved characteristics drawn from the same distribution (an assumption which is varied in sensitivity analysis). A linear predictor was created using an intercept, and the sum of products of characteristics with the corresponding effect size ($\beta_1, \dots, \beta_6$). This outcome model is shown mathematically below

$$Survival \sim Weibull(shape = \alpha, scale = \lambda)$$

where:

$$\alpha \sim Normal(mean = 1.3, standard\ deviation = 0.1)$$

$$\lambda = \frac{\exp(\beta_0 + \sum \beta_1 \mathrm{X}_1, \dots, \beta_6 \mathrm{X}_6)}{Treatment\ Hazard\ Ratio}$$

NB: β_0 is the intercept in the model, which is fixed as 2

$$\beta_1, \ldots, \beta_4 \sim Normal(mean = 0.5, standard\ deviation = 0.2)$$

$$\beta_5,\ \beta_6 \sim Normal(mean = 0, standard\ deviation = 0.2)$$

For the intervention (T^T):

$$X_1, \ldots, X_4 \sim Normal(mean = 0.3, standard\ deviation = 0.1)$$

$$Treatment\ Hazard\ Ratio = 0.75$$

For the control (T^C):

$$X_1, \ldots, X_4 \sim Normal(mean = 0.25, standard\ deviation = 0.1)$$

$$Treatment\ Hazard\ Ratio = 1$$

For both intervention & control:

$$X_5, X_6 \sim Normal(mean = 0.25, standard\ deviation = 0.1)$$

Six characteristics were selected as this was loosely informed by work on predictive characteristics in different cancers - for example three identified in bladder cancer by Bellmunt et al. (2010), and work showing a median of six characteristics were adjusted for in published MAICs (range 1-13) (Phillippo *et al.*, 2019).

In the base case, the characteristics were assumed to be independent. Although it is likely that some predictive and prognostic characteristics are linked (for instance in the paper by Bellmunt et al. patients with liver metastases are likely to have worse ECOG status), the degree of correlation is uncertain, and how this would impact results is also unknown – for transparency therefore these are assumed not to be correlated in the base case, which is varied in sensitivity analysis. This approach also allows an independent assessment of the impact of each issue that may affect an individual analysis, as there are examples where variables would not be expected to be correlated – for example lung cancer studies where there may exist the two uncorrelated prognostic characteristics of race and gender.

Each simulated characteristic ($X_1, \ldots, X_6$) was multiplied by a corresponding effect size ($\beta_1, \ldots, \beta_6$) with a linear predictor then created using the sum of products added to an intercept, which was set to 2 in the base case. This linear predictor was used in a Weibull proportional hazards survival model; a corresponding survival time (Y) was sampled for each

patient with and without receiving the intervention. For simplicity normal distributions were selected, which were varied in a sensitivity analysis.

In each simulation the patient characteristics, effect sizes, and resulting survival times were resampled. The intention of the simulation was to address whether the method [MAIC] performs well in survival models. The interest however is in the class of survival model rather than in the specific parameters of the distribution chosen (a shape of 1.3 being the mean of what is simulated in the Weibull distribution for example). By allowing parameters such as the shape to vary, we are able to ensure the result holds for each distribution in general (at least as would apply to survival outcomes), as opposed to only being valid with specific parameters. This approach does mean that more simulations are needed (as variability is introduced) however this seems a fair compromise for a more generalisable result.

To isolate the effect of MAIC (as opposed to just Monte Carlo error), a large number of patients ($n = 1000$) were sampled for both Population A and Population B and survival times were assumed to be fully observed with no missing data or censoring. This particular setup (a large number of patients, fully observed survival times, and no missing data) was selected to ensure the study assessed weighting methods – and was not overly influenced by the variability between samples, or approach to missing data; be that due to administrative censoring i.e. requiring extrapolation, or the type of 'missingness' assumed. In reality MAIC is likely to be conducted alongside techniques to account for missing data (Gabrio, Mason & Baio, 2019; Leurent *et al.*, 2018) and / or extrapolation of survival times (Latimer, 2011).

4.2 APPLICATION OF MATCHING ADJUSTED INDIRECT COMPARISON

As would be seen with a historical control used without any adjustment, a naïve comparison contrasts the outcome from Population A who received the intervention ($t = 0$), with the outcomes from Population B, the historical control, who were assumed to receive control ($t = 0$). In keeping with the terminology of Signorovitch et al. where treatment is denoted by 0 and control by 1, this is comparing $\bar{Y}_0^A$ with $\bar{Y}_1^B$. Due to the more favourable characteristics in Population A, such a comparison would be biased. Weighting methods (both MAIC and PSW), attempt to remove this bias by reweighting the patients in Population A to match those in Population B (assuming PSW was estimating the Average Effect on the Controls). If successful, the reweighting of $\bar{Y}_0^A$ would match the (unobserved - and unobservable) $\bar{Y}_0^B$, which can then be contrasted with $\bar{Y}_1^B$ to estimate the effect of the intervention in a similar population.

This comparison ($\bar{Y}_0^B$ vs $\bar{Y}_1^B$) is indeed what would be generated by an RCT in population B of the two treatments, through the use of exchangeable groups such that the only

meaningful difference between groups is treatment assignment. By comparing the estimated effect from MAIC to the (unobserved) true effect, the success of both MAIC and PSW in estimating this true effect can be assessed. Due to being a simulation study, we are able to perform this comparison as data generation mechanisms are known, and thus outcomes can be computed with and without the intervention for both treatment and control groups. By comparing the estimated effect to the (unobserved) true effect, the success of both MAIC and PSW in estimating the true effect can be assessed.

In the simulation study MAIC was implemented using two approaches

- Matching on first moments i.e. means of $X_1, \ldots, X_4$, referred to in results as $MAIC_{FM}$
- Matching on the means and higher moments i.e. matching on means and on standard deviations, as given as an option in the original paper by Signorovitch et al. (2010) which states

  > For example, given the baseline mean and standard deviation of age, it is straightforward to compute the mean of squared age, which can then be treated as a separate mean baseline characteristic for matching.

  In results this is referred to as $MAIC_{HM}$

PSW was also conducted using the same approach using inverse probability of treatment weighting (IPTW) with weights for all patients in both arms. Although not technically a comparator to MAIC (as calculating propensity scores requires full access to individual level data for both studies) PSW is a well-recognised approach with a long history, and a recognised standard in observational data. Its inclusion in the simulation study allows an assessment of how much accuracy is lost when patient level data are only available from one study.

4.3 OUTCOMES OF THE STUDY

A Cox Proportional Hazards (CPH) model was used to estimate the hazard ratio between the two arms in line with previous work in the field (Petto *et al.*, 2019). The mean underlying value for the CPH was known to be 0.75; however, due to the difference in patient characteristics, a naïve comparison would overestimate the effectiveness of treatment (a CPH estimated using $\bar{Y}_0^A$ and $\bar{Y}_1^B$). This is because the more favourable patient characteristics give a further benefit to Population A. Each weighting method was then applied and used to estimate the CPH ratio using the reweighted Population A to give the estimated outcome in Population B with treatment ($\widehat{\bar{Y}_0^B}$) compared with the observed outcomes for the control ($\bar{Y}_1^B$).

In each simulation a number of outcomes were calculated:

- The mean percentage error in the hazard ratio. If the methods are unbiased this should be approximately zero (with only Monte Carlo error keeping this from being zero); an unbiased method should be as likely to under-predict effectiveness as over-predict
- The mean absolute percentage error in the hazard ratio. This is a measure of the accuracy of methods. Whilst a method may be unbiased, if it regularly exhibits large prediction errors it would be imprecise, and unsuitable for use
- The coverage probability - whether the 95% interval for each estimated hazard ratio contained the 'true' value which is helpful in understanding how much belief can be placed in a method given any individual result
- In addition, for the weighting methods, whether the point estimate of the hazard ratio was more accurate than a naïve comparison was also calculated. Over all simulations if a method was likely to lead to an increase in bias compared to a naïve comparison, this would be a concern in recommending a method

4.4 SCENARIO ANALYSES PERFORMED

In order to understand how the method of MAIC performs under different circumstances, a number of scenario analyses was conducted. These included varying the setup of the study to ensure any findings were not specific to the simulation set up, testing how MAIC may apply under different conditions, and finally in violating the explicit and implicit assumptions of MAIC to understand the implications.

In varying the setup of the simulation study scenarios included:

- The use of binary variables $(X_1, \ldots, X_4)$., as would be seen with characteristics such as male / female
- An alternative survival function (the exponential); achieved by setting the shape (α) in the Weibull survival distribution to 1
- Changing the explanatory power of the patient characteristics $(X_1, \ldots, X_6)$., to give them a lesser or greater importance
- Reversing the direction of bias in $X_1, \ldots, X_4$, such that the historical data had more favourable patient characteristics

In testing MAIC under different conditions, the scenarios explored included:

- Setting some (or all) parameters ($X_1, \ldots, X_4$) to be nuisance variables, uncorrelated with outcomes
- Testing a non-linear effect of patient characteristics. This was done by taking the exponential of $X_1, \ldots, X_4$, then dividing each by 5, to ensure the mean scores remained the same
- Varying the degree of overlap in Population A and Population B from the base case (a 0.5 SD difference between groups) to have either little (1 SD) or large (0.1SD) overlap
- Allowing for variables used in weighting ($X_1, \ldots, X_4$) to be correlated. This was implemented by having an 'underlying fitness' parameter, defined as a normal variable with standard deviation 0.1, which had mean 0.3 for population A, and 0.25 for population B. Parameters $X_1, \ldots, X_4$ were then generated as the 'underlying fitness' plus sampling from a normal distribution with mean 0 and standard deviation 0.1

Finally, when violating the implicit / explicit assumptions in MAIC, the following scenarios were tested:

- Including imbalances in the unobserved characteristics of X_5 and X_6 with these characteristics being correlated, or uncorrelated with the observed patient characteristics. The same distributions were then used for these as for $X_1, \ldots, X_4$
- Sampling from non-normal distributions for Population A and Population B, as would be seen in the ages of patients enrolled in trials (many cancers having increasing incidence with age, though fewer people survive to older ages). This was implemented through the use of the lognormal distribution.
- Sampling from trimmed distributions for patient characteristics ($X_1, \ldots, X_4$) for Population A and Population B, as would be seen when inclusion criteria (such as a minimum or maximum age or performance status) are included in trial entry criteria.

How each scenario was implemented in the simulation study is shown in Table 4-1. In addition to these scenarios, to understand whether it was more important to have more patients available to weight with, or match to, the number of patients sampled for Population A and Population B (using the base case approach) were varied in combination using $n = 30$, $n = 300$, and $n = 3000$.

Table 4-1: Base case and sensitivity analysis parameters for the simulation study of the performance of MAIC

Scenario	Base case	Scenario setting
Varying the setup of the simulation study		
All variables are binary	Covariates 1 to 4: Population: A $X_A \sim Normal(mean = 0.3, SD = 0.1)$: Population: B $X_B \sim Normal(mean = 0.25, SD = 0.1)$:	Covariates 1 to 4: Population A: $X_A \sim Binomial(probability = 0.3)$ Population B: $X_B \sim Binomial(probability = 0.25)$
Exponential distribution used as the survival function	Survival: $Weibull\ shape = 1.3$	Survival: $Weibull\ shape = 1$
Explanatory variable power is low	Covariates 1:4: $\beta \sim Normal(mean = 0.5, SD = 0.2)$	Covariates 1:4: $\beta \sim Normal(mean = 0.1, SD = 0.05)$
Explanatory variable power is high		Covariates 1:4: $\beta \sim Normal(mean = 1.0, SD = 0.4)$
Treatment effect is low	$TreatmentHR = 0.75$	$TreatmentHR = 0.9$
Treatment effect is high		$TreatmentHR = 0.2$
Covariate sampling is reversed i.e. Population A less favourable	Covariates 1 to 4: Population A: $X_A \sim Normal(mean = 0.3, SD = 0.1)$	Covariates 1 to 4: Population A: $X_A \sim Normal(mean = 0.2, SD = 0.1)$
Testing MAIC under different conditions		
Half the matched parameters are nuisance parameters	Covariates 1:4: $\beta \sim Normal(mean = 0.5, SD = 0.2)$	Covariates 1:2: $\beta \sim Normal(mean = 1.0, SD = 0.2)$ Covariates 3:6: $\beta \sim Normal(mean = 0.0, SD = 0.2)$
All the matched parameters are nuisance parameters		Covariates 1:4: $\beta \sim Normal(mean = 0.0, SD = 0.2)$
The effect of parameters is non-linear	$\lambda = \frac{\exp(2 + \sum \beta_1 X_1, \ldots, \beta_6 X_6)}{Treatment\ Hazard\ Ratio}$	$\lambda = \frac{\exp(2 + \exp(\sum \beta_1 X_1, \ldots, \beta_6 X_6)/5)}{Treatment\ Hazard\ Ratio}$
Small difference is covariate sampling (0.1SD)	Covariates 1 to 4: Population A: $X_A \sim Normal(mean = 0.3, SD = 0.1)$	Covariates 1 to 4: Population A: $X_A \sim Normal(mean = 0.26, SD = 0.1)$
Large difference in covariate sampling (1SD)		Covariates 1 to 4: Population A: $X_A \sim Normal(mean = 0.35, SD = 0.1)$
Parameters correlated		Underlying health: Population A: $H_A \sim Normal(mean = 0.3, SD = 0.1)$

		Population B: $\mathrm{H_B} \sim Normal(mean = 0.25, \mathrm{SD} = 0.1)$ Covariates 1:4: $\mathrm{X} \sim Normal(mean = 0, \mathrm{SD} = 0.1) + \mathrm{H}$
Violating assumptions implicit / explicit assumptions		
Missing parameters correlated with observed parameters	Covariates 5 & 6: $\mathrm{X} \sim Normal(mean = 0, \mathrm{SD} = 0.2)$	Covariates 5 & 6: $X \sim \frac{\sum \mathrm{X_1, \ldots, X_4}}{4} + Normal(mean = 0, \mathrm{SD} = 0.1)$
Missing parameters uncorrelated with observed parameters		Covariates 5 & 6: Population A: $\mathrm{X_A} \sim Normal(mean = 0.3, \mathrm{SD} = 0.1)$ Population B: $\mathrm{X_B} \sim Normal(mean = 0.25, \mathrm{SD} = 0.1)$ Covariates 1:6: $\beta \sim Normal(mean = 0.35, \mathrm{SD} = 0.15)$
Non-normal distributions sampled in Population A	Covariates 1 to 4: Population A: $\mathrm{X_A} \sim Normal(mean = 0.3, \mathrm{SD} = 0.1)$ Population B: $\mathrm{X_B} \sim Normal(mean = 0.25, \mathrm{SD} = 0.1)$	Covariates 1 to 4: Population A: $\mathrm{X_A} \sim Lognormal(meanlog = \log(0.27), \mathrm{SDlog} = 0.5)$
Non-normal distributions sampled in Population B		Covariates 1 to 4: Population A: $\mathrm{X_B} \sim Lognormal(meanlog = \log(0.22), \mathrm{SDlog} = 0.5)$
Trimmed patient characteristics in Population A (no poor performers)		Covariates: $\mathrm{X_A}$ resampled if <0.2
Trimmed patient characteristics in Population B (no good performers)		Covariates: $\mathrm{X_B}$ resampled if >0.35

SD = standard deviation

4.5 IMPLEMENTATION IN SOFTWARE AND MODEL CONVERGENCE

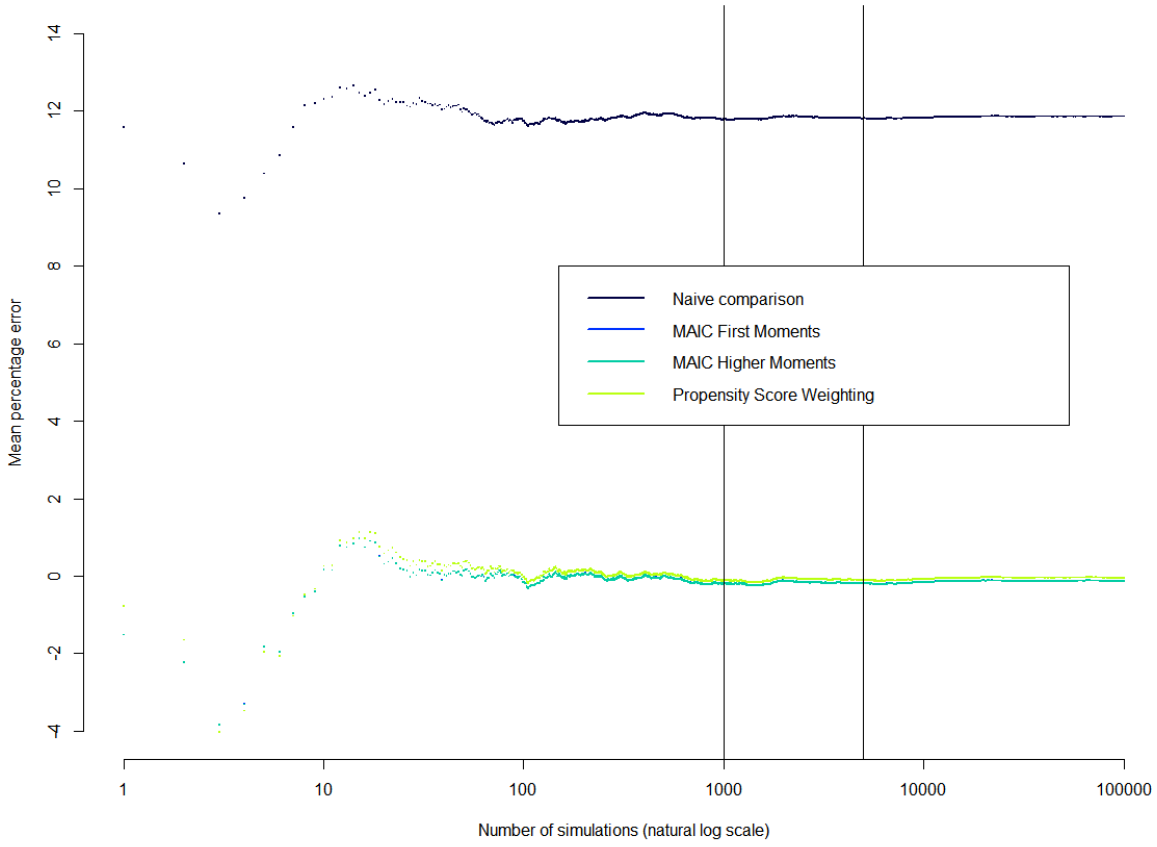
The simulation study was programmed in R version 3.6.1 (R Core Team, 2020) using the following packages

- *stats* to simulate survival curves
- *survival* to implement CPH models and robust standard errors)
- *ggplot2* and *ggsurvplot* to produce graphics
- *msm* to generate trimmed distributions for scenario analyses

To allow for reproducibility of the implementation of MAIC, the code snippet (including additional functionality such as cross-checks and validation) is presented in Appendix C.

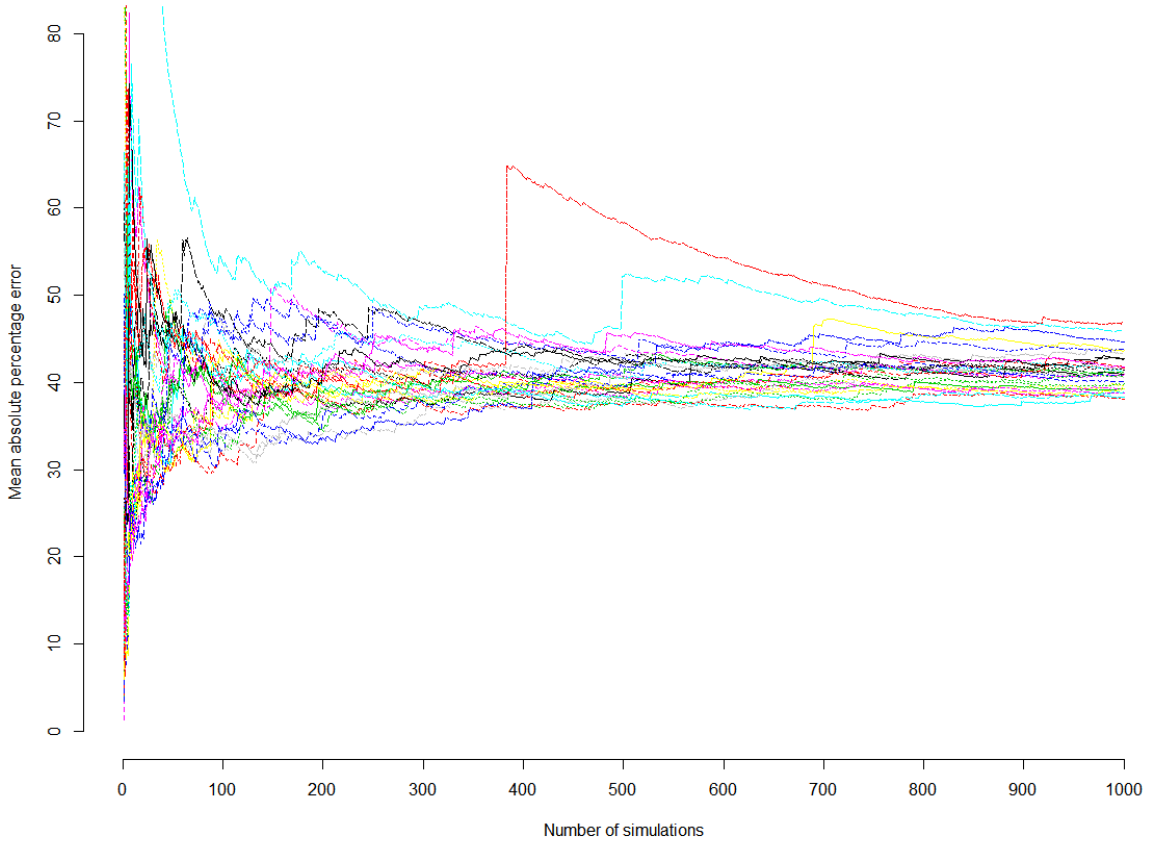
To ensure the results of the simulations were stable, 5000 simulations of each scenario were performed. This number was selected using two different methods. Firstly convergence plots were generated showing how different variables changed as the number of simulations increased (including going beyond the number of simulations used); an example is given below showing the mean percentage error which beyond the first several hundred simulations does not appear to change for the base case. In the example (the base case) the lines for each of the methods overlap around zero, as they appear to be unbiased

Figure 4-1: Convergence plot of the base case MAIC simulation study; Up to 100,000 simulations for mean percentage error of naïve comparison $MAIC_{FM}$, $MAIC_{HM}$ and PSW with vertical lines at 1,000 and 5,000 simulations



Some metrics for some scenarios (for instance the coverage probability when faced with a large effect size) were more uncertain, and took longer to stabilise. This was seen through running repeated batches of the same scenario and plotting the routes to convergence on the same graph. This 'route to convergence' is shown below for 30 runs of the mean absolute percentage error in the simulation with only 30 patients included in each arm.

Figure 4-2: Convergence plot of 30 runs of 1000 simulations for the base case mean absolute percentage error of $MAIC_{FM}$ in the scenario with n=30 in Population A and n=30 in Population B



The above graph indicates that the results appear to be close to stable by 1000 simulations, though as runtime is not a primary consideration in the study, a cautious approach was taken of performing 5000 simulations, which appeared to be well in excess of the largest number of simulations needed for results to stabilise.

A second approach was then taken at the request of a peer reviewer of the paper who desired a more objective measure of model convergence. This involved calculating Monte Carlo Standard Errors (MCSEs) using the *mcmcse* package. The values seen at 5,000 simulation (all <0.01) demonstrate that the samples drawn are representative of the underlying distributions; and that the findings are many times larger than the MCSEs indicates they are likely not due to chance.

4.6 FINDINGS FROM THE BASE CASE

In looking at the base case of the simulation study, the setup was such that the outcomes of the historical control (Population B) with untreated patients ($\bar{Y}_1^B$) had a mean survival of 11.4 months – had the patients received the intervention (with the associated hazard ratio), $\bar{Y}_1^B$,

this would have been 15.2 months. Because of the more favourable characteristics in Population A the observed survival ($\bar{Y}_0^A$) was 16.9 months; a bias of approximately 11% due to the healthier patients in Population A.

When weighting was performed on the simulated data, across the 5,000 simulations all weighting methods ($MAIC_{FM}$, $MAIC_{HM}$ and PSW) were unbiased (mean error close to zero), and had similar levels of accuracy in accounting for the bias inbuilt in the simulation i.e. had good accuracy. This is shown not only by the relatively low mean absolute percentage error, but importantly that few simulations (circa 1%) exacerbated the bias i.e. performed worse than a naïve comparison, with coverage probabilities over 90% (Table 4-2).

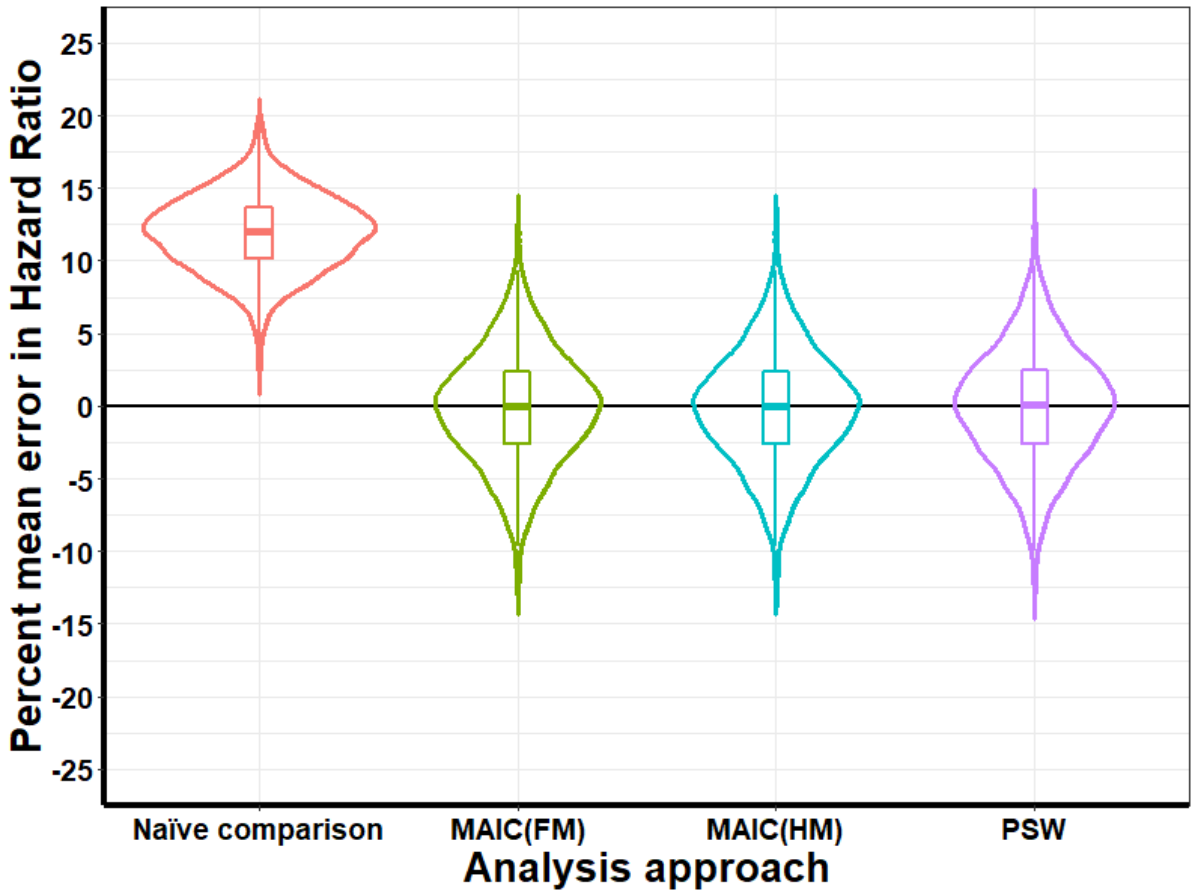
Table 4-2: Tabulated results of MAIC simulation study base case, 5,000 runs

Method	Mean Percentage Error (MCSE)	Absolute Percentage Error (MCSE)	Mean Standard Error	Coverage probability	Percent of scenarios worse than a naïve comparison
Base case					
Naïve comparison	11.8% (<0.01)	11.8% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.2% (<0.01)	2.6% (<0.01)	0.03	95%	2%
$MAIC_{HM}$	-0.2% (<0.01)	2.6% (<0.01)	0.03	95%	2%
PSW	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%

MCSE = Monte Carlo Standard Error, MAIC = Matching Adjusted Indirect Comparison, FM = First moments, HM = includes Higher moments, PSW = Propensity Score Weighting

The distribution of the error in the scenarios is shown in the violin plot in Figure 4-3, with overlaid box plots to short the quartiles of the error. This shows that all methods performed similarly given the setup of the base case scenario – as would be expected given the conditions are not set to challenge the way weighting is implemented.

Figure 4-3: Violin plot (with overlaid box plots) of the percent mean error in 5,000 runs of the MAIC simulation study base case



4.7 FINDINGS FROM SCENARIO ANALYSES

The results seen in the base case appear robust to altering the setup of the simulation study; whilst varying the characteristics of the study do results in some changes in the accuracy of the method, it remains unbiased. The only exception to this is the case of $MAIC_{HM}$ when binary variables are used (where it appears biased); this result however should be discarded as in such an implementation would never be performed (the square of a binary variable not being meaningful). Results of these scenarios are shown in Table 4-3 and Figure 4-4. In particular it is interesting to see that where the explanatory variable power is high i.e. explains a large proportion of the difference between trials, the use of some form of weighting becomes increasingly important (though the results are more uncertain due to the sampling variability). The final scenario included in this section also indicates that MAIC performs well regardless of the directionality in any bias i.e. the method works even if the patient characteristics in the data matched to are more favourable.

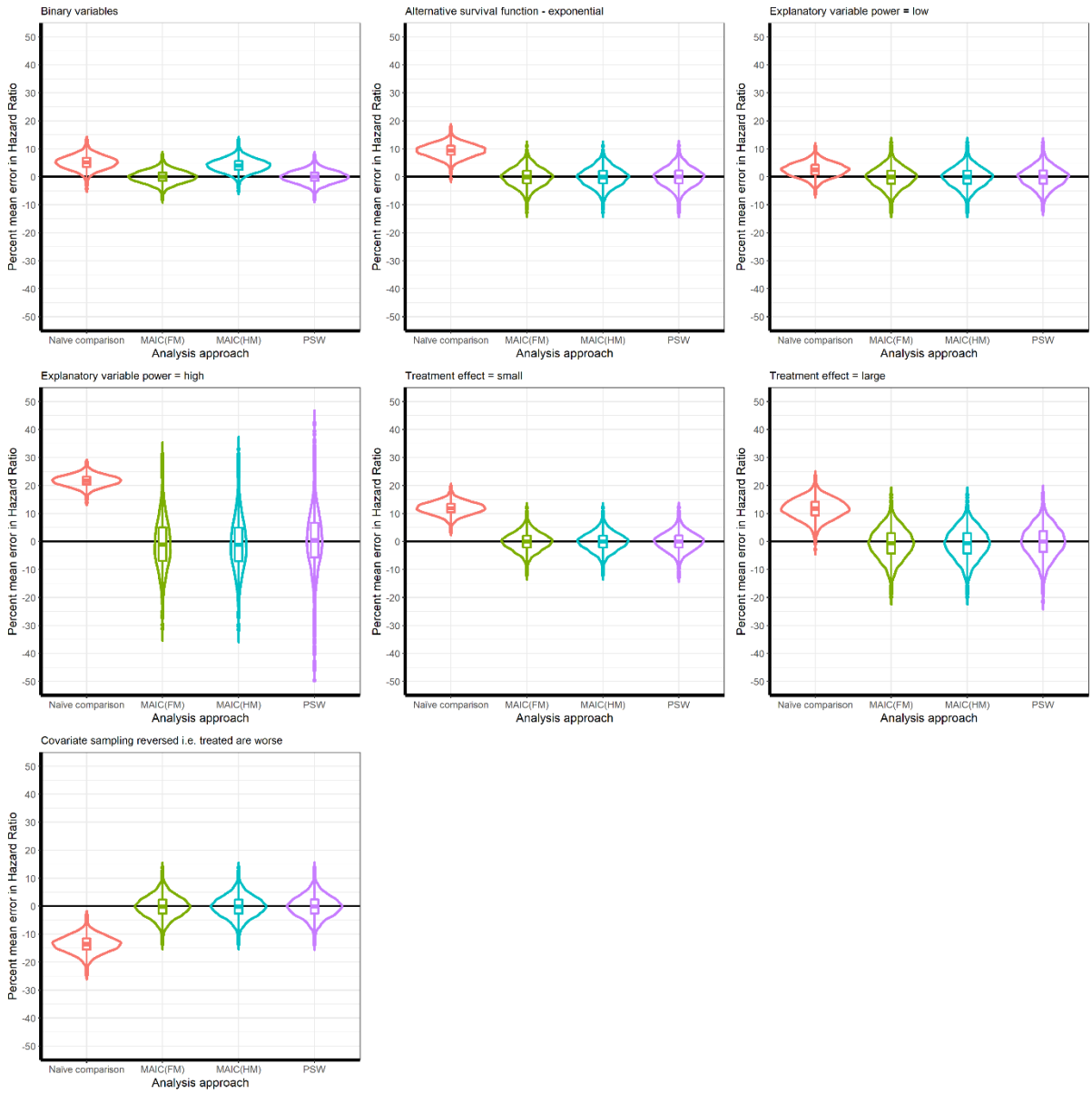
Table 4-3: Tabulated results of MAIC scenario analysis varying the setup of the simulation study, 5,000 runs

Method	Mean Percentage Error (MCSE)	Absolute Percentage Error (MCSE)	Mean Standard Error	Coverage probability	Percent of scenarios worse than a naïve comparison
Base case					
Naïve comparison	11.8% (<0.01)	11.8% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.2% (<0.01)	2.6% (<0.01)	0.03	95%	2%
$MAIC_{HM}$	-0.2% (<0.01)	2.6% (<0.01)	0.03	95%	2%
PSW	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%
All variables are binary					
Naïve comparison	5.2% (<0.01)	5.2% (<0.01)	0.03	48%	-
$MAIC_{FM}$	0% (<0.01)	1.8% (<0.01)	0.03	98%	12%
$MAIC_{HM}$	4.1% (<0.01)	4.2% (<0.01)	0.03	63%	4%
PSW	0% (<0.01)	1.8% (<0.01)	0.03	98%	12%
Exponential distribution used as the survival function					
Naïve comparison	9.4% (<0.01)	9.4% (<0.01)	0.03	3%	-
$MAIC_{FM}$	-0.1% (<0.01)	2.6% (<0.01)	0.03	95%	4%
$MAIC_{HM}$	-0.1% (<0.01)	2.6% (<0.01)	0.03	95%	4%
PSW	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	4%
Explanatory variable power is low					
Naïve comparison	2.5% (<0.01)	2.9% (<0.01)	0.03	84%	-
$MAIC_{FM}$	-0.1% (<0.01)	2.8% (<0.01)	0.04	95%	43%
$MAIC_{HM}$	-0.1% (<0.01)	2.8% (<0.01)	0.04	95%	43%
PSW	-0.1% (<0.01)	2.8% (<0.01)	0.04	95%	44%
Explanatory variable power is high					
Naïve comparison	21.7% (<0.01)	21.7% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.9% (<0.01)	7.1% (<0.01)	0.08	94%	2%
$MAIC_{HM}$	-1% (<0.01)	7.1% (<0.01)	0.08	94%	2%
PSW	0.2% (<0.01)	7.7% (<0.01)	0.09	94%	4%
Treatment effect is low (0.9 hazard ratio)					
Naïve comparison	11.9% (<0.01)	11.9% (<0.01)	0.03	0%	-
$MAIC_{FM}$	0% (<0.01)	2.5% (<0.01)	0.03	95%	1%
$MAIC_{HM}$	0% (<0.01)	2.5% (<0.01)	0.03	95%	1%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	95%	1%
Treatment effect is high (0.2 hazard ratio)					
Naïve comparison	11.7% (<0.01)	11.7% (<0.01)	0.04	10%	-
$MAIC_{FM}$	-0.8% (<0.01)	4.3% (<0.01)	0.05	94%	10%
$MAIC_{HM}$	-0.8% (<0.01)	4.3% (<0.01)	0.05	94%	10%
PSW	-0.1% (<0.01)	4.4% (<0.01)	0.05	94%	9%
Covariate sampling is reversed i.e. Population A are worse by 0.5SD					
Naïve comparison	-13.6% (<0.01)	13.6% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.2% (<0.01)	3% (<0.01)	0.04	95%	1%

$MAIC_{HM}$	-0.2% (<0.01)	3% (<0.01)	0.04	95%	1%
PSW	-0.1% (<0.01)	3% (<0.01)	0.04	95%	1%

MCSE = Monte Carlo Standard Error, MAIC = Matching Adjusted Indirect Comparison, MM = Method of moments, HM = includes Higher moments, PSW = Propensity Score Weighting

Figure 4-4: Violin plot (with overlaid box plots) of the percent mean error in 5,000 runs of the MAIC simulation study base case



When testing MAIC under different conditions, some interesting results can be noted – in particular that whilst the inclusion of all nuisance parameters that have no bearing on the outcome does (predictably) often lead to errors. However if even half of the variables used for weighting do influence outcomes, the approach remains accurate. Also reassuring is that correlated variables in the analysis not only performed well, but that the method performed better than in the base case.

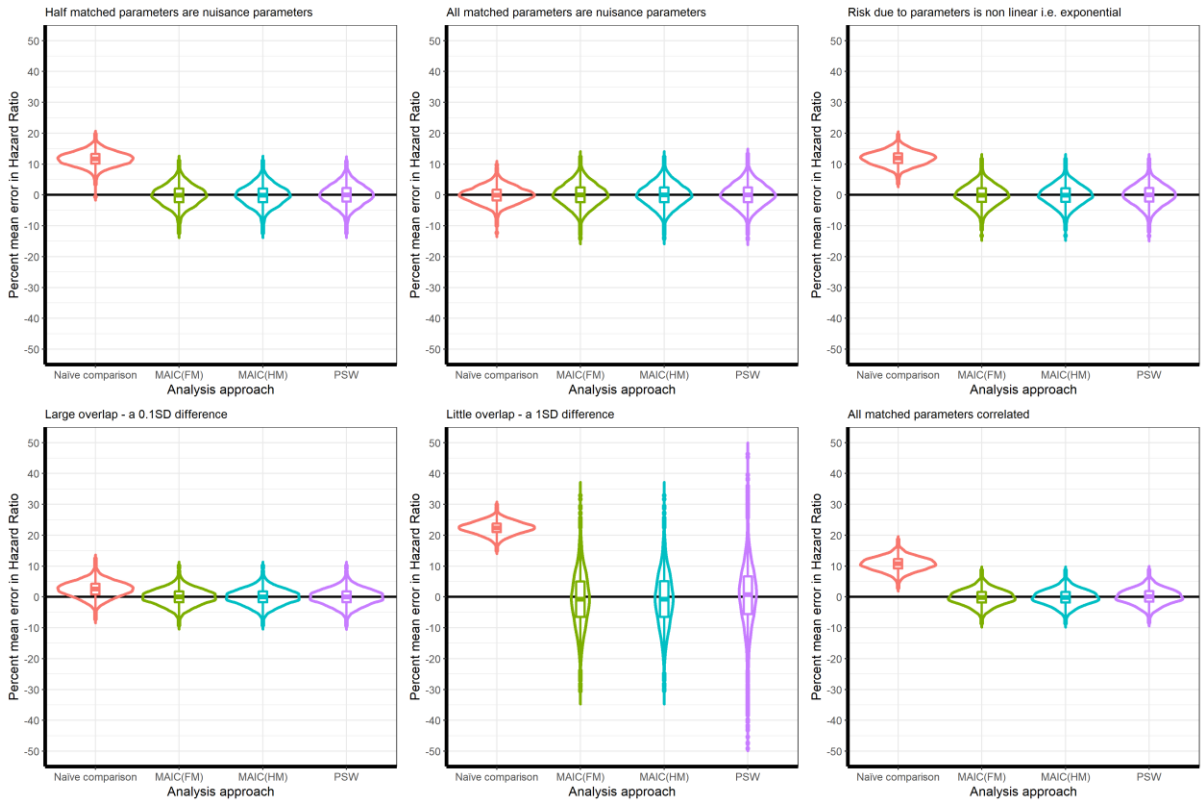
What should also be noted is the importance of overlap between the datasets – where the overlap is especially close (0.1SD was used in the simulation), a non-trivial number of simulations (circa 30%) resulted in an estimated HR more inaccurate than a naïve

comparison – although the method was unbiased, it did introduce a level of uncertainty. Conversely where the overlap was low (a mean of a 1SD difference between datasets) the method again remained unbiased and a big improvement from a naïve comparison (shown by the coverage probability and the percent of scenarios worse than a naïve comparison). In this scenario although the mean percentage error was near zero, the absolute error was substantially inflated – showing the importance of reasonably similar groups.

Table 4-4: Tabulated results of MAIC scenario analysis varying the conditions of the study, 5,000 runs

Method	Mean Percentage Error (MCSE)	Absolute Percentage Error (MCSE)	Mean Standard Error	Coverage probability	Percent of scenarios worse than a naïve comparison
Half the matched parameters are nuisance parameters					
Naïve comparison	11.7% (<0.01)	11.7% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.1% (<0.01)	2.6% (<0.01)	0.03	96%	2%
$MAIC_{HM}$	-0.1% (<0.01)	2.6% (<0.01)	0.03	96%	2%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	95%	2%
All the matched parameters are nuisance parameters					
Naïve comparison	0% (<0.01)	2.1% (<0.01)	0.03	95%	-
$MAIC_{FM}$	0% (<0.01)	2.9% (<0.01)	0.04	95%	64%
$MAIC_{HM}$	0% (<0.01)	2.9% (<0.01)	0.04	95%	64%
PSW	0% (<0.01)	2.9% (<0.01)	0.04	95%	64%
The effect of parameters is non-linear					
Naïve comparison	11.9% (<0.01)	11.9% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.1% (<0.01)	2.6% (<0.01)	0.03	96%	1%
$MAIC_{HM}$	-0.1% (<0.01)	2.6% (<0.01)	0.03	96%	1%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	96%	1%
Small difference is covariate sampling (0.1SD)					
Naïve comparison	2.5% (<0.01)	3% (<0.01)	0.03	84%	-
$MAIC_{FM}$	0% (<0.01)	2.1% (<0.01)	0.03	95%	31%
$MAIC_{HM}$	0% (<0.01)	2.1% (<0.01)	0.03	95%	31%
PSW	0% (<0.01)	2.1% (<0.01)	0.03	95%	31%
Large difference in covariate sampling (1SD)					
Naïve comparison	22.4% (<0.01)	22.4% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.7% (<0.01)	6.9% (<0.01)	0.08	95%	2%
$MAIC_{HM}$	-0.7% (<0.01)	6.9% (<0.01)	0.08	95%	2%
PSW	0.2% (<0.01)	7.7% (<0.01)	0.09	94%	4%
All parameters correlated					
Naïve comparison	10.7% (<0.01)	10.7% (<0.01)	0.03	1%	-
$MAIC_{FM}$	-0.1% (<0.01)	2% (<0.01)	0.03	96%	1%
$MAIC_{HM}$	-0.1% (<0.01)	2% (<0.01)	0.03	96%	1%
PSW	0.1% (<0.01)	2% (<0.01)	0.03	96%	1%
MCSE = Monte Carlo Standard Error, MAIC = Matching Adjusted Indirect Comparison, FM = Method of moments, HM = includes Higher moments, PSW = Propensity Score Weighting					

Figure 4-5: Violin plot (with overlaid box plots) of the percent mean error in 5,000 runs of the MAIC varying the conditions of the study



Where the assumptions underpinning MAIC are violated, the method (in many cases) performs surprisingly well. As would be expected if an important (and uncorrelated) variable is omitted from analyses, this was not accounted for in weighting, and estimates remained biased – this underlines the importance of the inclusion of all prognostic variables in datasets. Where this missing variable was correlated with observed variables, though the results were more inaccurate than the base case (higher absolute percentage error and lower confidence intervals), the approach was then unbiased.

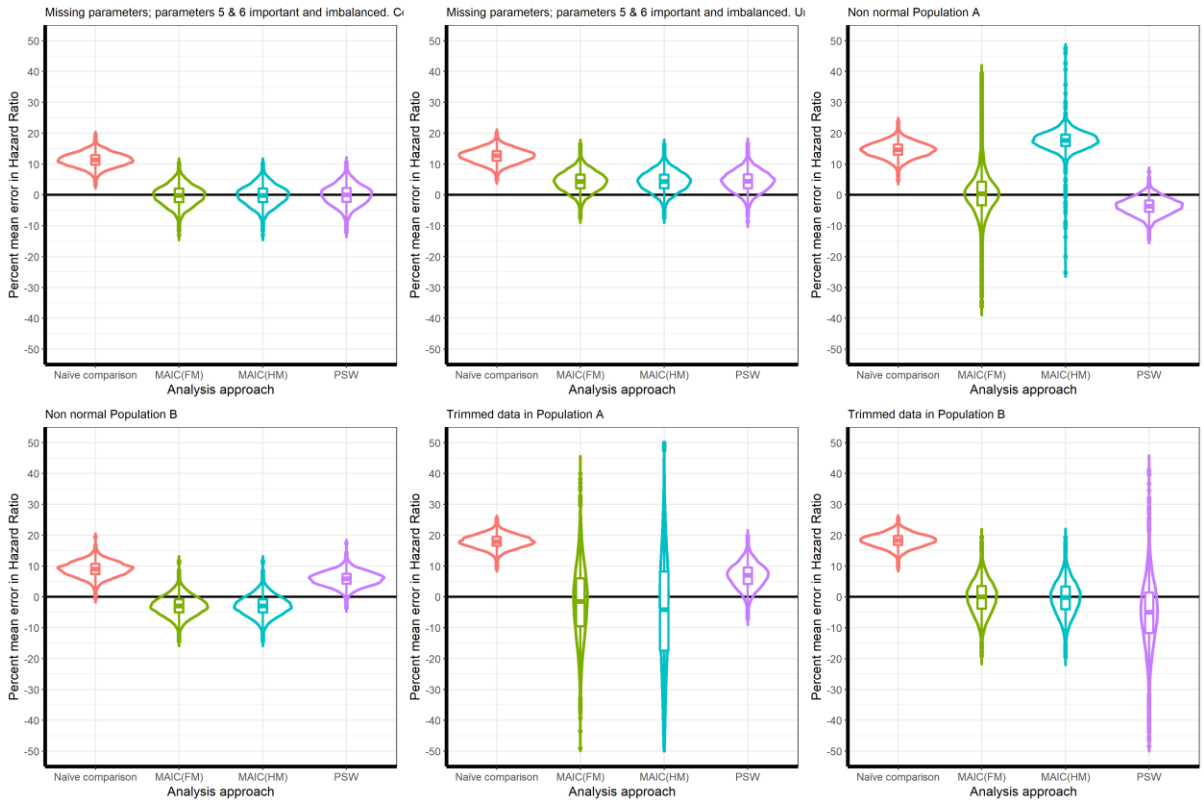
Where problems did become apparent was in differences between datasets related to inclusion criteria; in particular with $MAIC_{HM}$. The inclusion of patients for who had a probability of zero of being in the opposite dataset appears challenging - these have an impact on the distribution of patients, and thus on the higher moments which can be matched. Similarly non-normal distributions appear to be a problem for $MAIC_{HM}$, which implicitly assumes normality in moments (unless even higher moments such as kurtosis are available to match to).

Table 4-5: Tabulated results of MAIC scenario analysis with scenarios that violate assumptions implicit or explicit in MAIC, 5,000 runs

Method	Mean Percentage Error (MCSE)	Absolute Percentage Error (MCSE)	Mean Standard Error	Coverage probability	Percent of scenarios worse than a naïve comparison
Missing parameters correlated with observed parameters					
Naïve comparison	11.3% (<0.01)	11.3% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.2% (<0.01)	2.6% (<0.01)	0.03	96%	2%
$MAIC_{HM}$	-0.2% (<0.01)	2.6% (<0.01)	0.03	96%	2%
PSW	0% (<0.01)	2.6% (<0.01)	0.03	96%	2%
Missing parameters uncorrelated with observed parameters					
Naïve comparison	12.6% (<0.01)	12.6% (<0.01)	0.03	0%	-
$MAIC_{FM}$	4.3% (<0.01)	4.6% (<0.01)	0.03	75%	0%
$MAIC_{HM}$	4.3% (<0.01)	4.6% (<0.01)	0.03	75%	0%
PSW	4.4% (<0.01)	4.7% (<0.01)	0.03	74%	0%
Non-normal distributions sampled in Population A					
Naïve comparison	14.6% (<0.01)	14.6% (<0.01)	0.03	0%	-
$MAIC_{FM}$	1% (<0.01)	6.4% (<0.01)	0.03	63%	12%
$MAIC_{HM}$	17.6% (<0.01)	17.7% (<0.01)	0.03	1%	99%
PSW	-3.7% (<0.01)	3.9% (<0.01)	0.03	73%	1%
Non-normal distributions sampled in Population B					
Naïve comparison	9% (<0.01)	9% (<0.01)	0.03	6%	-
$MAIC_{FM}$	-2.9% (<0.01)	3.5% (<0.01)	0.03	88%	12%
$MAIC_{HM}$	-2.9% (<0.01)	3.5% (<0.01)	0.03	88%	12%
PSW	5.9% (<0.01)	5.9% (<0.01)	0.03	38%	0%
Trimmed patient characteristics in Population A (no poor performers)					
Naïve comparison	18% (<0.01)	18% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-1.8% (<0.01)	9.3% (<0.01)	0.11	92%	14%
$MAIC_{HM}$	-7.6% (<0.01)	17.8% (<0.01)	0.18	90%	38%
PSW	6.8% (<0.01)	7% (<0.01)	0.04	60%	0%
Trimmed patient characteristics in Population B (no good performers)					
Naïve comparison	18.3% (<0.01)	18.3% (<0.01)	0.03	0%	-
$MAIC_{FM}$	-0.1% (<0.01)	4.3% (<0.01)	0.05	95%	0%
$MAIC_{HM}$	-0.3% (<0.01)	4.3% (<0.01)	0.05	95%	0%
PSW	-5.5% (<0.01)	9.3% (<0.01)	0.09	87%	13%

MCSE = Monte Carlo Standard Error, MAIC = Matching Adjusted Indirect Comparison, FM = Method of moments, HM = includes Higher moments, PSW = Propensity Score Weighting

Figure 4-6: Violin plot (with overlaid box plots) of the percent mean error in 5,000 runs of the MAIC simulation study with scenarios that violate assumptions implicit or explicit in MAIC



4.8 FINDINGS FROM VARYING THE NUMBER OF PATIENTS AVAILABLE IN EACH DATASET

When the number of patients in each population is varied, it can be seen that with a relatively low number, in the context of the simulation study, the results of MAIC are inaccurate. As sample sizes in each population increase the level of error decreases - whilst no conclusions can be made about the required numbers for MAIC in the context of real world examples, a degree of caution should be used with matching to (or with) 'low' numbers.

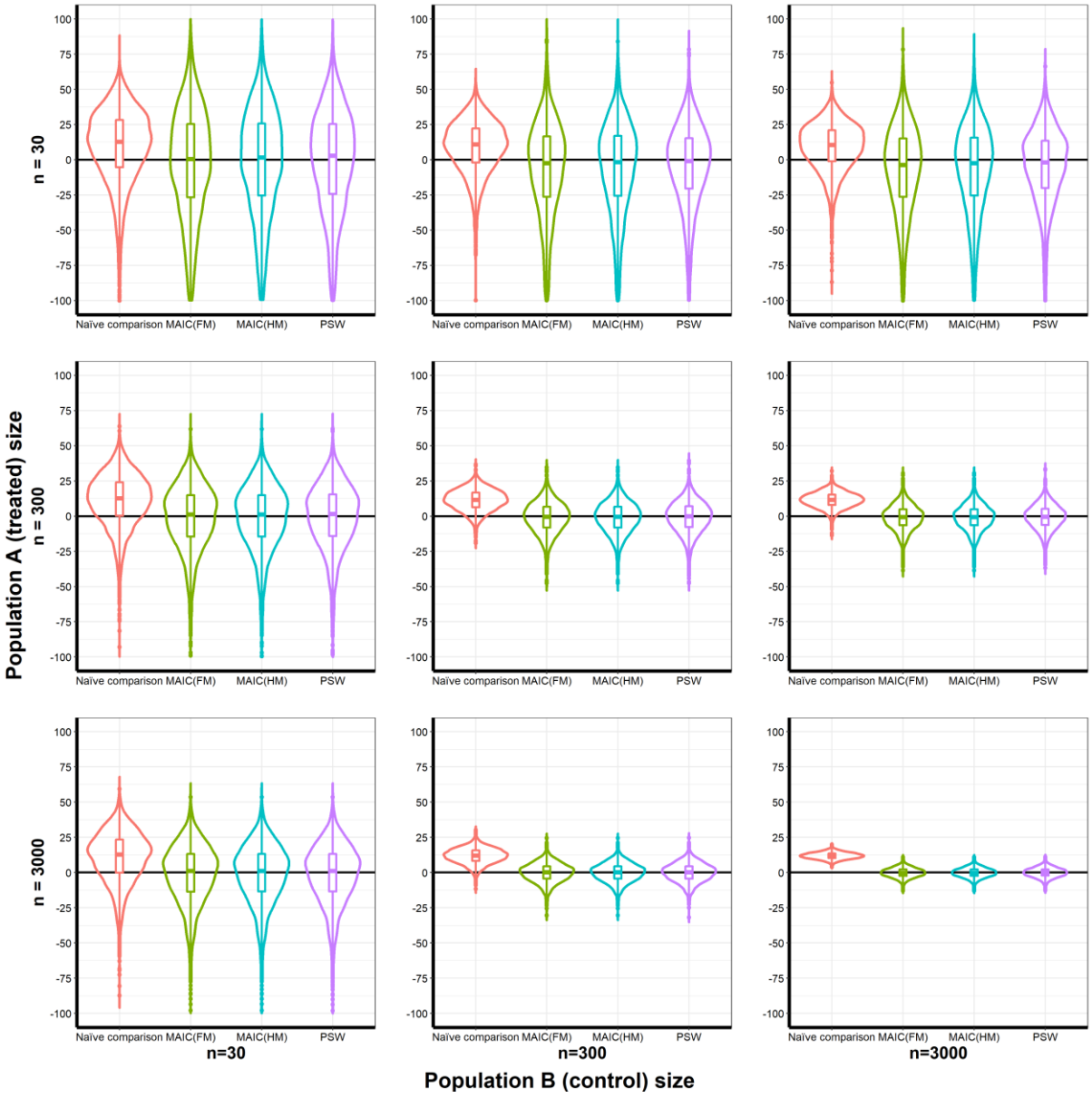
The relevant finding from this collection of scenarios however would appear to be that it is more important to have more patients to weight with (Population A) than to match to (Population B) – seen by comparing the results in Table 4-6, and visually inspecting the off-diagonals in Figure 4-7. This makes sense when considered, as the means to be matched to should stabilise relatively quickly, whereas the variability between patients means that more are needed (especially with multiple characteristics) to be able to adequately match even well established means.

Table 4-6: Tabulated results of MAIC scenario analysis varying the number of patients available in Population A and Population B, 5,000 runs

Method	Mean Percentage Error	Absolute Percentage Error	Mean Standard Error	Coverage probability	Percent of scenarios worse than a naïve comparison
Population A = 30, Population B = 30					
Naïve comparison	8.9% (<0.01)	23.1% (<0.01)	0.27	90%	-
$MAIC_{MM}$	-15.3% (0.012)	41.5% (0.011)	0.37	88%	62%
$MAIC_{HM}$	-12% (<0.01)	38.7% (<0.01)	0.36	88%	61%
PSW	-4.4% (<0.01)	31.4% (<0.01)	0.33	90%	60%
Population A = 30, Population B = 300					
Naïve comparison	8.8% (<0.01)	17.1% (<0.01)	0.19	90%	-
$MAIC_{MM}$	-12.5% (<0.01)	31.1% (<0.01)	0.28	86%	63%
$MAIC_{HM}$	-12% (<0.01)	30.8% (<0.01)	0.27	86%	62%
PSW	-6% (<0.01)	23.6% (<0.01)	0.24	89%	56%
Population A = 30, Population B = 3000					
Naïve comparison	8.9% (<0.01)	15.8% (<0.01)	0.18	89%	-
$MAIC_{MM}$	-11.9% (<0.01)	28.8% (<0.01)	0.25	86%	63%
$MAIC_{HM}$	-12.2% (<0.01)	29.4% (<0.01)	0.25	85%	62%
PSW	-6.1% (<0.01)	21.7% (<0.01)	0.22	90%	56%
Population A = 300, Population B = 30					
Naïve comparison	10.9% (<0.01)	18.1% (<0.01)	0.19	85%	-
$MAIC_{MM}$	-1.2% (<0.01)	18% (<0.01)	0.21	92%	41%
$MAIC_{HM}$	-1.2% (<0.01)	18.1% (<0.01)	0.21	92%	41%
PSW	-0.9% (<0.01)	18.1% (<0.01)	0.21	91%	41%
Population A = 300, Population B = 300					
Naïve comparison	11.4% (<0.01)	12% (<0.01)	0.08	68%	-
$MAIC_{MM}$	-1% (<0.01)	8.7% (<0.01)	0.11	95%	30%
$MAIC_{HM}$	-1% (<0.01)	8.7% (<0.01)	0.11	95%	30%
PSW	-0.7% (<0.01)	8.8% (<0.01)	0.11	94%	30%
Population A = 300, Population B = 3000					
Naïve comparison	11.4% (<0.01)	11.6% (<0.01)	0.06	47%	-
$MAIC_{MM}$	-0.9% (<0.01)	6.9% (<0.01)	0.09	94%	23%
$MAIC_{HM}$	-0.9% (<0.01)	6.9% (<0.01)	0.09	94%	23%
PSW	-0.7% (<0.01)	6.9% (<0.01)	0.09	94%	23%
Population A = 3000, Population B = 30					
Naïve comparison	10.7% (<0.01)	17.6% (<0.01)	0.18	83%	-
$MAIC_{MM}$	-1.4% (<0.01)	16.4% (<0.01)	0.18	90%	37%
$MAIC_{HM}$	-1.4% (<0.01)	16.4% (<0.01)	0.18	90%	37%
PSW	-1.4% (<0.01)	16.4% (<0.01)	0.18	90%	37%
Population A = 3000, Population B = 300					
Naïve comparison	11.8% (<0.01)	11.9% (<0.01)	0.06	45%	-

$MAIC_{MM}$	-0.1% (<0.01)	5.3% (<0.01)	0.07	94%	16%
$MAIC_{HM}$	-0.1% (<0.01)	5.3% (<0.01)	0.07	94%	16%
PSW	-0.1% (<0.01)	5.3% (<0.01)	0.07	94%	16%
Population A = 3000, Population B = 3000					
Naïve comparison	11.9% (<0.01)	11.9% (<0.01)	0.03	0%	-
$MAIC_{MM}$	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%
$MAIC_{HM}$	-0.1% (<0.01)	2.7% (<0.01)	0.03	95%	2%
PSW	0% (<0.01)	2.7% (<0.01)	0.03	95%	2%

Figure 4-7: Violin plot (with overlaid box plots) of the percent mean error in 5,000 runs of the MAIC simulation study varying the number of patients included in Population A and Population B



4.9 DISCUSSION ON THE MERITS AND APPROACH TO MAIC

Based on the results of the simulation study, MAIC would seem to be a reasonable approach to address between study differences. In the majority of cases it is unbiased, and for the most part, accurate. In terms of the approach used however, it is difficult to recommend the use of MAIC including matching on higher moments – in no cases does it perform appreciably better than $MAIC_{FM}$, however in many instances it does exhibit severe problems. These instances include scenarios extremely common in pharmaceutical clinical trials, such as truncated distributions, and non-normal distributions of patient characteristics.

Despite using on half the individual level data of PSW, MAIC did manage to perform similarly. The advantage of PSW approaches, however, was not truly utilised in this simulation study, for instance in having the ability to trim the data to align characteristics, and also the ability to choose which population to generate an estimate in. Despite these limitations, the inclusion of PSW does demonstrate that the performance of MAIC is acceptable, albeit in an artificial situation.

Whilst on the whole MAIC did perform well, caution should be noted based on scenario analyses under certain circumstances. Where there are limited patient numbers, poor overlap (or near perfect overlap) between studies, or highly trimmed patient characteristics, MAIC should be used subject to appropriate caveats. Also a reasonable interpretation is that variables to be included in weighting methods should also be plausibly linked to outcomes, similar to the approach for propensity score creation. Equally, for the results to be relied upon, any unobserved variables which are known to be important should be, as a minimum, be correlated with the observed patient characteristics. The results of scenario analyses from this study also indicate that if an intervention has a very large effect size, MAIC may indeed not be needed, and may introduce more bias than it resolves – particularly if linked with other issues (such as low sample size). In such circumstances it may be better to let the ‘obviousness’ of results speak for themselves, as opposed to present an analysis that is likely to be highly susceptible to the role of chance. The same would seem to apply if the studies have similar patient characteristics, where weighting may introduce more bias than it eliminates.

4.10 THE USE OF MATCHING ADJUSTED INDIRECT COMPARISON OUTSIDE OF TIME TO EVENT OUTCOMES

Although this simulation has been rooted in a single time to event outcome, the results should generalise to other outcomes for instance multiple time to event outcomes (i.e. time on treatment, progression free survival, and overall survival), response rates, or categorical

outcomes. The reason for this is that the technique of MAIC is performed only on baseline characteristics, with weights then produced for each patient to match between the datasets. Using these weights multiple outcomes could be calculated for each patient – unlike STC outcomes are not included in the model.

The structure employed in the simulation study however may need to change should other data be simulated; the use of a CPH based model allowed a convenient metric for the (in)accuracy in estimates. If others were to replicate this study in other settings using a different outcome model - for instance a non-proportional hazards survival model or alternative endpoint, a different metric would be required to judge the success of the method.

4.11 SUMMARY OF FINDINGS

The work in this chapter demonstrates that although the limitations of MAIC should be noted, on the whole it does represent a substantial improvement on the use of naive comparisons.

The main area of focus for analysts should be to ensure the data available (and the associated structure) meets the basic requirements for the technique, and the problem is carefully considered. Provided these criteria are broadly met however, the method is both accurate and unbiased when matching on first moments. Due to the poor performance of matching on higher moments however (with few, if any advantages) this method cannot be recommended.

5 NOVEL METHODS FOR THE CREATION OF HISTORICAL CONTROLS

In addition to the use of MAIC to account for differences between studies, an additional need I identified in my review of models (Section 3.2) was the complexity in estimating comparative effectiveness in the absence of historical data to act as a point of reference. In this chapter I propose three methods for the estimation of counterfactual outcomes based on historical data.

5.1 THE USE OF NON-RESPONDERS AS A CONTROL ARM

The method described in this section (5.1) was published in Cost Effectiveness and Resource Allocation (Hatswell *et al.*, 2017)

When treatments are licensed on the basis of uncontrolled study data, this is often linked to the response rate seen in trials (indeed early stage oncology trials often use a 20% response rate as a benchmark) – mirrored in ICH guidance (ICH Harmonised Tripartite, 2000). The assumption implicit in this is that a response (usually defined as improvement in disease measurement by a given percentage) would lead to better outcomes. By extension this assumes a non-responder had no appreciable benefit from the intervention.

In such instances - where patients can be neatly divided into groups of responders and non-responders, it may be possible to use the outcomes need in non-responders as a proxy for the outcomes that would have been seen in the absence of treatment. There are a number of assumptions inherent in the approach which if not met however could introduce bias (some in favour of the intervention, and some acting against it).

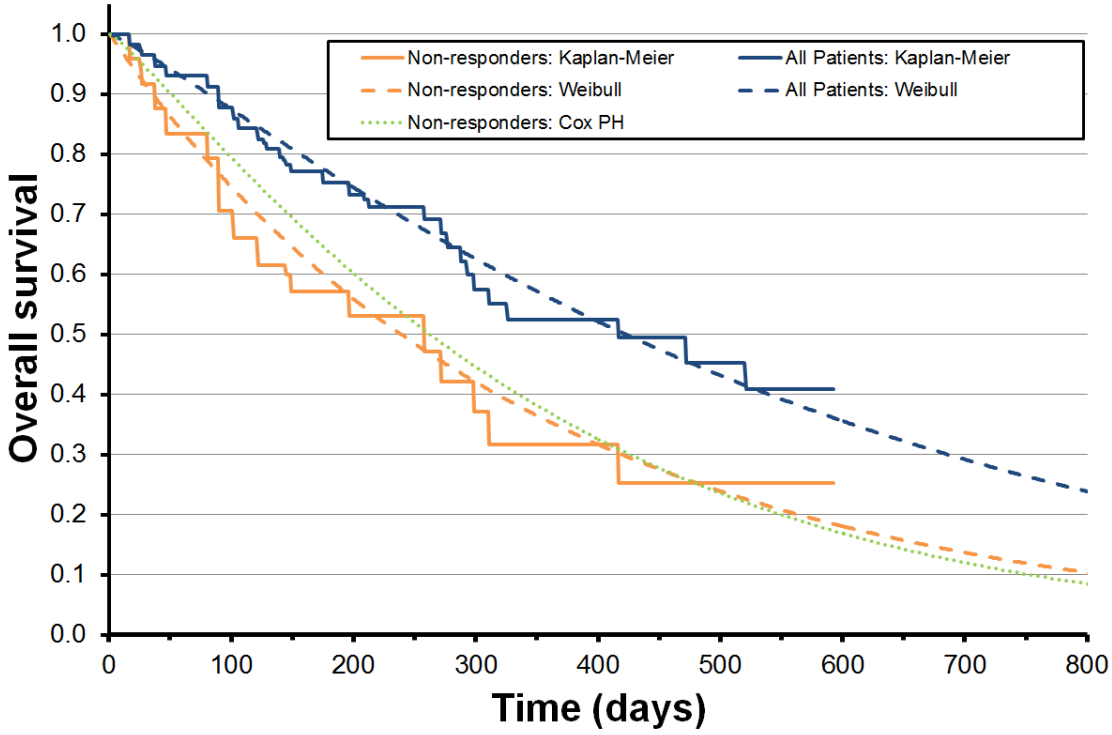
If this approach is taken, should some patient characteristics not just be predictive of treatment success, but also prognostic (predictive of outcomes), it may be the case that patients who achieved a response would have performed regardless (and thus introduce a bias in the analysis in favour of the intervention). Similarly, it may be that patients who did not receive a response not only failed to benefit from treatment, but also experienced side effects from the intervention that worsened their outcomes. The approach however may bias against an intervention if other conditions are met, for instance if non-responders did receive a benefit (for instance in holding their disease stable) which did not reach the threshold for a clinical benefit.

The utility of the approach can be demonstrated using the Hx-CD20-406 Study for the drug ofatumumab for the treatment of double refractory chronic lymphocytic leukaemia. In the study, 59 patients received treatment with ofatumumab, with an Overall Response Rate (ORR) of 58% and median OS of 13.7 months being sufficient for regulators to grant a

license (Wierda *et al.*, 2010). As expected, the outcomes for non-responders were substantially worse than for responders; their median OS was 9.8 months, whereas this was not reached in responders.

In order to estimate the comparative effectiveness of ofatumumab versus no treatment (using non-responders as a proxy for this), a comparison was made between all patients and the non-responders. All patients were used in the comparison (and not simply responders), as it is not possible to identify a priori which patients would have a response to treatment. To perform this modelling parametric survival curves were fitted to the data, with the best fit (according to AIC and visual inspection) being a Weibull model, with a Cox Proportional Hazards (CPH) model used to estimate outcomes in non-responders (Figure 5-1) including the effects of number of age, sex, Rai (disease) stage, ECOG performance status, number of prior therapies, years since diagnosis, and prognostic chromosomal deletions (11p and 17q). The effect of this was to reduce the unadjusted Hazard Ratio from 0.49, to an adjusted hazard ratio of 0.53.

Figure 5-1: Estimated overall survival in all patients and non-responders for ofatumumab in double refractory Chronic Lymphocytic Leukaemia



Using this approach, ofatumumab was estimated to produce a gain of 0.550 LYs (1.494 vs 0.945) using the CPH model. If instead independent curve fits were specified, this benefit reduced to 0.542 LYs (1.494 vs 0.952). The approach proposed is subject to a number of

important assumptions (not all of which are fully testable), however may allow the production of estimates where no suitable historical data can be identified.

5.2 THE USE OF EXTRAPOLATION TO CREATE A HISTORICAL CONTROL

Beyond the use of external data, it is also apparent that further use could be made of data from prior lines of therapy – the two methods proposed below allow for such methods to be used to create historical control arms.

In order to aid the implementation of these methods, example R code (using simulated data) is presented in Appendix D.

5.2.1 EXTRAPOLATION OF DATA FROM THE PREVIOUS LINE OF TREATMENT TO ESTIMATE THE COUNTERFACTUAL

The methods approach described in this section (5.2.1) was published in Statistical Methods in Medical Research (Hatswell & Sullivan, 2019)

In some disease areas, patients receive repeated treatment lines to control their disease, beginning with evidence-based treatment lines, and once all licensed treatment options are exhausted (either through non-response, intolerance, or unavailability in the patient's region), they are treated with off label and experimental treatments. Despite the lack of evidence, such treatments may appear in treatment guidelines with a low evidence rating, due to the desire of physicians to control disease symptoms or extend life. Examples of diseases like this are typically found in haematology, such as multiple myeloma, Chronic Lymphocytic Leukaemia, and Non-Hodgkin's Lymphoma.

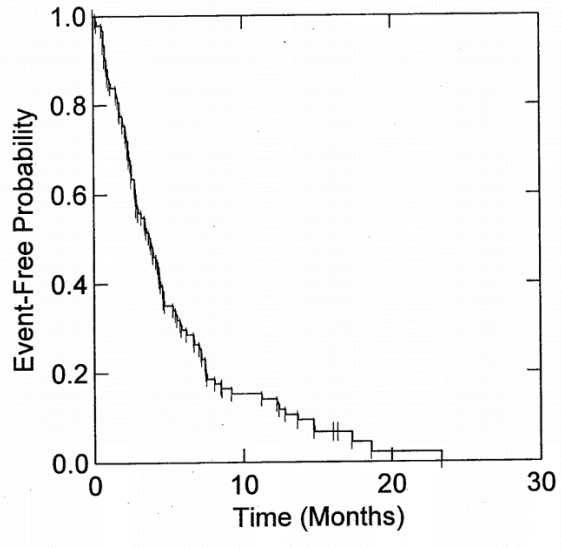
Where a patient has received all evidence-based treatments, the variety in treatment strategies is likely to increase dramatically as there is unlikely to be a consensus regarding the appropriate care. In this environment the acceptability of approvals without a controlled clinical trial increases as there is unlikely to be a consensus around what constitutes 'standard of care' against which a drug can be assessed (alternative trial designs may include a 'physician's choice' arm). In such circumstances, the lack of a defined standard of care adds an extra complexity in estimating the counterfactual outcomes – evidence is by definition sparse, and even if it is possible to identify historical control information for a specific treatment, it is uncertain how many will receive each.

When faced with such evidence (an uncontrolled study with uncertain counterfactual outcomes), one promising area that could be used to estimate outcomes, is the trial of the previous line of therapy. Whilst treatments may be licensed where there is no standard of

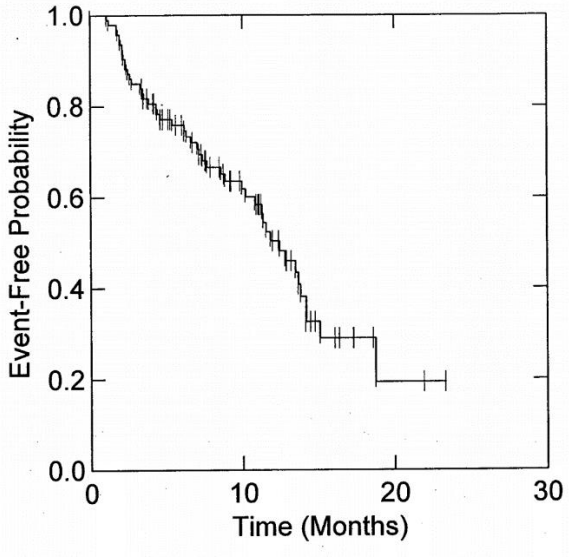
care, typically the previous treatment line is the last data available. This can be seen with the example of ofatumumab, which was licensed for the treatment of double refractory Chronic Lymphocytic Leukaemia (DR-CLL) – all patients being refractory to treatment with first line fludarabine, and subsequent alemtuzumab. In the case of alemtuzumab, the main evidence for its use was taken from the CAM211 study (Keating *et al.*, 2002), which enrolled 93 patients with CLL refractory to fludarabine treatment. This uncontrolled study had a primary endpoint of Disease Free Survival (DFS), with a secondary outcome of Overall Survival, the results of which are shown in Figure 5-2; having been sourced from the FDA review of the product (Food and Drug Administration, 2001).

Figure 5-2: Progression free survival and overall survival from the CAM211 study in refractory Chronic Lymphocytic Leukaemia extracted from the FDA review (Food and Drug Administration, 2001)

Disease Free Survival:



Overall survival:

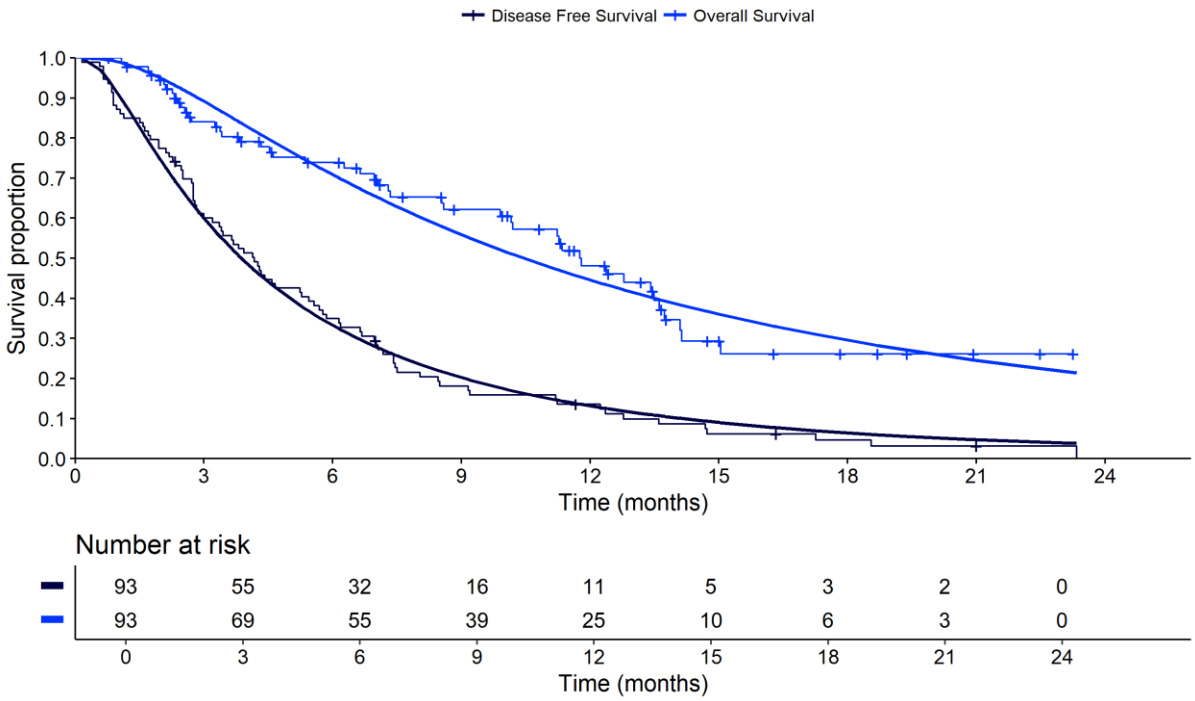


Whilst this evidence was designed to support the approval of alemtuzumab, it does however offer insight into the subsequent outcomes of patients after becoming refractory to alemtuzumab and having their disease return i.e. DR-CLL. This evidence can be seen in the gap between OS and DFS, if plotted together – in the period post DFS but before death, a patient is living with DFS, and being treated with standard of care. Whilst incomplete (patients are not followed until death), it is possible to estimate the outcomes over time using extrapolation, as is commonly used in health technology appraisal.

To estimate the period between death, I digitised the Kaplan-Meir data using the method of Guyot et al. (2012). Parametric curves were fitted to the disease-free, and overall survival

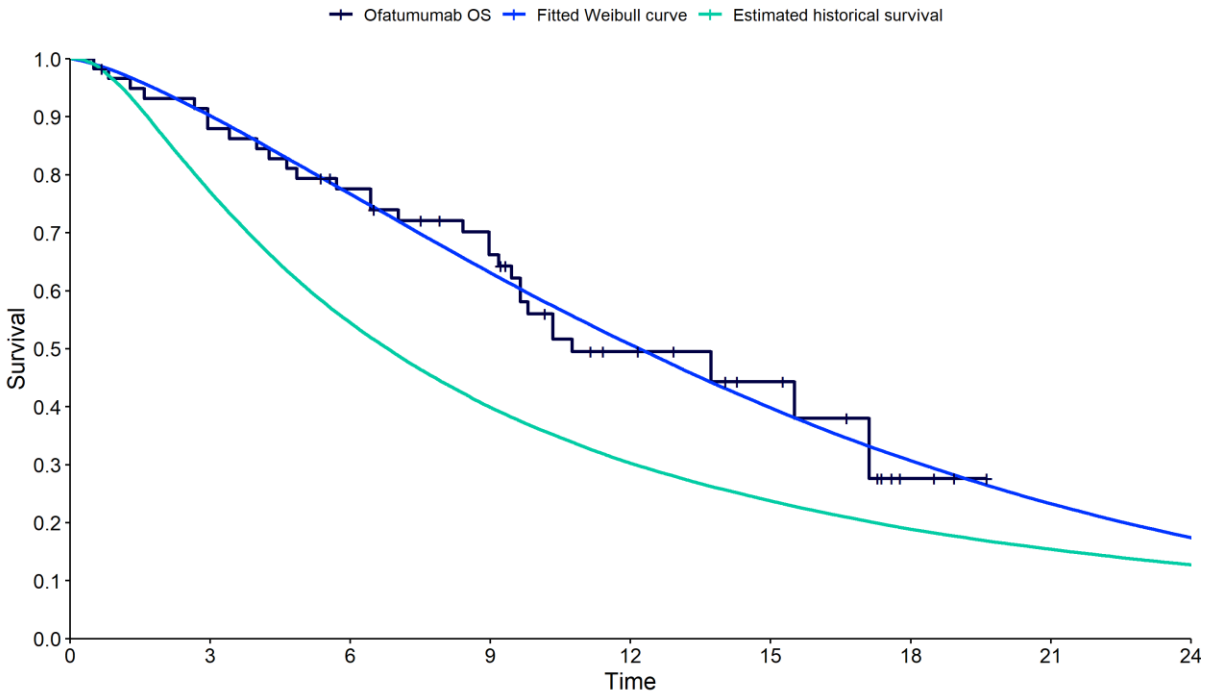
curves, selecting the most appropriate fit according to the guidelines published by Latimer (2011). A lognormal curve fit was found to be the best fitting curve based on AIC, BIC and visual inspection – as shown in Figure 5-3.

Figure 5-3: Recreated Disease-Free Survival and Overall Survival from the CAM211 study, with fitted (lognormal) parametric curve fits overlaid



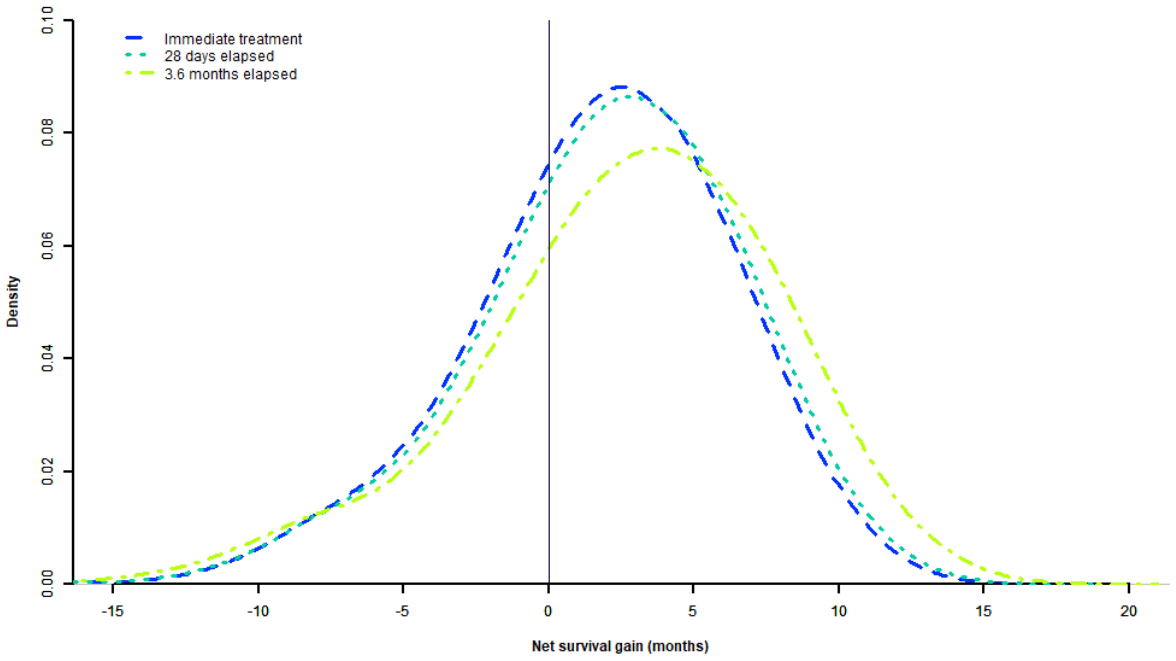
The resulting difference between DFS and OS was then estimated. To account for uncertainty in the curve fits, 100,000 bootstrap samples were taken of the curve fit parameters, with patients exhibiting a mean of 13 months survival post DFS with alemtuzumab. This estimate can then be used to compare with the survival of patients treated with ofatumumab obtained through a similar digitisation of data from the HX-C20-406 study (Wierda *et al.*, 2010) to which a Weibull curve was fitted (shown in Figure 5-4)

Figure 5-4: Digitized ofatumumab Kaplan-Meier survival data from Hx-CD20-406 and Weibull curve fit, plotted against survival post disease progression from the CAM211 study



From this analysis, the overall survival gain of ofatumumab can be estimated, compared to our outcomes from a time prior to the availability ofatumumab, when by definition patients could only receive standard of care. Exactly how large the gain in OS is however, depends on the assumption used for how long elapsed between disease recurrence, and treatment with ofatumumab. In the protocol for the ofatumumab study, patients must have been treatment free for 28 days prior to beginning the study, with the mean patient having been treatment free for 3.6 months. The range of likely outcomes (calculated via 100,000 bootstraps) is therefore shown in Figure 5-5. Using the approach (and accounting for uncertainty in fitted curves) if a 3.6-month delay in subsequent treatment is assumed (as seems most plausible), ofatumumab is estimated to increase survival in 72% of simulations and provide a mean net survival gain of 2.7 months. Also presented is a curve showing if each patient had only been disease free for the minimum interval required by the study protocol (28 days).

Figure 5-5: Density plot of estimated mean survival gain of ofatumumab over historical control, using assumptions of immediate treatment with ofatumumab, 28-day delay, and 3.6-month delay



Whilst this method may allow for estimation of counterfactual outcomes where data are otherwise unavailable, it does however have several limitations. The first of these is the obvious need for data to be available at the previous line (in this instance fortunately the relevant time to event information was presented in the FDA review, as it was not in the trial publication). Where a treatment is for newly diagnosed patients for instance, this approach will not be possible to implement. Other assumptions are around the comparability of patient populations; whilst this may be possible to assess the overlap between inclusion criteria of studies, these may not remain the same at study exit of the prior line. For instance it may be that at the time of disease progression in the prior study, their performance status had deteriorated such that they would no longer have been eligible for the novel intervention (for example if their performance status had worsened). Even if patient level data are available (I had to estimate this from published information), such detail is unlikely to have been captured in the trial.

5.2.2 USING A PATIENT'S OUTCOMES FROM A PRIOR LINE OF TREATMENT TO ESTIMATE COUNTERFACTUAL OUTCOMES

The methods approach described in this section (5.2.2) was published in Statistical Methods in Medical Research (Hatswell & Sullivan, 2019) *and used by Gilead in NICE Appraisal ID1379, though I was not involved in this*

In some conditions, where the overall objective is of disease control (and whilst their disease is controlled, their risk of events is low), a further option for estimating the benefit of treatment may be found using a patient's own treatment history.

An example of this can be seen in Follicular Lymphoma (FL) – a condition of the lymphatic system where white blood cells multiply (creating abnormal B-cells), which collect in lymph nodes. The disease is treatable, though patients will eventually become resistant to the drugs used. At the point patients become resistant to treatment, their white blood cell count will begin to rise (again), with common symptoms of tiredness, weight loss, and fever.

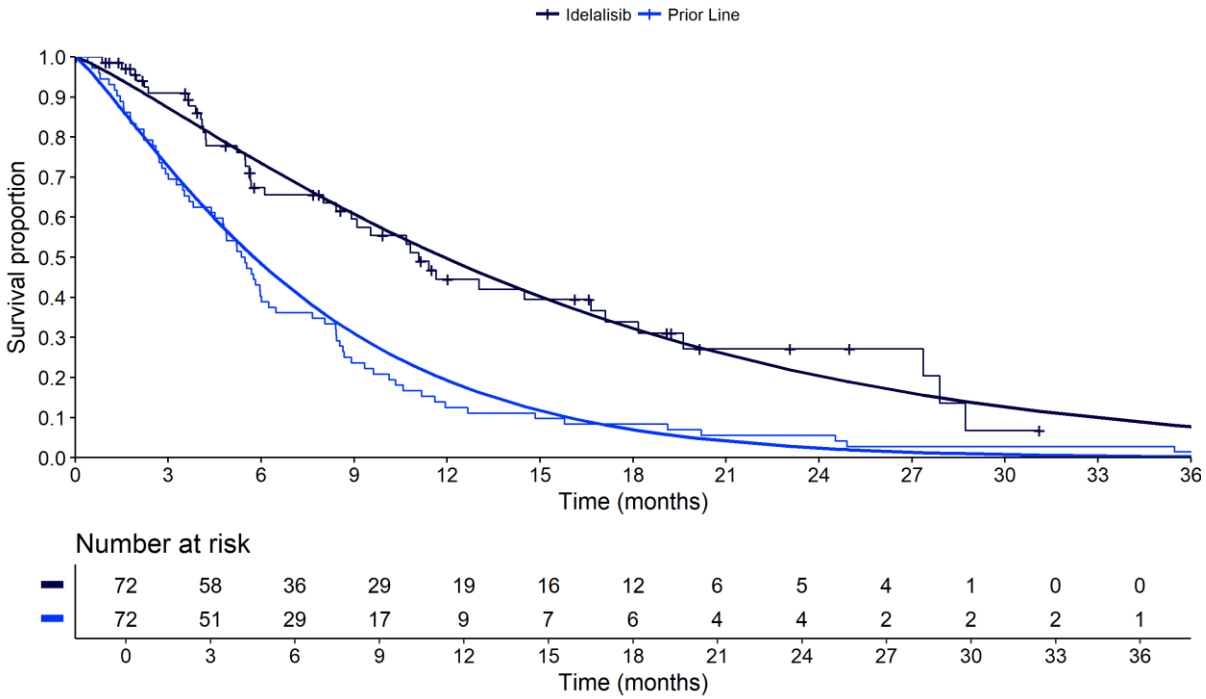
It is in the context of highly pre-treated patients that the drug idelalisib was licensed on the basis of a Phase 2 study (Gopal *et al.*, 2014) which included 72 patients with FL (from a total of 125). All had been heavily pre-treated (a median of four prior regimens) and were now at the point where no single standard of care was agreed upon with patients generally receiving a variety of off-label chemotherapies. As the results of the Phase 2 study were promising (an overall response rate of over 50%), no further studies were conducted and the benefit-risk of the product deemed sufficient in the eyes of regulators in both Europe and the FDA.

Whilst regulatory approval is a necessary condition for use in HTA driven countries, it is not sufficient, and the need to estimate the comparative effectiveness prompted a need to look for other methods, in the absence of a suitable historical control. Given the patients had all received previous treatments, and would in the absence of idelalisib, be given a different regiment to what they had previously (as there is no standard of care) – these treatments act as the counterfactual. The approach suggested is therefore to compare the outcomes at the previous line of treatment, as these may be broadly reflective of the outcomes achieved in this patient group.

With access to patient level data from the study, to perform the analysis would be relatively straightforward. Without such access, in order to estimate the comparative effectiveness of idelalisib a different approach was needed. Whilst the main publication by Gopal et al. only contains outcomes for all patients in the study (not FL only), a kin paper by Salles et al. (2017) includes data on OS, and time to progression (TTP) for patients. Whilst medians for prior treatment (to allow validation) are not reported in the published study, this was found in

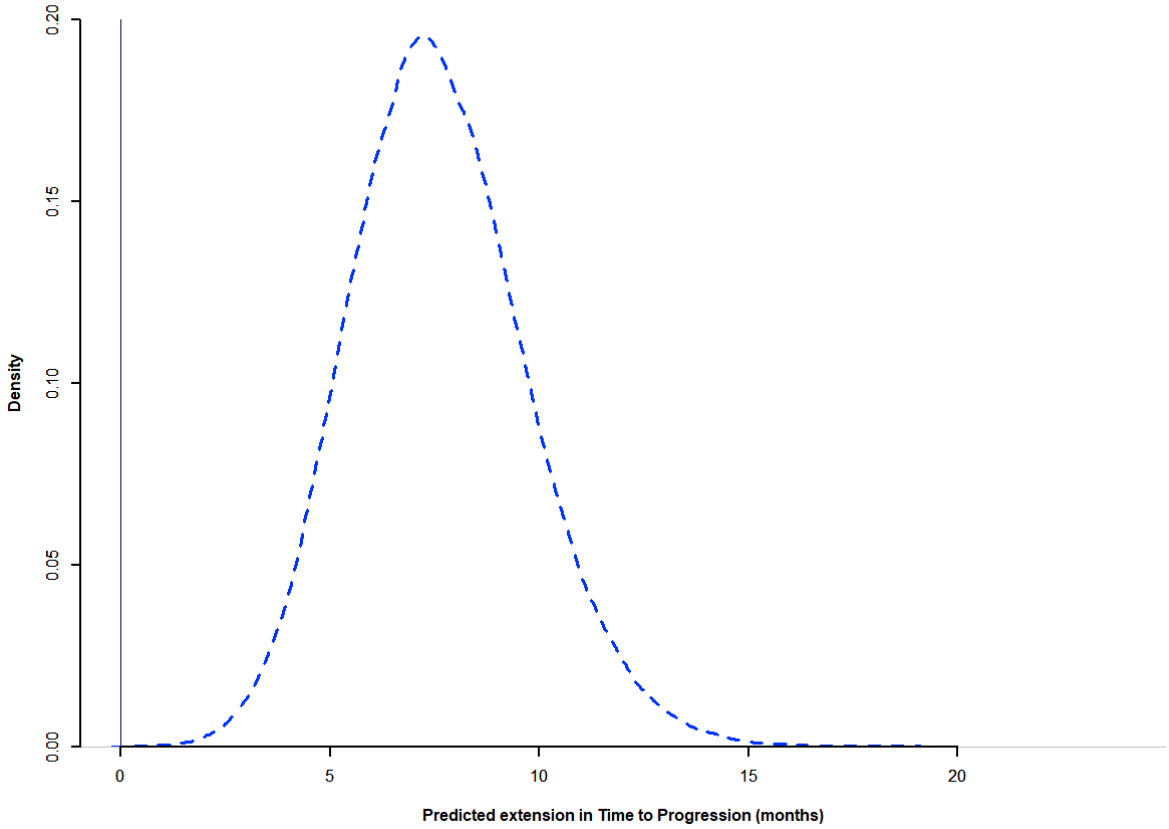
publicly available information from PBAC, and the FDA (Miller, Przepiorka & de Claro, 2014; PBAC, 2015). The available data were digitised and used to recreate pseudo patient level data for TTP for idelalisib, and the previous line of treatment to which parametric curves could be fitted (using the Latimer algorithm, independent Weibull curves provided the best fit to data). The fitted data was a close approximation to the live data, with medians of 11.1 and 5.4 months TTP, compared to the reported 11.0 and 5.1 months.

Figure 5-6: Digitized Kaplan-Meier data from idelalisib study 101-09, including fitted (Weibull) parametric survival curves



The area between the curves represents the predicted extension in time to progression of idelalisib. In order to estimate the expected gain in TTP (including uncertainty) the fitted survival curves were bootstrapped 100,000 times, with the area under each curve summed, and the difference taken. The results of this analysis are shown in Figure 5-7, and estimate a mean gain of 7.6 months for idelalisib – given the prior line of therapy achieved a mean of only 7.4 months, this appears an impressive result, and explains how in all bootstraps idelalisib was superior.

Figure 5-7: Density plot of estimated Time To Progression gain of idelalisib over the previous line of treatment



It should be noted that there is a number of assumptions / limitations around the approach. The first (and main) of these being that the data available for the prior line of therapy is, by definition, TTP and not the more widely used endpoint of PFS (in TTP deaths are counted as censors, whereas in PFS they are classified as events). This means that no data are available on how often patients died whilst on the prior line of therapy, or waiting to begin the next line – an assumption must therefore be made around the rate of deaths whilst on treatment (this could be informed by the difference between TTP and PFS in the contemporary study).

The second major limitation is how survival gain must be estimated – the proposed approach gives an estimated gain in time on treatment, it does not inform what may happen to survival after treatment. Should there be differences in post treatment survival between lines, this is not something that is able to be estimated within this approach.

5.3 SUMMARY OF NOVEL METHODS PROPOSED

The work in this chapter presents three novel methods for the creation of historical controls. Although not suitable on every occasion their use may facilitate the use of existing data to allow comparisons – indeed two of the methods have already been used in practice.

6 PRACTICAL EXAMPLES OF THE IMPLEMENTATION OF METHODS

In part due to the work discussed in previous chapters, I have been fortunate enough to have the opportunity to use some of the methods presented for HTA submissions, two of these pieces of work are presented as case studies below. They demonstrate how the different approaches can be implemented using publicly available information.

6.1 AVELUMAB FOR THE TREATMENT OF MERKEL CELL CARCINOMA

The approach described in this section were used to estimate comparative effectiveness in the NICE appraisal of avelumab in Merkel Cell Carcinoma (NICE TA517) and subsequently published in Bullement et al. (2019)

Markel Cell Carcinoma is a rare aggressive skin cancer which is more common in older people and those with immunosuppression. Historically off label chemotherapy has been used in first line disease, with patients who progress potentially receiving chemotherapy, or going untreated. In this disease area the PD-1 immune checkpoint inhibitor avelumab was studied in 88 2nd line patients with no control arm (Kaufman *et al.*, 2016).

As this was an uncontrolled study for HTA the comparative effectiveness needed to be estimated. To do so the license holder of avelumab (Merck) performed a retrospective database study of patients who had previously been diagnosed with Merkel Cell Carcinoma, and gone on to receive second line therapy, identifying 20 patients in the EU (Cowey, Becker & Bharmal, 2016) and 34 patients in the US (Becker *et al.*, 2017). A literature search was also performed (Nghiem *et al.*, 2017), identifying a study by Iyer et al. (2016) which reports outcomes for 30 patients – a further study included 14/23 second line patients, but outcomes were not reported by line (Samlowski *et al.*, 2010).

To establish which patient characteristics are prognostic, and thus would be important to use for propensity scoring (in case of the database study where individual data was available) or MAIC (in the case of Iyer et al., where individual level data was not available), the data from the database study was interrogated. This was done using a variety of techniques; univariate regression, multivariate regression, and Kaplan-Meier plots (which may help identify trends, even if these do not reach statistical significance). This exercise was conducted for both PFS and OS. Candidate variables were selected with input from clinicians with experience in the disease area.

The first stage in the analysis was to fit a variety of parametric survival curves to the data, selecting the most appropriate functional form for any regression analysis. In both visual inspection and AIC, the Weibull curve was preferred for both PFS and OS, and was

therefore used for multivariable and univariable regression. In the multivariate regression, none of the suggested patient characteristics (stage at diagnosis, gender, age, immunosuppression status, ECOG) were predictive of outcome. Where results began to approach statistical significance at the threshold of p=0.1, this was in the opposite direction to what would be expected; for instance older patients (over the age of 75) having better survival than younger patients.

Based on the analysis performed, no matching or weighting methods were therefore used in the Merck submission to NICE – in performing matching between the trial data and historical controls, we would have been unlikely to improve on a naïve comparison, and may indeed have introduced further bias and uncertainty. Whilst this was questioned in the NICE process, the end result was the acceptance of the case made by Merck, and the approval of the product for use.

6.2 BRENTUXIMAB VEDOTIN FOR THE TREATMENT OF HODGKIN'S LYMPHOMA

The approach described in this section was used to estimate comparative effectiveness in the NICE appraisals of brentuximab vedotin for Hodgkin's Lymphoma (NICE TA446 & TA524) and are currently under review for publication

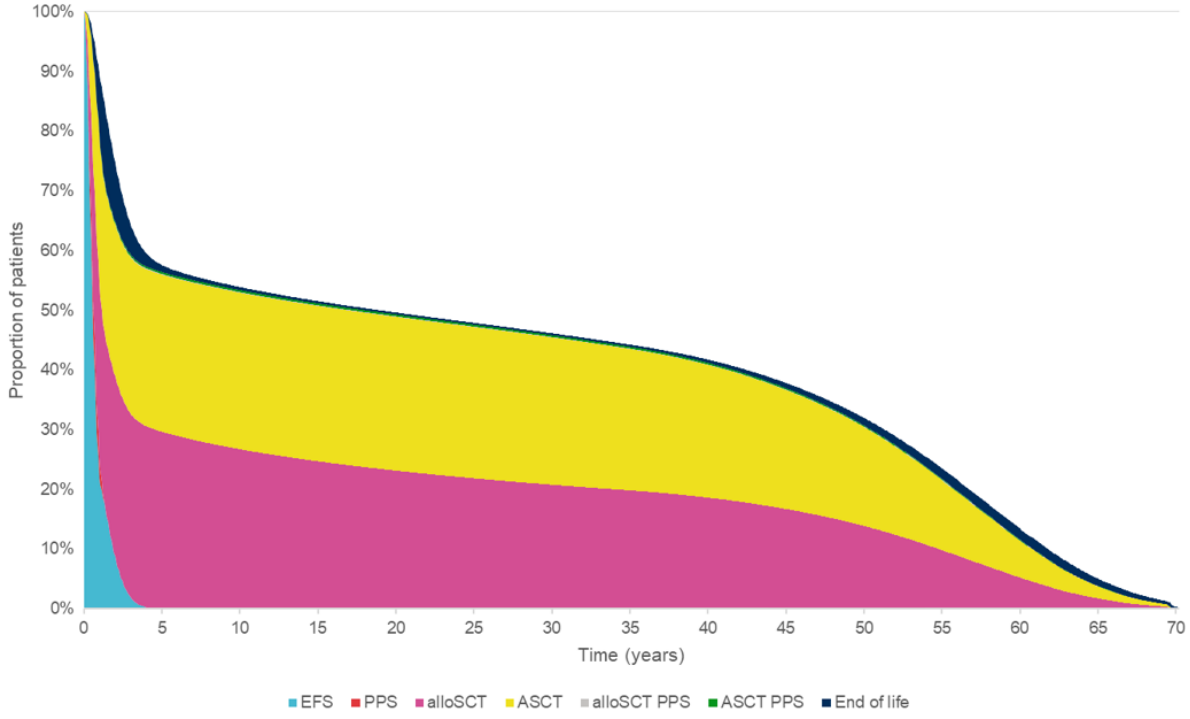
Brentuximab vedotin was studied in several positions in Hodgkin's lymphoma, including in randomised clinical studies (Moskowitz *et al.*, 2015), one indication it received however, was for use in patients who were refractory to two prior lines of chemotherapy, and unsuitable for multi-agent chemotherapy. The data collected in this group was observational in nature, and involved small numbers of patients (40) – standard of care would likely have been single agent chemotherapy, which was known to have poor outcomes in this group (who had failed all prior treatments).

At the time brentuximab vedotin was licensed (2010) the UK had a mechanism known as the 'Cancer Drugs Fund' (CDF) in which treatments needed to show only clinical efficacy, and not cost-effectiveness, to gain routine use. When this system was revised, all medicines currently approved were then re-appraised by NICE for their cost-effectiveness including any new evidence generated under the older system (Grieve *et al.*, 2016). At this point Takeda (the manufacturer of brentuximab vedotin) were required to demonstrate the comparative effectiveness of brentuximab vedotin.

Fortunately for Takeda, whilst the drug had been available on the NHS, data had been collected by a collaboration of clinicians in the UK, which was subsequently published (Eyre *et al.*, 2017). This included data on 99 patients treated with brentuximab vedotin, and

allowed the modelling of outcomes using parametric curve fitting to account for the movement between health states (Event Free Survival, Post Progression Survival, and SCT). External (published) data was then used to model the outcomes of SCT (Sureda *et al.*, 2012, 2001). Applying these curves in a Microsoft Excel model led to the Markov Trace shown in Figure 6-1.

Figure 6-1: Markov trace of modelled patient survival with brentuximab vedotin in Hodgkin's lymphoma

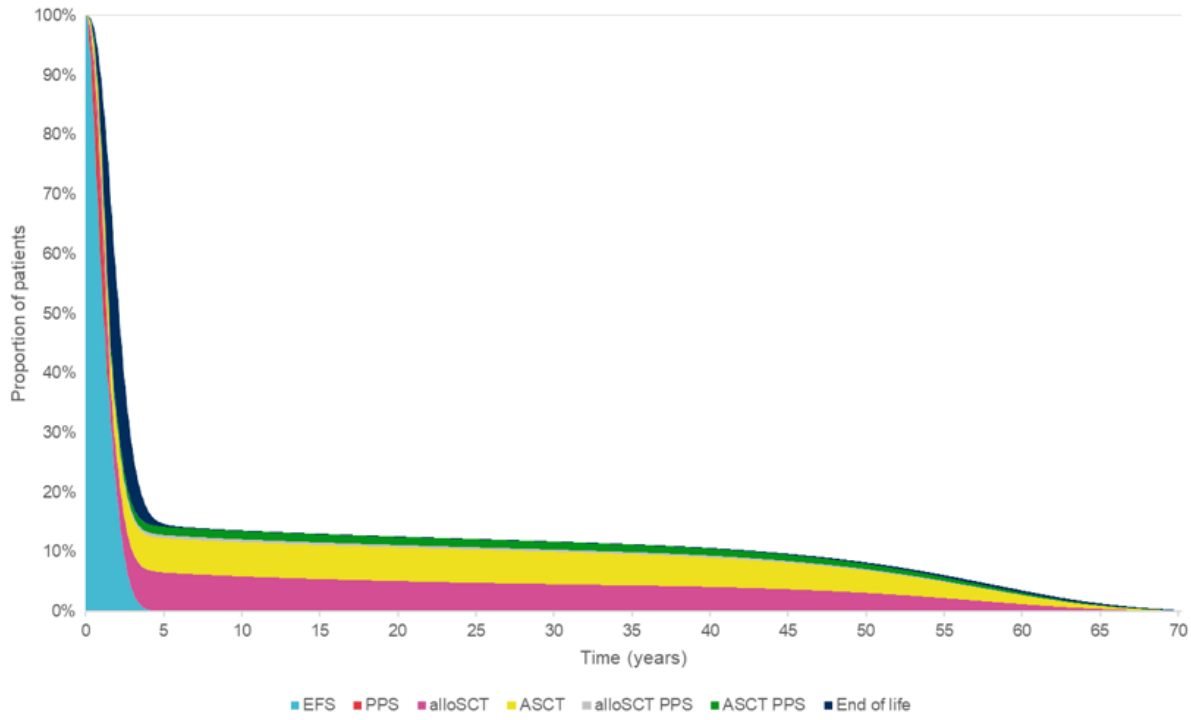


To model the outcomes of standard of care, a literature search performed for the benefit of regulators identified four publications giving historical outcomes for single agent chemotherapy (Haim, Ben-Shahar & Epelbaum, 1995; Little *et al.*, 1998; Mead *et al.*, 1982; Zinzani *et al.*, 2000). One of these studies (Mead et al.) however reported data only on patients who had survived for at least six months, thus introducing 'immortal time bias' (Lévesque *et al.*, 2010) meaning it had to be excluded. Due to the age of the publications, little information was reported on patient baseline characteristics, with only one study (Little et al.) reporting time to event endpoints – and even then, in only 17 patients, thus preventing direct modelling of outcomes.

To overcome this limitation, a surrogate outcomes approach was used. The ORR and SCT rates (which were reported in each of the publications) were meta-analysed using a random effects model in the R package *meta*. This gave an estimated ORR rate of 64.3%, and estimated SCT rate of 7.8% (higher rates than would have been seen with a naïve pooling). To model the effects of these outcomes, patients were assumed to follow the same

outcomes as with brentuximab vedotin for a given level of response e.g. the same EFS to PPS distribution, weighted for the number of responders and SCTs. Following this approach led to the Markov Trace shown in Figure 6-2.

Figure 6-2: Markov trace of modelled patient survival with single agent chemotherapy in Hodgkin's lymphoma



When the outcomes are compared on a per health state basis, it can be seen that the large survival gain is primarily given by the increased rate of SCT for brentuximab vedotin (41.0% versus 7.8%) – as SCT is curative in around a half of patients this leads to a substantial improvement in outcomes.

Table 6-1: Estimated life years in each health state for brentuximab vedotin and single agent chemotherapy in Hodgkin's lymphoma

Health state	Brentuximab vedotin	Standard of Care	Incremental
EFS	0.8	1.3	-0.5
SCT	25.2	5.7	19.5
PPS	1.0	0.3	0.7
Palliative care	0.8	0.6	0.2
Total	**27.7**	**7.9**	**19.9**

The weaknesses of the study were noted in the NICE appraisal - most notably the potential for different patient characteristics, and the age of the historical data. This meant there was the potential the data were no longer relevant due to the 'drift' in patients and outcomes. Despite these limitations, in the first appraisal (TA446) brentuximab vedotin was given 'access with evidence development' in the revised CDF, before being fully approved in TA524 after further evidence on the rate SCT seen in practice was collected.

7 DISCUSSION

7.1 THE USE OF UNCONTROLLED CLINICAL STUDIES IN PHARMACEUTICAL LICENSING

The literature review performed as a part of this project was completed in 2014, and covered approximately 15 years of pharmaceutical licensing in Europe and the US. Throughout the period approvals were made on the basis of uncontrolled studies, with no clear time pattern; any cluster of approvals was more linked to a drug being approved with multiple indications, as opposed to a number of drugs arriving simultaneously. Whilst it is possible this pattern may have changed (or indeed may change in the future) with novel mechanisms promising large 'obvious' improvements (such as chimeric antigen receptor and gene therapies), the framework of regulators to assess benefit-risk seems adequately equipped to handle these. Indeed when the review I conducted was updated by others (Goring *et al.*, 2019), no differences were seen in approval types or patterns.

One interesting aspect to the review completed was that it does seem the FDA are more accepting of uncontrolled studies; of the 44 comparable applications made, 43 were approved, whilst the EMA approved only 35. Similarly the review time was much shorter from the FDA; whilst this may be to a degree process driven, the magnitude of difference in time from application to approval (8.7 vs 15.5 months, a 6.8-month difference) indicates that the EMA asked more questions of manufacturers, and were ultimately more reluctant to give approval. Fundamentally however, the majority of applications to regulators for marketing authorisation were successful, showing that in many instances the benefit-risk was deemed to be favourable. Given the indications investigated (mainly end stage cancers with a poor prognosis) this is unsurprising, but does demonstrate the need for methods to estimate the comparative effectiveness of products for payers.

Although not containing novel arguments (other papers are referred to), one of the best summaries of the position of uncontrolled studies in drug approval was given by Byar et al. (1990), when discussing the appropriateness of placebo controls for AIDS. The paper came from a discussion in 1989 at the height of the epidemic, with the recent licensing of AZT (zidovudine), the first antiretroviral. The authors state that randomisation is 'essential' in AIDS studies, and high levels of evidence will do the most good:

> Previous research in chronic diseases has taught us that many proposed treatments were of no benefit, some did more harm than good, and those that were beneficial generally yielded only moderate gains. Unrealistic expectations of benefits may lead to unrealistic research strategies that might be suitable to detect large benefits but not moderate ones. Although in special situations uncontrolled or historically controlled trials might be considered, we believe that

> progress in the treatment of AIDS will be most rapid and certain if researchers devote their energies chiefly to the design of randomized trials.

Yet despite this position some of the clearest (and most frequently cited) examples of uncontrolled studies come in AIDS research; ganciclovir is cited as an example by the authors (separately it is used by Rawlins et al. as an example of where RCTs are not needed due to the 'obvious' benefit – Section1.5.1). The drug however was licensed for use in AIDS linked cytomegalovirus reactivation, on the basis of data from only 41 patients over a five year period (Food and Drug Administration, 1989). Similarly one of the best examples to explain the necessity of uncontrolled study approvals in the systematic review detailed in Chapter 3 was paclitaxel for Kaposi's sarcoma (a symptom of AIDS). For this treatment a RCT could not be conducted for several reasons including equipoise - the prognosis was exceptionally poor for patients, the results from early studies were impressive, and the drug's safety established in other indications. More relevant however was the development of antiretrovirals (such as zidovudine) which had led to the incidence of the disease plummeting. A RCT would not recruit sufficient patients as within two years it was anticipated that given the fall in incidence, there would be few Kaposi's sarcoma patients to benefit from the findings (which would likely be confirmation of the efficacy of paclitaxel).

These examples show the paradox inherent in such approvals; whilst they are not the preferred approach (and are discouraged by the majority of authors), there remain situations where the evidence on effectiveness is sufficiently compelling that despite the lack of control arm, a rational regulator will be convinced of the benefit-risk of a product. At this stage payers (amongst others) will need to quantify the magnitude of benefit in order to make coverage decisions – regardless of whether such information is optimal.

7.2 TECHNIQUES FOR ESTIMATING EFFECTIVENESS BASED ON UNCONTROLLED CLINICAL STUDIES

The results of my review of models used to estimate comparative effectiveness found that by far the most common approach used was that of a historical control, seen in around 80% of approaches. At the time however few attempted to control for any differences between datasets, or even investigated any potential imbalances between studies i.e. they assumed populations were exchangeable. Given previous work in the area, this appears inappropriate.

Although only anecdotal, it does seem like this situation is improving; the technique of MAIC was first published in 2010 (towards the end of the period covered by my literature search for models), but has been then used in over 50 publications since (shown in Chapter 4). Though seemingly not as widely used, the same applies to the technique of STC.

Fundamentally however the understanding of the limitations of historical controls is now more widespread (in part due to the proliferation of methods and guidance about how to account for these differences). These initiatives, such as the NICE TSDs, do then suggest the assessment of similarity of datasets, ultimately meaning some may be deemed not suitable for use. The work performed in this thesis around MAIC may also help in this regard; in showing some differences which may be problematic – such as non-overlapping populations.

In my opinion, other methods I highlight should also be more widely used, such as the E-value, and threshold analysis. These may help to provide interpretation of uncontrolled studies, and highlight the strength (or fragility) of any conclusions drawn, in a way similar to which the p-value, for all its flaws, allows interpretation of the strength of findings from RCTs.

7.3 IMPLICATIONS FOR THE DESIGN AND CONDUCT OF UNCONTROLLED STUDIES

The work I have conducted has been around the application of methods for the estimation of comparative effectiveness from uncontrolled studies – separate to the question of whether randomised studies should be performed, which is addressed in detail elsewhere in the literature. There are however implications of the work I have performed for data collection by pharmaceutical companies which happens both in, and alongside uncontrolled studies.

Statistical modelling of efficacy forms the basis for health technology appraisal (HTA) which determines reimbursement in many countries. Even where HTA is not required for market access, plausible estimates of the incremental benefit of treatment can aid in gaining regulatory approval and promoting uptake amongst clinicians. Indeed the increased sophistication of regulatory agencies may also result in statistical analyses being requested to provide context to an application dossier – to this end companies should be mindful of the data required to construct such estimates.

A considered approach would look at the availability of historical data prior to the final study design if there is a possibility that a trial may lead to a regulatory application, a similar approach to ‘Threshold-crossing’ proposed by some regulators (Eichler *et al.*, 2016). Should a suitable historical control be available, then the study design data collection should be carefully considered to facilitate comparisons between the data sources. For example although a metric (such as an older version of a disease staging tool) may not be relevant for understanding the behaviour of the new intervention, if it is available in the historical control then collecting this data may reassure reviewers as to the similarity of patients (or provide data for matching methods).

Conversely, if no historical control is available that would match with a study design, adequate planning would allow this to be addressed in some form. Approaches that could be used include commissioning observational studies to collect such data – either prospective in locations where the trial is not recruiting, or retrospective using databases and hospital records. An example here are the database studies conducted by Merck in Merkel Cell Carcinoma in parallel to the uncontrolled clinical study of avelumab, where inclusion criteria for both studies were very similar with the aim of producing a good basis for comparison (Becker *et al.*, 2016; Cowey, Becker & Bharmal, 2016). Alternative approaches include those discussed in Chapter 5 to create historical controls - if such methods are appropriate, further data collection within the uncontrolled study may also be merited. For example if the intention is to create a control based on patients' previous line of therapy, this requires information to be collected with as much fidelity possible from patient histories at the time of the study.

The conclusion of my research therefore is that although there are statistical techniques which may help to provide robust estimates of comparative effectiveness, their use needs to be adequately planned for. Should companies engage ahead of time, it may be possible to gain access to ILD for historical trials, or as a minimum, ensure that there are historical studies which could be compared against, should their study merit such a comparison. A larger focus on planning for such eventualities may avoid the need for some of the more speculative comparisons that are presently required.

7.4 POTENTIAL IMPROVEMENTS AND MODIFICATIONS TO TRIAL DESIGNS

In addition to data collection efforts, improvements could also be made to the way trials are designed – particularly if there is a chance they will be used for regulatory applications. Ideally a systematic approach to the setup of uncontrolled studies (as with the Threshold Crossing approach), but at the very least having a 'run in' phase to clinical studies, where data are collected from patients who would be eligible, but before the drug is made available. This could provide some data on the natural history of patients and could be performed whilst manufacturing or ethics approval for administration of the investigational drug is being finalised (though ethics approval would be required for the observational period). A similar approach was taken in haemophilia for the drug emicizumab (Pipe *et al.*, 2019) where patient histories (or varying lengths) were available before patients crossed over to receive treatment. Were the length of this 'run in' period to be randomised on a per patient level, depending on the disease area, it could also be amenable to analysis using an interrupted time series (Section 2.1.1.3) to provide more formal estimates of the treatment effect.

Other approaches could be further improved to leverage data; Jiao et al. (2019) based at the FDA, suggest the use of 'platform' trials, in which control arms are shared between different products. These are similar to the multi-arm studies proposed by others to allow the assessment of multiple products simultaneously (Parmar *et al.*, 2017; Zeissler *et al.*, 2020; Grayling & Wason, 2020). Whilst this would be advantageous (in reducing ethical concerns but maintaining power), it has practical limitations in that companies are unlikely to wish their products to be investigated in the same studies as even if underpowered an implicit head to head comparison will be made between treatments (with a non-trivial chance of giving an incorrect point estimate of relative effectiveness). A development to the approach would be removing the explicit link within a single study between treatments. This could be done by having continuous recruitment, meaning there would be a high correlation between the effective control arms for each product, but without them being exactly the same, and thus avoiding direct comparisons being able to be made. This approach would be a variation on 'master protocol' studies though would only be possible in well-defined disease areas where inclusion and exclusion criteria could be shared, and sufficient patients enrolled. Although complex the approach would be similar to the 'cohort multiple randomised trial' design of Relton et al. (2010) whereby a cohort study is conducted at the same time as patients are enrolled from the same population for uncontrolled studies.

7.5 SUGGESTED DECISION PROCESS FOR METHOD SELECTION

In reviewing the methods available for the estimation of comparative effectiveness, depending on the circumstances (and data) available, different methods would seem appropriate depending on circumstances. As seen in the review of existing approaches (both in those used, but also those available), this mainly revolves around the availability of historical control(s), and the level of access to individual patient data available for the datasets of interest.

Where individual level data are available for all studies of interest, the methods of NICE DSU TSD 17 would seem applicable, and the flowchart in that document should be followed. Similarly where no patient level data are available the route forward seems relatively straightforward; meta-regression can be used if sufficient studies are available (Following the guidance in NICE DSU TSD 3), but more likely a narrative conclusion should be given.

If limited individual patient data are available (i.e. data from one, but not both studies), options to minimise bias should be considered. These will vary depending on the circumstances, but could include MAIC, STC, and surrogate outcome based approaches.

Where no historical information is available consideration should be given to identifying one, either through the use of literature searching, emulation of a trial from other data, or statistical approaches such as those presented in Chapter 4. After any historical controls have been analysed with appropriate methods, presenting results of both naïve and adjusted comparisons seems appropriate. No method can be conclusively shown to remove all potential issues with data, and thus given the limitations in data and methods, the impact made by any statistical approaches (for instance reweighting) is likely to be important for decision makers to be aware.

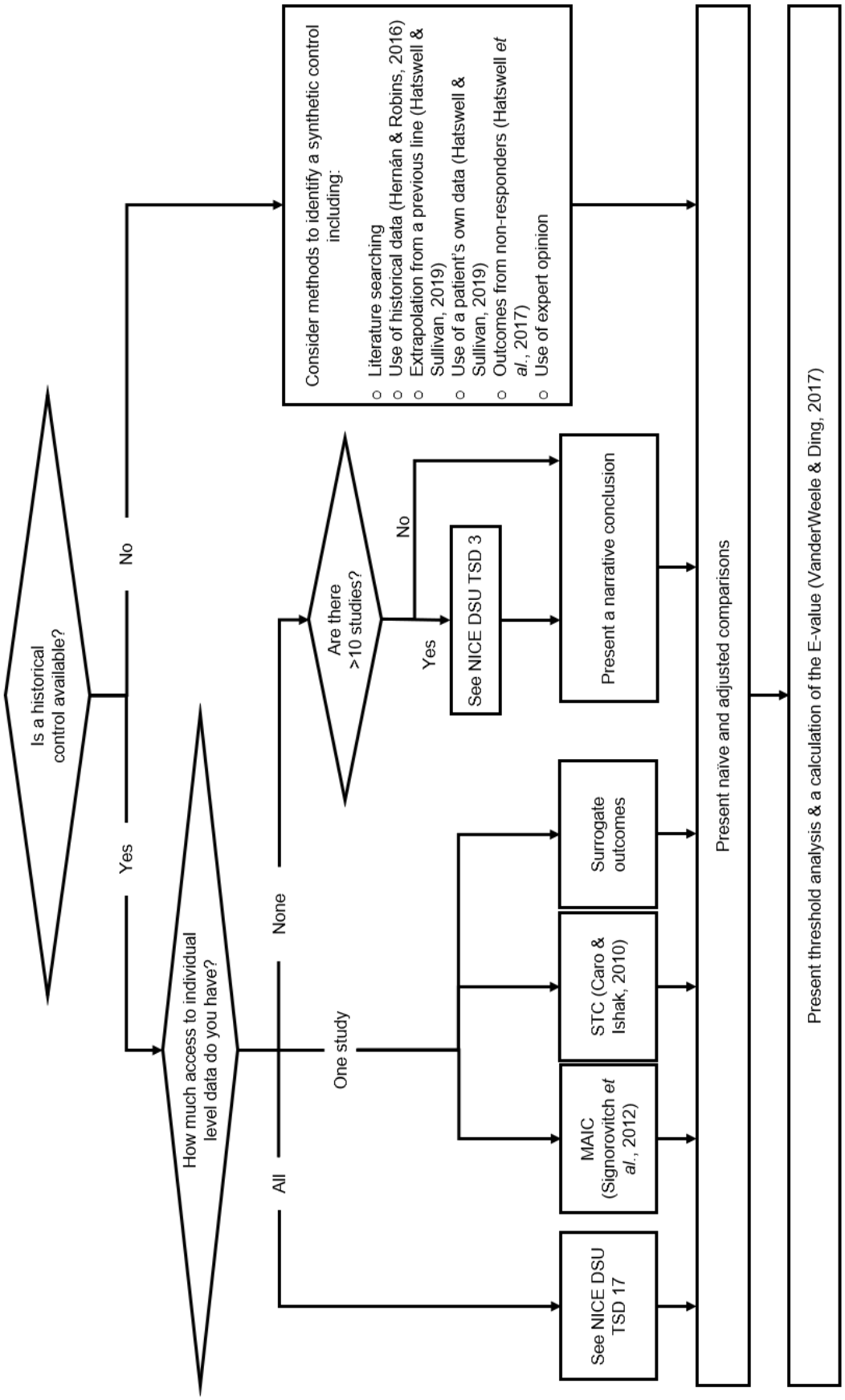
Finally, regardless of whether a historical control is available, consideration should be given to threshold analysis, and the use of the E-value (VanderWeele & Ding, 2017). Threshold analysis allows decision makers to understand how much a parameter (such as hazard ratio) may need to change by for the decision to be incorrect, whilst the E-value performs a similar function in helping decision makers to understand the robustness of the evidence they are being asked to appraise. To understand how much 'clearance' there is in any given decision may allow a more open discussion of the uncertainties around comparisons, and avoid a focus on point estimates.

This suggested approach is laid out as a flowchart in Figure 7-1, and incorporates the findings of the thesis, in considering what circumstances each of the methods will be required (Chapter 3), categorising the approaches based on the type and level of data available (Chapter 2), the suitability of MAIC (Chapter 4), incorporating the novel methods proposed (Chapter 5) as well as other methods identified in the process of completing the work (Section 2.3 and Section 2.4).

Although it is not possible (due to the nature of the evidence) to be conclusive, I believe the use of this flowchart as a decision aid would represent best practice, and ensure opportunities for improvements in the evidence presented to decision makers is optimal. Optimal in this context being allowing the most accurate estimate of comparative effectiveness, and also an understanding of the strength of findings.

In due course it may also be possible to create a flowchart for how to synthesize multiple estimate of effectiveness. Some form of synthesis would seem appropriate - however if not all studies have the same face validity or applicability to the decision problem, a level of down weighting to those studies i.e. the power prior approach, would seem appropriate. At this point in time however it is not clear how the level of down weighting should be determined.

Figure 7-1: Suggested algorithm for estimation of effectiveness based on uncontrolled clinical studies



7.6 MY CONTRIBUTION TO THE LITERATURE

There are several areas where I feel my work has advanced the literature. The first of these in defining the scale of the issue of drug approvals without RCT data. Anecdotally I had been told, including by my supervisors, that such approvals were rare but increasing in number. The work in reviewing drug approvals (which has been widely cited) demonstrates such approvals are not a recent phenomenon, not necessarily rare, and with a rate of use that does not have a clear trend (Section 3.1). That others have updated this literature review I believe emphases its relevance.

Subsequently, being able to investigate approaches taken to modelling these approvals (Section 3.2) has emphasised the role of historical controls, but also the potential biases involved in using a naïve comparison to historical data. In this area I have then outlined the relevant approaches depending on the availability of historical data, signposting the relevant work where needed (Chapter 2). In identifying a gap relating to the suitability of one of the more widely used techniques to adjust for imbalances seen with historical controls (MAIC) I then performed a simulation study to understand what assumptions must be met for the method to be both accurate an unbiased (Chapter 4).

In addition to examining existing techniques, I also propose three new methods (with associated examples) for the estimation of comparative effectiveness where historical control data are not available (Section 5.1 and Section 5.2). I then give two (real world) examples of using the methods and approaches I have identified (Chapter 6).

In summarising my findings, I then propose a flowchart for what I consider to the be options available to an analyst, which is presented in Section 7.5.

7.7 LIMITATIONS OF THE WORK PERFORMED

Whilst I have attempted to be comprehensive in the work I have performed, there are a number of limitations which relate to the scale of the issue, and in defining the scope of the thesis.

One limitation is in my review of approvals granted by the EMA – whilst the vast majority of drugs now go down this route, individual countries in the EU may have approved treatments on the basis of uncontrolled studies (which are then able to be used in other countries via the process of mutual recognition). Whilst this is mitigated by the (intentional) inclusion of FDA approvals in my review, it does mean there may have been some treatments,

particularly earlier in the review period when the EMA processes were still being established, which were not included.

A second limitation is that reality decisions on benefit-risk, and indeed clinical development are seldom as neat as being based entirely on uncontrolled data versus using RCTs. In the majority of cases even a large Phase 3 RCT would usually be supported by some level of uncontrolled data – for instance an open label extension trial which collects further data on sustained safety and effectiveness. Whilst in some ways a limitation (I do not look at the role of uncontrolled data in all areas), it does mean that the work I have done is not limited in utility simply to uncontrolled studies - it does have broader implications. As an example the work on the appropriateness of MAIC would be helpful in interventions that have RCTs , but no complete network to perform network meta-analysis (for instance due to different comparator arms).

Another limitation is that decisions on the use of uncontrolled trials for drug approval are made for a variety of reasons which may include clinical need, financial, competitive, and strategic reasons. As drugs themselves are not randomised to being investigated in uncontrolled versus controlled trials, this limits what can be inferred about them, for instance their precise effectiveness. If such drugs are seen as being highly efficacious early in development then the existing literature, for example on bias in historical controls (developed where historical, and randomised controlled estimates are available), may not apply.

A final limitation is that there are, necessarily, many areas which are out of scope, or would be developments of the work I have performed which I feel deserve further research. These areas I have I have discussed below.

7.8 FURTHER RESEARCH

Based on the work performed, there is a number of areas where I believe further research would be useful. These can be separated in to areas relating to methodology, areas relating to clinical data, and finally research on decision making.

In terms of statistical methods, both MAIC and STC would seem in need of further work to establish when they should be used, and how they should be used:

- For MAIC I believe it is important to understand the interplay between sample size, and number of characteristics that can / should be included in any matching. For instance it would be good to know (given the limited size of many studies), when it would be important to omit a variable from matching if it is linked to outcomes, but only weakly

- STC is less investigated, but in being a different (regression based) approach I believe research would be helpful on how STC should be conducted with multiple outcomes (such as PFS and OS). Each outcome could be modelling with different explanatory variables (as determined by clinical input, and statistical relevance), or the same covariates used, regardless of say statistical significance. Which of the two approaches should be used appears relevant

- Both MAIC and STC would benefit from the availability of standardised approaches – for instance open source code, or software implementations (such as an R package)

- Both MAIC and STC result in estimates of treatment effectiveness for the intervention with individual level data, but in the population for which only aggregate information is available. When multiple sets of aggregate level data are available, it is not clear how these varied estimates should be synthesized (if indeed, they should be synthesized at all)

Further methodological research around the use of historical data would also be helpful, in particular how multiple sets of historical data should be synthesized. This thesis highlights meta-analytic predictive and power priors, but how these should be used without contemporary controls to determine acceptability is uncertain. Specifically how weights should be assigned to each set of historical data represents an area of uncertainty which does not appear to have been resolved in the literature.

In terms of clinical data, further research in to the differences between case series, registries, and clinical trials would be helpful. Whilst I have identified a number of papers that demonstrate differences between outcomes, these do not attempt to provide a systematic approach, nor help to pinpoint the source of this difference. For instance are different patients enrolled in trials vs registries? Are outcome measurements taken on different schedules, or does investigator versus independent assessment affect the metrics recorded?

Similarly with the increasing availability of electronic health records (particularly in the US), it may become increasingly possible to generate synthetic control arms at low cost. Whilst this has been done to a limited extent by companies such as Flatiron Health (Carrigan *et al.*, 2019; Davies *et al.*, 2018) the approach may help reduce the uncertainty inherent in decisions made by payers and clinicians, but first must be appropriately validated before being relied upon for decision making.

The final area where I believe further research is needed is in helping to decide which approaches should be recommended to estimate comparative effectiveness. I have highlighted the methods available throughout this thesis, and provided a diagram of which may be suitable (depending on the various levels of access to data) in Section 7.5, but further research is needed to understand which methods are most appropriate for a given decision problem. For instance what level of model fit is needed for STC to be preferred to MAIC? The availability of such recommendations would help to reduce the variability seen in assessments, and ensure decisions on access to treatments are made on the best available evidence and analysis.

REFERENCES

Aiello, E.C., Muszbek, N., Richardet, E., Lingua, A., et al. (2007) *Cost-effectiveness of new targeted therapy sunitinib malate as second line treatment in metastatic renal cell carcinoma in Argentina*.

All Wales Medicines Strategy Group (2013) *Form B Guidance Notes*. [Online]. Available from: http://www.awmsg.org/docs/awmsg/appraisaldocs/inforandforms/Form%20B%20guidance%20notes.pdf .

Anderson, D.R. (1991) Umbrellas and lions. *Journal of Clinical Epidemiology*. [Online] 44 (3), 335–337. Available from: doi:10.1016/0895-4356(91)90045-b.

Anderson, M., Naci, H., Morrison, D., Osipenko, L., et al. (2019) A review of NICE appraisals of pharmaceuticals 2000–2016 found variation in establishing comparative clinical effectiveness. *Journal of Clinical Epidemiology*. [Online] 105, 50–59. Available from: doi:10.1016/j.jclinepi.2018.09.003.

Annemans, L., Van Cutsem, E., Humblet, Y., Van Laethem, J.L., et al. (2007) Cost-effectiveness of cetuximab in combination with irinotecan compared with current care in metastatic colorectal cancer after failure on irinotecan--a Belgian analysis. *Acta Clinica Belgica*. [Online] 62 (6), 419–425. Available from: doi:10.1179/acb.2007.061.

Antoñanzas, F., Terkola, R. & Postma, M. (2016) The Value of Medicines: A Crucial but Vague Concept. *PharmacoEconomics*. [Online] 34 (12), 1227–1239. Available from: doi:10.1007/s40273-016-0434-8.

Austin, P.C. & Steyerberg, E.W. (2015) The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*. [Online] 68 (6), 627–636. Available from: doi:10.1016/j.jclinepi.2014.12.014.

Australian Government Department of Health (2013) *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee Version 4.4*. [Online]. Available from: http://www.pbac.pbs.gov.au/content/information/printable-files/pbacg-book.pdf [Accessed: 5 March 2015].

AWMSG (2012) *AWMSG Secretariat Assessment Report – Advice No. 4312 Argatroban (Exembol®)*. [Online]. Available from: http://www.awmsg.org/awmsgonline/grabber;jsessionid=1f631736a0503b59c6ed8236c4b8?resId=308.

Bakolis, I., Kelly, R., Fecht, D., Best, N., et al. (2016) Protective Effects of Smoke-free Legislation on Birth Outcomes in England: A Regression Discontinuity Design. *Epidemiology (Cambridge, Mass.)*. [Online] 27 (6), 810–818. Available from: doi:10.1097/EDE.0000000000000534.

Banbeta, A., van Rosmalen, J., Dejardin, D. & Lesaffre, E. (2019) Modified power prior with multiple historical trials for binary endpoints. *Statistics in Medicine*. [Online] 38 (7), 1147–1169. Available from: doi:10.1002/sim.8019.

Barker, I., Lloyd, T. & Steventon, A. (2016) Effect of a national requirement to introduce named accountable general practitioners for patients aged 75 or older in England: regression discontinuity analysis of general practice utilisation and continuity of care. *BMJ open*. [Online] 6 (9), e011422. Available from: doi:10.1136/bmjopen-2016-011422.

Becker, J., Lorenz, E., Haas, G., Helwig, C., et al. (2016) Evaluation of real world treatment outcomes in patients with metastatic merkel cell carcinoma (MCC) following second line chemotherapy. *Annals of Oncology*. [Online] 27 (suppl_6). Available from: doi:10.1093/annonc/mdw379.48 [Accessed: 13 April 2017].

Becker, J.C., Lorenz, E., Ugurel, S., Eigentler, T.K., et al. (2017) Evaluation of real-world treatment outcomes in patients with distant metastatic Merkel cell carcinoma following second-line chemotherapy in Europe. *Oncotarget*. [Online] 8 (45), 79731–79741. Available from: doi:10.18632/oncotarget.19218.

Bellmunt, J., Choueiri, T.K., Fougeray, R., Schutz, F.A.B., et al. (2010) Prognostic Factors in Patients With Advanced Transitional Cell Carcinoma of the Urothelial Tract Experiencing Treatment Failure With Platinum-Containing Regimens. *Journal of Clinical Oncology*. [Online] 28 (11), 1850–1855. Available from: doi:10.1200/JCO.2009.25.4599.

Benson, K. & Hartz, A.J. (2000) A Comparison of Observational Studies and Randomized, Controlled Trials. *New England Journal of Medicine*. [Online] 342 (25), 1878–1886. Available from: doi:10.1056/NEJM200006223422506.

Berger, M.L., Sox, H., Willke, R.J., Brixner, D.L., et al. (2017) Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value in Health*. [Online] 20 (8), 1003–1008. Available from: doi:10.1016/j.jval.2017.08.3019.

Black, N. (1994) Experimental and observational methods of evaluation. *BMJ*. [Online] 309 (6953), 540. Available from: doi:10.1136/bmj.309.6953.540a.

Borenstein, M., Hedges, L. & Rothstein, H. (2007) Meta-analysis: fixed effect vs. random effects. *Meta-Analysis. com*. [Online] Available from: http://www.meta-analysis.com/downloads/Meta-analysis%20fixed%20effect%20vs%20random%20effects.pdf [Accessed: 8 January 2016].

Briggs, A.H. & O'Brien, B.J. (2001) The death of cost-minimization analysis? *Health economics*. 10 (2), 179–184.

Britton, A., McKee, M., Black, N., McPherson, K., et al. (1998) Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment (Winchester, England)*. 2 (13), i–iv, 1–124.

Bullement, A., Nathan, P., Willis, A., Amin, A., et al. (2019) Cost Effectiveness of Avelumab for Metastatic Merkel Cell Carcinoma. *PharmacoEconomics - Open*. [Online] Available from: doi:10.1007/s41669-018-0115-y [Accessed: 25 January 2019].

Buxton, M.J., Drummond, M.F., Van Hout, B.A., Prince, R.L., et al. (1997) Modelling in economic evaluation: an unavoidable fact of life. *Health economics*. 6 (3), 217–227.

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., et al. (2016) Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*. [Online] 58 (1), 104–132. Available from: doi:10.1002/bimj.201400049.

Byar, D.P. (1980) Why Data Bases Should Not Replace Randomized Clinical Trials. *Biometrics*. [Online] 36 (2), 337. Available from: doi:10.2307/2529989.

Byar, D.P., Schoenfeld, D.A., Green, S.B., Amato, D.A., et al. (1990) Design Considerations for AIDS Trials. *New England Journal of Medicine*. [Online] 323 (19), 1343–1348. Available from: doi:10.1056/NEJM199011083231912.

Canadian Agency for Drugs and Technologies in Health (2009) *Addendum to CADTH's Guidelines for the Economic Evaluation of Health Technologies: Specific Guidance for Oncology Products*. [Online]. Available from: http://www.cadth.ca/media/pdf/H0405_Guidance_for_Oncology_Prodcuts_gr_e.pdf.

Canadian Agency for Drugs and Technologies in Health (2006) *Guidelines for the Economic Evaluation of Health Technologies: Canada, 3rd Edition*. [Online]. Available from: http://www.cadth.ca/media/pdf/186_EconomicGuidelines_e.pdf.

Caro, J.J., Briggs, A.H., Siebert, U. & Kuntz, K.M. (2012) Modeling Good Research Practices—Overview A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–1. *Medical Decision Making*. [Online] 32 (5), 667–677. Available from: doi:10.1177/0272989X12454577.

Caro, J.J. & Ishak, K.J. (2010) No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics*. 28 (10), 957–967.

Carrigan, G., Whipple, S., Capra, W.B., Taylor, M.D., et al. (2019) Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. *Clinical Pharmacology & Therapeutics*. [Online] Available from: doi:10.1002/cpt.1586 [Accessed: 5 August 2019].

Center for Drug Evaluation and Research (1997) *Guidance for industry - Guideline for the format and content of the microbiology section of an application*. [Online]. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM075101.pdf.

Chow, S.-C. & Chang, Y.-W. (2019) Statistical considerations for rare diseases drug development. *Journal of Biopharmaceutical Statistics*. [Online] 29 (5), 874–886. Available from: doi:10.1080/10543406.2019.1657441.

Ciociola, A.A., Cohen, L.B., Kulkarni, P. & FDA-Related Matters Committee of the American College of Gastroenterology (2014) How drugs are developed and approved by the FDA: current process and future directions. *The American Journal of Gastroenterology*. [Online] 109 (5), 620–623. Available from: doi:10.1038/ajg.2013.407.

Clark, L. (2017) The hidden, potentially life-threatening, cost Brexit could have on the UK. *Wired UK*. [Online]. Available from: https://www.wired.co.uk/article/brexit-impact-on-uk-drug-prices [Accessed: 1 February 2019].

Cochrane, A.L. (1972) *Effectiveness and efficiency: random reflections on health services*. Nuffield Provincial Hospitals Trust.

Colditz, G.A., Miller, J.N. & Mosteller, F. (1989) How study design affects outcomes in comparisons of therapy. I: Medical. *Statistics in Medicine*. [Online] 8 (4), 441–454. Available from: doi:10.1002/sim.4780080408.

Cowey, L., Becker, J. & Bharmal, M. (2016) *Retrospective Observational Study to Evaluate Treatment Outcomes in Patients with Metastatic Merkel Cell Carcinoma Following Chemotherapy Observational Study Protocol/Analysis Plan*.

CRASH Trial Collaborators (2004) Effect of intravenous corticosteroids on death within 14 days in 10 008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *The Lancet*. [Online] 364 (9442), 1321–1328. Available from: doi:10.1016/S0140-6736(04)17188-2.

Crofton, J. (2006) The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line. *Journal of the Royal Society of Medicine*. 99 (10), 531–534.

Cundy, T.P., Sierakowski, K., Manna, A., Cooper, C.M., et al. (2016) Fast-track surgery for uncomplicated appendicitis in children: a matched case-control study: Fast-track emergency surgery in children. *ANZ Journal of Surgery*. [Online] Available from: doi:10.1111/ans.13744 [Accessed: 22 September 2016].

Dagher, R., Johnson, J., Williams, G., Keegan, P., et al. (2004) Accelerated Approval of Oncology Products: A Decade of Experience. *JNCI Journal of the National Cancer Institute*. [Online] 96 (20), 1500–1509. Available from: doi:10.1093/jnci/djh279.

Dahabreh, I.J. & Kent, D.M. (2014) Can the Learning Health Care System Be Educated With Observational Data? *JAMA*. [Online] 312 (2), 129–130. Available from: doi:10.1001/jama.2014.4364.

Dakin, H. & Wordsworth, S. (2013) Cost-minimisation analysis versus cost-effectiveness analysis, revisited. *Health Economics*. [Online] 22 (1), 22–34. Available from: doi:10.1002/hec.1812.

Dallow, N., Best, N. & Montague, T.H. (2018) Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics*. [Online] 17 (4), 301–316. Available from: doi:10.1002/pst.1854.

Davies, J., Martinec, M., Delmar, P., Coudert, M., et al. (2018) Comparative effectiveness from a single-arm trial and real-world data: alectinib versus ceritinib. *Journal of Comparative Effectiveness Research*. [Online] 7 (9), 855–865. Available from: doi:10.2217/cer-2018-0032.

Davis, C., Naci, H., Gurpinar, E., Poplavska, E., et al. (2017) Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13. *BMJ*. [Online] 359, j4530. Available from: doi:10.1136/bmj.j4530.

Dejardin, D., van Rosmalen, J. & Lesaffre, E. (2014) *Including historical data in the analysis of clinical trials using the modified power prior: Practical considerations for survival models*.

Desai, J.R., Bowen, E.A., Danielson, M.M., Allam, R.R., et al. (2013) Creation and implementation of a historical controls database from randomized clinical trials. *Journal of the American Medical Informatics Association*. [Online] 20 (e1), e162–e168. Available from: doi:10.1136/amiajnl-2012-001257.

Dias, S., Sutton, A.J., Welton, N.J. & Ades, A.E. (2011) NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. *Sheffield: Decision Support Unit ScHARR*. 1–24.

Diehl, L.F. & Perry, D.J. (1986) A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? *Journal of Clinical Oncology*. 4 (7), 1114–1120.

Dimick, J.B. & Ryan, A.M. (2014) Methods for Evaluating Changes in Health Care Policy: The Difference-in-Differences Approach. *JAMA*. [Online] 312 (22), 2401. Available from: doi:10.1001/jama.2014.16153.

Djulbegovic, B., Glasziou, P., Klocksieben, F.A., Reljic, T., et al. (2018) Larger effect sizes in nonrandomized studies are associated with higher rates of EMA licensing approval. *Journal of Clinical Epidemiology*. [Online] 98, 24–32. Available from: doi:10/gdp3zt.

Downing, N.S., Aminawung, J.A., Shah, N.D., Braunstein, J.B., et al. (2012) Regulatory Review of Novel Therapeutics — Comparison of Three Regulatory Agencies. *New England Journal of Medicine*. [Online] 366 (24), 2284–2293. Available from: doi:10.1056/NEJMsa1200223.

Dron, L., Golchi, S., Hsu, G. & Thorlund, K. (2019) Minimizing control group allocation in randomized trials using dynamic borrowing of external control data – An application to second line therapy for non-small cell lung cancer. *Contemporary Clinical Trials Communications*. [Online] 16, 100446. Available from: doi:10.1016/j.conctc.2019.100446.

Duan, Y. (2005) *A modified bayesian power prior approach with applications in water quality evaluation*. [Online]. Virginia Polytechnic Institute and State University. Available from: http://scholar.lib.vt.edu/theses/available/etd-12072005-133505/ [Accessed: 10 June 2015].

Eckermann, S. & Pekarsky, B. (2014) Can the Real Opportunity Cost Stand Up: Displaced Services, the Straw Man Outside the Room. *PharmacoEconomics*. [Online] Available from: doi:10.1007/s40273-014-0140-3 [Accessed: 20 March 2014].

Egger, M., Moons, K.G.M., Fletcher, C. & GetReal Workpackage 4 (2016) GetReal: from efficacy in clinical trials to relative effectiveness in the real world: From Efficacy to Real-world Effectiveness. *Research Synthesis Methods*. [Online] 7 (3), 278–281. Available from: doi:10.1002/jrsm.1207.

Eichler, H.-G., Bloechl-Daum, B., Bauer, P., Bretz, F., et al. (2016) 'Threshold-crossing': A Useful Way to Establish the Counterfactual in Clinical Trials? *Clinical Pharmacology and Therapeutics*. [Online] 100 (6), 699–712. Available from: doi:10.1002/cpt.515.

Ellenberg, S.S. & Temple, R. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Annals of Internal Medicine*. 133 (6), 464–470.

Ellis, S.E., Speroff, T., Dittus, R.S., Brown, A., et al. (2004) Diabetes patient education: a meta-analysis and meta-regression. *Patient Education and Counseling*. 52 (1), 97–105.

Elze, M.C., Gregson, J., Baber, U., Williamson, E., et al. (2017) Comparison of Propensity Score Methods and Covariate Adjustment. *Journal of the American College of Cardiology*. [Online] 69 (3), 345–357. Available from: doi:10.1016/j.jacc.2016.10.060.

Emmanuel, E.J. & Miller, F.G. (2001) THE ETHICS OF PLACEBO-CONTROLLED TRIALS—AMiddle GROUND. *N Engl J Med*. [Online] 345 (12). Available from: http://or.org/pdf/Placebo_Ethics.pdf [Accessed: 20 March 2014].

European Medicines Agency (2014a) *Ask EMA - (ASK-3426) Potential error in lipegfilgrastim SPC*.

European Medicines Agency (2014b) *Lipegfilgrastim (Lonquex) SmPC - WC500148380*. [Online]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/002556/WC500148380.pdf [Accessed: 13 August 2014].

Evans, D. (2003) Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of clinical nursing*. 12 (1), 77–84.

Eyre, T.A., Phillips, E.H., Linton, K.M., Arumainathan, A., et al. (2017) Results of a multicentre UK-wide retrospective study evaluating the efficacy of brentuximab vedotin in relapsed, refractory classical

Hodgkin lymphoma in the transplant naive setting. *British Journal of Haematology*. [Online] 179 (3), 471–479. Available from: doi:10.1111/bjh.14898.

Faria, R., Alava, M.H., Manca, A. & Wailoo, A.J. (2015) *NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data*. [Online] Available from: http://www.nicedsu.org.uk/TSD17%20-%20DSU%20Observational%20data%20FINAL.pdf.

Fellow, J.P.H.S.S.V. & Director, S.G.F. (2008) *Cochrane Handbook for Systematic Reviews of Interventions*. [Online]. John Wiley & Sons, Ltd. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780470712184.fmatter/summary [Accessed: 31 October 2016].

Fisher, S.R.A. (1935) *The Design of Experiments*. Oliver and Boyd.

Food and Drug Administration (2001) *Complete review: BLA 99-0786 Campath*. [Online]. Available from: http://www.accessdata.fda.gov/drugsatfda_docs/nda/2000/103948_0000_Campath_Medr.pdf.

Food and Drug Administration (1989) *Ganciclovir Injection, for intravenous use. Prescribing Information*. [Online]. Available from: https://img.thebody.com/nih/prof/ganciclovir.pdf [Accessed: 31 December 2019].

Food and Drug Administration (2007) *Guidance for industry - Clinical trials endpoints for the approval of cancer drugs and biologics.pdf*.

Food and Drug Administration (2014) *History of the FDA*. [Online]. November 2014. Available from: http://www.fda.gov/aboutFDA/whatwedo/history/default.htm [Accessed: 7 November 2014].

Freedman, B. (1987) Equipoise and the ethics of clinical research. *The New England journal of medicine*. [Online] 317 (3), 141–145. Available from: doi:10.1056/NEJM198707163170304.

Freemantle, N., Marston, L., Walters, K., Wood, J., et al. (2013) Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ*. [Online] 347 (nov11 3), f6409–f6409. Available from: doi:10.1136/bmj.f6409.

French, J.A., Temkin, N.R., Shneker, B.F., Hammer, A.E., et al. (2011) Lamotrigine XR Conversion to Monotherapy: First Study Using a Historical Control Group. *Neurotherapeutics*. [Online] 9 (1), 176–184. Available from: doi:10.1007/s13311-011-0088-3.

French, J.A., Wang, S., Warnock, B. & Temkin, N. (2010) Historical control monotherapy design in the treatment of epilepsy: Historical Control Monotherapy. *Epilepsia*. [Online] 51 (10), 1936–1943. Available from: doi:10.1111/j.1528-1167.2010.02650.x.

Fugh-Berman, A. & Melnick, D. (2008) Off-Label Promotion, On-Target Sales. *PLoS Medicine*. [Online] 5 (10). Available from: doi:10.1371/journal.pmed.0050210 [Accessed: 30 November 2015].

Gabrio, A., Mason, A.J. & Baio, G. (2019) A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data: Handling structural ones and missingness in economic evaluations. *Statistics in Medicine*. [Online] 38 (8), 1399–1420. Available from: doi:10.1002/sim.8045.

Gaddipati, H., Liu, K., Pariser, A. & Pazdur, R. (2012) Rare Cancer Trial Design: Lessons from FDA Approvals. *Clinical Cancer Research*. [Online] 18 (19), 5172–5178. Available from: doi:10.1158/1078-0432.CCR-12-1135.

Geneletti, S., O'Keeffe, A.G., Sharples, L.D., Richardson, S., et al. (2015) Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine*. [Online] 34 (15), 2334–2352. Available from: doi:10.1002/sim.6486.

Geneletti, S., Ricciardi, F., O'Keeffe, A.G. & Baio, G. (2019) Bayesian modelling for binary outcomes in the regression discontinuity design. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. [Online] 182 (3), 983–1002. Available from: doi:10.1111/rssa.12440.

Gerstein, H.C., McMurray, J. & Holman, R.R. (2019) Real-world studies no substitute for RCTs in establishing efficacy. *The Lancet*. [Online] 393 (10168), 210–211. Available from: doi:10.1016/S0140-6736(18)32840-X.

Glasziou, P., Chalmers, I., Rawlins, M. & McCulloch, P. (2007) When Are Randomised Trials Unnecessary? Picking Signal from Noise. *BMJ: British Medical Journal*. 334 (7589), 349–351.

Goldacre, B. (2009) *Bad science*. London, Fourth estate.

Goodman, S.N. (1999) Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*. 130 (12), 995–1004.

Gopal, A.K., Kahl, B.S., de Vos, S., Wagner-Johnston, N.D., et al. (2014) PI3Kδ Inhibition by Idelalisib in Patients with Relapsed Indolent Lymphoma. *New England Journal of Medicine*. [Online] 370 (11), 1008–1018. Available from: doi:10.1056/NEJMoa1314583.

Goring, S., Taylor, A., Müller, K., Li, T.J.J., et al. (2019) Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. *BMJ Open*. [Online] 9 (2), e024895. Available from: doi:10.1136/bmjopen-2018-024895.

Gosling, J.P. (2018) SHELF: The Sheffield Elicitation Framework. In: Luis C. Dias, Alec Morton, & John Quigley (eds.). *Elicitation: The Science and Art of Structuring Judgement*. International Series in Operations Research & Management Science. [Online]. Cham, Springer International Publishing. pp. 61–93. Available from: doi:10.1007/978-3-319-65052-4_4 [Accessed: 5 August 2019].

Grayling, M.J., Dimairo, M., Mander, A.P. & Jaki, T.F. (2019) A Review of Perspectives on the Use of Randomization in Phase II Oncology Trials. *JNCI: Journal of the National Cancer Institute*. [Online] 111 (12), 1255–1262. Available from: doi:10.1093/jnci/djz126.

Grayling, M.J. & Mander, A.P. (2016) Do single-arm trials have a role in drug development plans incorporating randomised trials? *Pharmaceutical Statistics*. [Online] 15 (2), 143–151. Available from: doi:10.1002/pst.1726.

Grayling, M.J. & Wason, J.M. (2020) A web application for the design of multi-arm clinical trials. *BMC cancer*. [Online] 20 (1), 80. Available from: doi:10.1186/s12885-020-6525-0.

Greenland, S. & Robins, J.M. (2009) Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*. [Online] 6 (1), 1–9. Available from: doi:10.1186/1742-5573-6-4.

Greenland, S. & Robins, J.M. (1986) Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*. [Online] 15 (3), 413–419. Available from: doi:10.1093/ije/15.3.413.

Grieve, R., Abrams, K., Claxton, K., Goldacre, B., et al. (2016) Cancer Drugs Fund requires further reform. *BMJ*. [Online] i5090. Available from: doi:10.1136/bmj.i5090.

Griffiths, E.A., Macaulay, R., Vadlamudi, N.K., Uddin, J., et al. (2017) The Role of Noncomparative Evidence in Health Technology Assessment Decisions. *Value in Health*. [Online] 20 (10), 1245–1251. Available from: doi:10.1016/j.jval.2017.06.015.

Grigore, B., Peters, J., Hyde, C. & Stein, K. (2016) A comparison of two methods for expert elicitation in health technology assessments. *BMC Medical Research Methodology*. [Online] 16, 85. Available from: doi:10.1186/s12874-016-0186-3.

Gsteiger, S., Neuenschwander, B., Mercier, F. & Schmidli, H. (2013) Using historical control information for the design and analysis of clinical trials with overdispersed count data. *Statistics in Medicine*. [Online] 32 (21), 3609–3622. Available from: doi:10.1002/sim.5851.

Guyot, P., Ades, A., Ouwens, M.J. & Welton, N.J. (2012) Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*. [Online] 12, 9. Available from: doi:10.1186/1471-2288-12-9.

Haim, N., Ben-Shahar, M. & Epelbaum, R. (1995) Prolonged daily administration of oral etoposide in lymphoma following prior therapy with adriamycin, an ifosfamide-containing salvage combination, and intravenous etoposide. *Cancer Chemotherapy and Pharmacology*. 36 (4), 352–355.

Hasler, J. (2010) *5 Tales of Survival from Extreme Falls*. [Online]. 29 January 2010. Popular Mechanics. Available from: http://www.popularmechanics.com/outdoors/survival/stories/4344037 [Accessed: 7 November 2014].

Hatswell, A.J. (2017) How do we avoid disaster when exiting the European Medicines Agency? Making the most of Brexit in pharmaceutical regulation. *ecancermedicalscience*. [Online] 11. Available from: doi:10.3332/ecancer.2017.ed67.

Hatswell, A.J., Baio, G., Berlin, J.A., Irs, A., et al. (2016) Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. *BMJ Open*. [Online] 6 (6), e011666. Available from: doi:10.1136/bmjopen-2016-011666.

Hatswell, A.J., Baio, G. & Freemantle, N. (2017) *Research Report number 327: A description of the circumstances surrounding pharmaceutical approvals by the FDA and EMA from 1999 to 2014 made without randomised control trial data*. [Online]. Available from: https://www.ucl.ac.uk/drupal/site_statistics/sites/statistics/files/migrated-files/rr327.pdf.

Hatswell, A.J., Bardou, M., Gallagher, M. & Beckerman, R. (2014) Modeling Alchemy: The Impact of Unorthodox Trial Design on Health Technology Appraisal Strategy. *ISPOR Connections*. 20 (4), 6–9.

Hatswell, A.J., Bullement, A., Briggs, A., Paulden, M., et al. (2018) Probabilistic Sensitivity Analysis in Cost-Effectiveness Models: Determining Model Convergence in Cohort Models. *PharmacoEconomics*. [Online] 36 (12), 1421–1426. Available from: doi:10.1007/s40273-018-0697-3.

Hatswell, A.J., Burns, D., Baio, G. & Wadelin, F. (2019) Frequentist and Bayesian meta-regression of health state utilities for multiple myeloma incorporating systematic review and analysis of individual patient data. *Health Economics*. [Online] Available from: doi:10.1002/hec.3871 [Accessed: 25 February 2019].

Hatswell, A.J., Freemantle, N. & Baio, G. (2017a) Economic Evaluations of Pharmaceuticals Granted a Marketing Authorisation Without the Results of Randomised Trials: A Systematic Review and Taxonomy. *PharmacoEconomics*. [Online] 35 (2), 163–176. Available from: doi:10.1007/s40273-016-0460-6.

Hatswell, A.J., Freemantle, N. & Baio, G. (2017b) *Research Report number 326: A description of economic models constructed for pharmaceuticals granted a marketing authorisation without a randomised controlled trial by the FDA and EMA from 1999 to 2014*. [Online]. Available from: https://www.ucl.ac.uk/drupal/site_statistics/sites/statistics/files/migrated-files/rr326.pdf.

Hatswell, A.J., Freemantle, N. & Baio, G. (2020) The Effects of Model Misspecification in Unanchored Matching-Adjusted Indirect Comparison (MAIC): Results of a Simulation Study. *Value in Health*. [Online] 0 (0). Available from: doi:10.1016/j.jval.2020.02.008 [Accessed: 27 May 2020].

Hatswell, A.J. & Porter, J.K. (2018) Reducing Drug Wastage in Pharmaceuticals Dosed by Weight or Body Surface Areas by Optimising Vial Sizes. *Applied Health Economics and Health Policy*. [Online] Available from: doi:10.1007/s40258-018-0444-0 [Accessed: 11 January 2019].

Hatswell, A.J. & Sullivan, W.G. (2019) Creating historical controls using data from a previous line of treatment – Two non-standard approaches. *Statistical Methods in Medical Research*. [Online] 0962280219826609. Available from: doi:10.1177/0962280219826609.

Hatswell, A.J., Thompson, G.J., Maroudas, P.A., Sofrygin, O., et al. (2017) Estimating outcomes and cost effectiveness using a single-arm clinical trial: ofatumumab for double-refractory chronic lymphocytic leukemia. *Cost Effectiveness and Resource Allocation*. [Online] 15, 8. Available from: doi:10.1186/s12962-017-0071-x.

Hernán, M.A. & Robins, J.M. (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology*. [Online] kwv254. Available from: doi:10.1093/aje/kwv254.

Hill, A.B. (1965) The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*. 58 (5), 295–300.

Ho, D.E., Imai, K., King, G. & Stuart, E.A. (2011) **MatchIt** : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*. [Online] 42 (8). Available from: doi:10.18637/jss.v042.i08 [Accessed: 27 May 2020].

Hobbs, B.P., Sargent, D.J. & Carlin, B.P. (2012) Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis*. [Online] 7 (3), 639–674. Available from: doi:10.1214/12-BA722.

Hopper, C., Niziol, C. & Sidhu, M. (2004) The cost-effectiveness of Foscan mediated photodynamic therapy (Foscan-PDT) compared with extensive palliative surgery and palliative chemotherapy for patients with advanced head and neck cancer in the UK. *Oral Oncology*. [Online] 40 (4), 372–382. Available from: doi:10.1016/j.oraloncology.2003.09.003.

Hoyle, M., Snowsill, T., Haasova, M., Cooper, C., et al. (2013) *Bosutinib for previously treated chronic myeloid laukaemia: a single technology appraisal*. [Online]. Available from: https://njl-admin.nihr.ac.uk/document/download/2006138.

Hrobjartsson, A., Thomsen, A.S.S., Emanuelsson, F., Tendal, B., et al. (2012) Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*. [Online] 344 (feb27 2), e1119–e1119. Available from: doi:10.1136/bmj.e1119.

Ibrahim, J.G. & Chen, M.-H. (2000) Power prior distributions for regression models. *Statistical Science*. [Online] 15 (1), 46–60. Available from: doi:10.1214/ss/1009212673.

Ibrahim, J.G., Chen, M.-H., Gwon, Y. & Chen, F. (2015) The power prior: theory and applications. *Statistics in Medicine*. [Online] 34 (28), 3724–3749. Available from: doi:10.1002/sim.6728.

ICH Harmonised Tripartite (2000) ICH Topic E 10 - Choice of Control Group and Related Issues in Clinical Trials. *Choice*. E10.

Ioannidis, J.P., Haidich, A.B., Pappa, M., Pantazis, N., et al. (2001) Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 286 (7), 821–830.

Ip, S., Paulus, J.K., Balk, E.M., Dahabreh, I.J., et al. (2013) *Role of Single Group Studies in Agency for Healthcare Research and Quality Comparative Effectiveness Reviews*. AHRQ Methods for Effective Health Care. [Online]. Rockville (MD), Agency for Healthcare Research and Quality (US). Available from: http://www.ncbi.nlm.nih.gov/books/NBK121314/ [Accessed: 25 March 2015].

Ishak, K.J., Proskorovsky, I. & Benedict, A. (2015) Simulation and Matching-Based Approaches for Indirect Comparison of Treatments. *PharmacoEconomics*. [Online] Available from: doi:10.1007/s40273-015-0271-1 [Accessed: 13 April 2015].

Ishak, K.J., Proskorovsky, I., Korytowsky, B., Sandin, R., et al. (2014) Methods for Adjusting for Bias Due to Crossover in Oncology Trials. *PharmacoEconomics*. [Online] 32 (6), 533–546. Available from: doi:10.1007/s40273-014-0145-y.

Isogawa, N., Takeda, K., Maruo, K. & Daimon, T. (2019) A Comparison Between a Meta-analytic Approach and Power Prior Approach to Using Historical Control Information in Clinical Trials With Binary Endpoints. *Therapeutic Innovation & Regulatory Science*. [Online] 2168479019862531. Available from: doi:10.1177/2168479019862531.

ISPOR (2020) *ISPOR Meetings Profile*. [Online]. March 2020. Available from: http://www.ispor.org/MeetingsProfile.asp [Accessed: 8 August 2014].

Iyer, J.G., Blom, A., Doumani, R., Lewis, C., et al. (2016) Response rates and durability of chemotherapy among 62 patients with metastatic Merkel cell carcinoma. *Cancer Medicine*. [Online] 5 (9), 2294–2301. Available from: doi:10.1002/cam4.815.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An Introduction to Statistical Learning*. Springer Texts in Statistics. [Online]. New York, NY, Springer New York. Available from: doi:10.1007/978-1-4614-7138-7 [Accessed: 20 March 2017].

Jefferys, D.B. & Jones, K.H. (1995) EMEA and the new pharmaceutical procedures for Europe. European Medicines Evaluation Agency. *European Journal of Clinical Pharmacology*. 47 (6), 471–476.

Jiao, F., Tu, W., Jimenez, S., Crentsil, V., et al. (2019) Utilizing shared internal control arms and historical information in small-sized platform clinical trials. *Journal of Biopharmaceutical Statistics*. [Online] 1–15. Available from: doi:10.1080/10543406.2019.1657132.

Johnston, I. (2017) Brexit could have 'catastrophic' effect on Britons' access to life-saving drugs, expert warns. *The Independent*. [Online] 27 June. Available from: http://www.independent.co.uk/news/uk/politics/brexit-latest-news-life-saving-drugs-access-catastrophic-effect-healthcare-treatment-expert-ema-mhra-a7810896.html [Accessed: 1 February 2019].

Jonsson, B., Martinalbo, J. & Pignatti, F. (2017) European Medicines Agency Perspective on Oncology Study Design for Marketing Authorization and Beyond. *Clinical Pharmacology & Therapeutics*. [Online] 101 (5), 577–579. Available from: doi:10.1002/cpt.612.

Kaptchuk, T.J. (2001) The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? *Journal of clinical epidemiology*. 54 (6), 541–549.

Kaufman, H.L., Russell, J., Hamid, O., Bhatia, S., et al. (2016) Avelumab in patients with chemotherapy-refractory metastatic Merkel cell carcinoma: a multicentre, single-group, open-label, phase 2 trial. *The Lancet Oncology*. [Online] 17 (10), 1374–1385. Available from: doi:10.1016/S1470-2045(16)30364-3.

Keating, M.J., Flinn, I., Jain, V., Binet, J.-L., et al. (2002) Therapeutic role of alemtuzumab (Campath-1H) in patients who have failed fludarabine: results of a large international study. *Blood*. 99 (10), 3554–3561.

Kish, T., Aziz, A. & Sorio, M. (2017) Hepatitis C in a New Era: A Review of Current Therapies. *Pharmacy and Therapeutics*. 42 (5), 316–329.

Korn, E.L., Liu, P.-Y., Lee, S.J., Chapman, J.-A.W., et al. (2008) Meta-Analysis of Phase II Cooperative Group Trials in Metastatic Stage IV Melanoma to Determine Progression-Free and Overall Survival Benchmarks for Future Phase II Trials. *Journal of Clinical Oncology*. [Online] 26 (4), 527–534. Available from: doi:10.1200/JCO.2007.12.7837.

Kübler, A., Niziol, C., Sidhu, M., Dünne, A., et al. (2005) Eine Kosten-Effektivitäts-Analyse der photodynamischen Therapie mit Foscan® (Foscan®-PDT) im Vergleich zu einer palliativen Chemotherapie bei Patienten mit fortgeschrittenen Kopf-Halstumoren in Deutschland. *Laryngo-Rhino-Otologie*. [Online] 84 (10), 725–732. Available from: doi:10.1055/s-2005-861048.

Kuhn, T.S. (1996) *The Structure of Scientific Revolutions*. New ed of 3 Revised ed edition. Chicago, IL, University of Chicago Press.

Lachin, J.M. (1981) Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*. [Online] 2 (2), 93–113. Available from: doi:10.1016/0197-2456(81)90001-5.

Latimer, N. (2011) *NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data*. [Online]. Available from: http://www.nicedsu.org.uk/NICE%20DSU%20TSD%20Survival%20analysis.updated%20March%202013.v2.pdf [Accessed: 8 January 2016].

Leurent, B., Gomes, M., Faria, R., Morris, S., et al. (2018) Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. *PharmacoEconomics*. [Online] 36 (8), 889–901. Available from: doi:10.1007/s40273-018-0650-5.

Lévesque, L.E., Hanley, J.A., Kezouh, A. & Suissa, S. (2010) Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. [Online] 340. Available from: doi:10.1136/bmj.b5087 [Accessed: 10 March 2020].

Lewis, C.J., Sarkar, S., Zhu, J. & Carlin, B.P. (2019) Borrowing From Historical Control Data in Cancer Drug Development: A Cautionary Tale and Practical Guidelines. *Statistics in Biopharmaceutical Research*. [Online] 11 (1), 67–78. Available from: doi:10.1080/19466315.2018.1497533.

Light, D.W. & Lexchin, J. (2015) Why do cancer drugs get such an easy ride? *BMJ*. [Online] 350 (apr23 1), h2068–h2068. Available from: doi:10.1136/bmj.h2068.

Lim, J., Walley, R., Yuan, J., Liu, J., et al. (2018) Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities.

Therapeutic Innovation & Regulatory Science. [Online] 52 (5), 546–559. Available from: doi:10.1177/2168479018778282.

Lipsky, M.S. & Sharp, L.K. (2001) From idea to market: the drug approval process. *The Journal of the American Board of Family Practice*. 14 (5), 362–367.

Little, R., Wittes, R.E., Longo, D.L. & Wilson, W.H. (1998) Vinblastine for recurrent Hodgkin's disease following autologous bone marrow transplant. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*. [Online] 16 (2), 584–588. Available from: doi:10.1200/JCO.1998.16.2.584.

MacLehose, R.R., Reeves, B.C., Harvey, I.M., Sheldon, T.A., et al. (2000) A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment (Winchester, England)*. 4 (34), 1–154.

Makady, A., Ham, R.T., de Boer, A., Hillege, H., et al. (2017) Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*. [Online] 20 (4), 520–532. Available from: doi:10.1016/j.jval.2016.12.003.

Mathes, T. & Pieper, D. (2017) Clarifying the distinction between case series and cohort studies in systematic reviews of comparative studies: potential impact on body of evidence and workload. *BMC Medical Research Methodology*. [Online] 17 (1). Available from: doi:10.1186/s12874-017-0391-8 [Accessed: 17 August 2017].

Mazumdar, M., Fazzari, M. & Panageas, K.S. (2001) A standardization method to adjust for the effect of patient selection in phase II clinical trials. *Statistics in Medicine*. [Online] 20 (6), 883–892. Available from: doi:10.1002/sim.706.

McConnell, K.J. & Lindner, S. (2019) Estimating treatment effects with machine learning. *Health Services Research*. [Online] 0 (0). Available from: doi:10.1111/1475-6773.13212 [Accessed: 19 October 2019].

Mead, G., Harker, G., Kushlan, P. & Rosenberg, S. (1982) Single agent palliative chemotherapy for end-stage Hodgkin's disease. *Cancer*. 50, 829–835.

Meier, P. (1975) Statistics and Medical Experimentation. *Biometrics*. [Online] 31 (2), 511. Available from: doi:10.2307/2529434.

Miguel, L.S., Augustin, U., Busse, R., Knai, C., et al. (2014) Recognition of pharmaceutical prescriptions across the European Union: A comparison of five Member States' policies and practices. *Health Policy*. [Online] 116 (2), 206–213. Available from: doi:10.1016/j.healthpol.2013.11.003.

Miller, B.W., Przepiorka, D. & de Claro, R.A. (2014) *Clinical Review NDA 205858*. [Online]. May 2014. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/205858Orig1s000MedR.pdf [Accessed: 24 March 2018].

Miller, F.G. & Joffe, S. (2011) Equipoise and the Dilemma of Randomized Clinical Trials. *New England Journal of Medicine*. [Online] 364 (5), 476–480. Available from: doi:10.1056/NEJMsb1011301.

Milne, I. (2012) Who was James Lind, and what exactly did he achieve. *Journal of the Royal Society of Medicine*. [Online] 105 (12), 503–508. Available from: doi:10.1258/jrsm.2012.12k090.

Moroz, V., Wilson, J.S., Kearns, P. & Wheatley, K. (2014) Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment. *Trials*. 15 (1), 481.

Moskowitz, C.H., Nademanee, A., Masszi, T., Agura, E., et al. (2015) Brentuximab vedotin as consolidation therapy after autologous stem-cell transplantation in patients with Hodgkin's lymphoma at risk of relapse or progression (AETHERA): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet (London, England)*. [Online] 385 (9980), 1853–1862. Available from: doi:10.1016/S0140-6736(15)60165-9.

National Institute for Health (2014) *Fact Sheet - MEDLINE, PubMed, and PMC (PubMed Central): How are they different?* [Online]. 7 May 2014. Available from: http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html [Accessed: 9 November 2014].

Netzer, T. (2006) European Union centralised procedure for marketing authorisation of oncology drugs: An in-depth review of its efficiency. *European Journal of Cancer*. [Online] 42 (4), 446–455. Available from: doi:10.1016/j.ejca.2005.04.045.

Neuenschwander, B., Branson, M. & Spiegelhalter, D.J. (2009) A note on the power prior. *Statistics in Medicine*. [Online] 28 (28), 3562–3566. Available from: doi:10.1002/sim.3722.

New Zealand Government (2015) *Guidelines for funding applications to PHARMAC*. [Online]. New Zealand Government. Available from: http://www.pharmac.govt.nz/2010/02/11/Guidelines%20for%20Suppliers%20Submissions.pdf.

Nghiem, P., Kaufman, H.L., Bharmal, M., Mahnke, L., et al. (2017) Systematic literature review of efficacy, safety and tolerability outcomes of chemotherapy regimens in patients with metastatic Merkel cell carcinoma. *Future Oncology*. [Online] Available from: doi:10.2217/fon-2017-0072 [Accessed: 4 May 2017].

NICE (2013) *Guide to the methods of technology appraisal 2013*. [Online]. NICE. Available from: https://www.nice.org.uk/process/pmg9/chapter/foreword.

Nikolakopoulos, S., Tweel, I. van der & Roes, K.C.B. (2018) Dynamic borrowing through empirical power priors that control type I error. *Biometrics*. [Online] 74 (3), 874–880. Available from: doi:10.1111/biom.12835.

Normington, J., Zhu, J., Mattiello, F., Sarkar, S., et al. (2019) An efficient Bayesian platform trial design for borrowing adaptively from historical control data in lymphoma. *Contemporary Clinical Trials*. [Online] 105890. Available from: doi:10.1016/j.cct.2019.105890.

Ollier, A., Morita, S., Ursino, M. & Zohar, S. (2019) An adaptive power prior for sequential clinical trials – Application to bridging studies. *Statistical Methods in Medical Research*. [Online] 0962280219886609. Available from: doi:10.1177/0962280219886609.

Oxford Dictionaries (2010) *Oxford Dictionary of English*. 3rd edition. New York, NY, OUP Oxford.

Parmar, M.K., Sydes, M.R., Cafferty, F.H., Choodari-Oskooei, B., et al. (2017) Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. *Clinical Trials (London, England)*. [Online] 14 (5), 451–461. Available from: doi:10.1177/1740774517725697.

Paulden, M., McCabe, C. & Karnon, J. (2014) Achieving Allocative Efficiency in Healthcare: Nice in Theory, not so NICE in Practice? *PharmacoEconomics*. [Online] 32 (4), 315–318. Available from: doi:10.1007/s40273-014-0146-x.

Paz-Ares, L., del Muro, J.G., Grande, E. & Díaz, S. (2010) A cost-effectiveness analysis of sunitinib in patients with metastatic renal cell carcinoma intolerant to or experiencing disease progression on immunotherapy: perspective of the Spanish National Health System. *Journal of Clinical Pharmacy and Therapeutics*. [Online] Available from: doi:10.1111/j.1365-2710.2009.01135.x [Accessed: 27 November 2014].

PBAC (2015) *Public summary document - November 2015 Meeting - Idelalisib*. [Online]. Available from: http://www.pbs.gov.au/industry/listing/elements/pbac-meetings/psd/2015-11/files/idelalisib-follicular-lymphoma-psd-november-2015.pdf [Accessed: 21 December 2017].

Pearce, W., Raman, S. & Turner, A. (2015) Randomised trials in context: practical problems and social aspects of evidence-based medicine and policy. *Trials*. [Online] 16 (1), 394. Available from: doi:10.1186/s13063-015-0917-5.

Penrose, S.R. (2006) *The Road To Reality: A Complete Guide to the Laws of the Universe*. New Ed edition. London, Vintage.

Petto, H., Kadziola, Z., Brnabic, A., Saure, D., et al. (2019) Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons via a Common Comparator. *Value in Health*. [Online] 22 (1), 85–91. Available from: doi:10.1016/j.jval.2018.06.018.

Pharmaceutical Management Agency (2012) *Prescription for Pharmacoeconomic analysis - Methods for cost-utility analysis version 2.1*. [Online]. Available from: https://www.pharmac.health.nz/assets/pfpa-final.pdf.

Phillippo, D., Ades, A.E., Dias, S., Palmer, S., et al. (2016) *NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE*.

Phillippo, D.M., Ades, A.E., Dias, S., Palmer, S., et al. (2017) Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. *Medical Decision Making*. [Online] 0272989X17725740. Available from: doi:10.1177/0272989X17725740.

Phillippo, D.M., Dias, S., Elsada, A., Ades, A.E., et al. (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*. [Online] 35 (3), 221–228. Available from: doi:10.1017/S0266462319000333.

Phillips, B., Ball, C., Sackett, D., Badenoch, D., et al. (2009) Oxford Centre for Evidence-based Medicine - Levels of Evidence. *CEBM*. [Online]. Available from: http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/ [Accessed: 25 February 2015].

Pipe, S.W., Shima, M., Lehle, M., Shapiro, A., et al. (2019) Efficacy, safety, and pharmacokinetics of emicizumab prophylaxis given every 4 weeks in people with haemophilia A (HAVEN 4): a multicentre, open-label, non-randomised phase 3 study. *The Lancet Haematology*. [Online] 6 (6), e295–e305. Available from: doi:10.1016/S2352-3026(19)30054-7.

Pocock, S.J. (1976) The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*. 29 (3), 175–188.

Powell, M. (2000) Licensing new antibacterial agents - a European perspective. *International Journal of Antimicrobial Agents*. 16 (3), 199–203.

Prasad, V. & Oseran, A. (2015) Do we need randomised trials for rare cancers? *European Journal of Cancer*. [Online] 51 (11), 1355–1357. Available from: doi:10.1016/j.ejca.2015.04.015.

Project Data Sphere (2015) *Mission & Vision*. [Online]. January 2015. Available from: https://projectdatasphere.org/projectdatasphere/html/mission [Accessed: 23 January 2015].

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. [Online]. Vienna, Austria, R Foundation for Statistical Computing. Available from: https://www.R-project.org.

Rafia, R., Simpson, E., Stevenson, M. & Papaioannou, D. (2013) Trabectedin for the Treatment of Advanced Metastatic Soft Tissue Sarcoma: A NICE Single Technology Appraisal. *PharmacoEconomics*. [Online] 31 (6), 471–478. Available from: doi:10.1007/s40273-013-0044-7.

Rawlins, M. (2008) De Testimonio: on the evidence for decisions about the use of therapeutic interventions. *Clinical Medicine*. [Online] 8 (6), 579–588. Available from: doi:10.7861/clinmedicine.8-6-579.

Rawlins, P.S.M. (2013) *What Constitutes Credible Evidence of Effectiveness*. [Online]. Available from: https://www.ohe.org/system/files/private/publications/385%20-%20What%20Constitutes%20Credible%20Evidence%202012AnnLec%20Rawlins%202013p.pdf?download=1.

Relton, C., Torgerson, D., O'Cathain, A. & Nicholl, J. (2010) Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. *BMJ*. [Online] 340. Available from: doi:10.1136/bmj.c1066 [Accessed: 31 December 2019].

Roberts, S.A., Allen, J.D. & Sigal, E.V. (2011) Despite criticism of the FDA review process, new cancer drugs reach patients sooner in the United States than in Europe. *Health Affairs (Project Hope)*. [Online] 30 (7), 1375–1381. Available from: doi:10.1377/hlthaff.2011.0231.

Rosenbaum, P.R. & Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. [Online] 70 (1), 41. Available from: doi:10.2307/2335942.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., et al. (2018) Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*. [Online] 27 (10), 3167–3182. Available from: doi:10.1177/0962280217694506.

Rovithis, D. (2013) Do health economic evaluations using observational data provide reliable assessment of treatment effects? *Health economics review*. 3 (1), 1–7.

Rubinstein, L., Crowley, J., Ivy, P., LeBlanc, M., et al. (2009) Randomized Phase II Designs. *Clinical Cancer Research*. [Online] 15 (6), 1883–1890. Available from: doi:10.1158/1078-0432.CCR-08-2031.

Saccà, L. (2010) The uncontrolled clinical trial: scientific, ethical, and practical reasons for being. *Internal and Emergency Medicine*. [Online] 5 (3), 201–204. Available from: doi:10.1007/s11739-010-0355-z.

Sacks, Henry., Chalmers, T.C.. & Smith Jr., Harry. (1982) Randomized versus historical controls for clinical trials. *The American Journal of Medicine*. [Online] 72 (2), 233–240. Available from: doi:10.1016/0002-9343(82)90815-4.

Salles, G., Schuster, S.J., de Vos, S., Wagner-Johnston, N.D., et al. (2017) Efficacy and safety of idelalisib in patients with relapsed, rituximab- and alkylating agent-refractory follicular lymphoma: a subgroup analysis of a phase 2 study. *Haematologica*. [Online] 102 (4), e156–e159. Available from: doi:10.3324/haematol.2016.151738.

Samlowski, W.E., Moon, J., Tuthill, R.J., Heinrich, M.C., et al. (2010) A phase II trial of imatinib mesylate in merkel cell carcinoma (neuroendocrine carcinoma of the skin): A Southwest Oncology Group study (S0331). *American Journal of Clinical Oncology*. [Online] 33 (5), 495–499. Available from: doi:10.1097/COC.0b013e3181b9cf04.

Schmidli, H., Häring, D., Thomas, M., Cassidy, A., et al. (2019) Beyond randomized clinical trials: Use of external controls. *Clinical Pharmacology and Therapeutics*. [Online] Available from: doi:10.1002/cpt.1723.

Schmidli, H., Wandel, S. & Neuenschwander, B. (2012) The network meta-analytic-predictive approach to non-inferiority trials. *Statistical Methods in Medical Research*. [Online] 22 (2), 219–240. Available from: doi:10.1177/0962280211432512.

Schoenfeld, D.A., Finkelstein, D.M., Macklin, E., Zach, N., et al. (2019) Design and analysis of a clinical trial using previous trials as historical control. *Clinical Trials*. [Online] 1740774519858914. Available from: doi:10.1177/1740774519858914.

Schulz KF, Chalmers I, Hayes RJ & Altman DG (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. [Online] 273 (5), 408–412. Available from: doi:10.1001/jama.1995.03520290060030.

Schulz, K.F. & Grimes, D.A. (2002) Blinding in randomised trials: hiding who got what. *The Lancet*. 359 (9307), 696–700.

Scottish Medicines Consortium (2011) *2nd Re-Submission In Confidence - Trabectedin, 0.25 and 1mg powder for concentrate for solution for infusion (Yondelis)*. [Online]. Available from: https://www.scottishmedicines.org.uk/files/advice/trabectedin_Yondelis_2ND_RESUBMISSION_FINAL_JUNE_2011_for_website.pdf.

Scottish Medicines Consortium (2012) *Argatroban, 100mg/ml, concentrate for solution for infusion (Exembol) SMC No. (812/12)*. [Online]. Available from: https://www.scottishmedicines.org.uk/files/advice/argatroban_Exembol_FINAL_October_2012_for_website.pdf.

Scottish Medicines Consortium (2014) *Guidance to Manufacturers for Completion of New Product Assessment Form (NPAF)*. [Online]. Available from: https://www.scottishmedicines.org.uk/files/submissionprocess/Guidance_on_NPAF_Final_October_2014.doc.

Scottish Medicines Consortium (2007) *Sunitinib 50mg capsules (Sutent) No. (343/07)*. [Online]. Available from: https://www.scottishmedicines.org.uk/files/sunitinib_Sutent_MRCC_343_07.pdf.

Scottish Medicines Consortium (2010) *Trabectedin, 0.25 and 1mg powder for concentrate for solution for infusion (Yondelis)*. [Online]. Available from: https://www.scottishmedicines.org.uk/files/advice/trabectedin_Yondelis_RESUBMISSION_FINAL_October_2010.doc_for_website.pdf.

Senderowicz, A.M. & Pfaff, O. (2014) Similarities and differences in the oncology drug approval process between FDA and European Union with emphasis on in vitro companion diagnostics. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*. [Online] 20 (6), 1445–1452. Available from: doi:10.1158/1078-0432.CCR-13-1761.

Seymour, L., Ivy, S.P., Sargent, D., Spriggs, D., et al. (2010) The Design of Phase II Clinical Trials Testing Cancer Therapeutics: Consensus Recommendations from the Clinical Trial Design Task Force of the National Cancer Institute Investigational Drug Steering Committee. *Clinical Cancer Research*. [Online] 16 (6), 1764–1769. Available from: doi:10.1158/1078-0432.CCR-09-3287.

Shadish, W.R. (2013) Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology*. [Online] 9 (2), 129–144. Available from: doi:10.1007/s11292-012-9166-8.

Shafrin, J., Shrestha, A., Chandra, A., Erder, M.H., et al. (2017) Evaluating Matching-Adjusted Indirect Comparisons in Practice: A Case Study of Patients with Attention-Deficit/Hyperactivity Disorder. *Health Economics*. [Online] 26 (11), 1459–1466. Available from: doi:10.1002/hec.3408.

Shah, R.R., Roberts, S.A. & Shah, D.R. (2013) A fresh perspective on comparing the FDA and the CHMP/EMA: approval of antineoplastic tyrosine kinase inhibitors: FDA *vs* . EU approvals of TKIs. *British Journal of Clinical Pharmacology*. [Online] 76 (3), 396–411. Available from: doi:10.1111/bcp.12085.

Shepshelovich, D., Tibau, A., Goldvaser, H., Molto, C., et al. (2018) Postmarketing Modifications of Drug Labels for Cancer Drugs Approved by the US Food and Drug Administration Between 2006 and 2016 With and Without Supporting Randomized Controlled Trials. *Journal of Clinical Oncology*. [Online] 36 (18), 1798–1804. Available from: doi:10.1200/JCO.2017.77.5593.

Siegel, J.E., Torrance, G.W., Russell, L.B., Luce, B.R., et al. (1997) Guidelines for pharmacoeconomic studies. *Pharmacoeconomics*. 11 (2), 159–168.

SIGN (2011) *SIGN 50: A Guideline Developer's Handbook*. First published 2008, revised 2011.

Signorovitch, J.E., Wu, E.Q., Andrew, P.Y., Gerrits, C.M., et al. (2010) Comparative Effectiveness Without Head-to-Head Trials. *Pharmacoeconomics*. 28 (10), 935–945.

Signorovitch, J.E., Wu, E.Q., Swallow, E., Kantor, E., et al. (2011) Comparative efficacy of vildagliptin and sitagliptin in Japanese patients with type 2 diabetes mellitus. *Clinical drug investigation*. 31 (9), 665–674.

Simpson, E., Rafia, R. & Stevenson, M. (2009) *Trabectedin for the treatment of advanced metastatic soft tissue sarcoma - Evidence review Group Report*. [Online]. Available from: https://www.nice.org.uk/guidance/TA185/documents/evidence-review-group-report2.

Simpson, E., Rafia, R., Stevenson, M. & Papaioannou, D. (2010) Trabectedin for the treatment of advanced metastatic soft tissue sarcoma. *Health Technol Assess*. [Online] 14 Suppl 1. Available from: doi:10.3310/hta14suppl1/09 [Accessed: 25 November 2014].

Simpson, J. (2014) Eating Too Much Rice Almost Sank the Japanese Navy. *War Is Boring*. [Online]. Available from: https://medium.com/war-is-boring/eating-too-much-rice-almost-sank-the-japanese-navy-f985772c81a6 [Accessed: 1 June 2019].

Smith, G.C.S. & Pell, J.P. (2003) Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ*. [Online] 327 (7429), 1459–1461. Available from: doi:10.1136/bmj.327.7429.1459.

Snyders, K., Cho, D., Hong, J.H., Lord, S., et al. (2019) Benchmarking single-arm studies against historical controls from non-small cell lung cancer trials - an empirical analysis of bias. *Acta Oncologica (Stockholm, Sweden)*. [Online] 1–6. Available from: doi:10.1080/0284186X.2019.1674452.

Sormani, M.P. (2009) The Will Rogers phenomenon: the effect of different diagnostic criteria. *Journal of the Neurological Sciences*. [Online] 287 Suppl 1, S46-49. Available from: doi:10.1016/S0022-510X(09)71300-0.

Sperber, D., Mortimer, D., Lorgelly, P. & Berlowitz, D. (2013) An Expert on Every Street Corner? Methods for Eliciting Distributions in Geographically Dispersed Opinion Pools. *Value in Health*. [Online] 16 (2), 434–437. Available from: doi:10.1016/j.jval.2012.10.011.

Stanley, J.C. (1966) The Influence of Fisher's "The Design of Experiments" on Educational Research Thirty Years Later. *American Educational Research Journal*. [Online] 3 (3), 223–229. Available from: doi:10.3102/00028312003003223.

Sterne, J.A. & Smith, G.D. (2001) Sifting the evidence—what's wrong with significance tests? *Physical Therapy*. 81 (8), 1464–1469.

Strimenopoulou, F. & Walley, R. (2014) *Clinical and pre-clinical applications of Bayesian methods at UCB*.

Sugiyama, Y. & Seita, A. (2013) Kanehiro Takaki and the control of beriberi in the Japanese Navy. *Journal of the Royal Society of Medicine*. [Online] 106 (8), 332–334. Available from: doi:10.1177/0141076813497889.

Sureda, A., Arranz, R., Iriondo, A., Carreras, E., et al. (2001) Autologous stem-cell transplantation for Hodgkin's disease: results and prognostic factors in 494 patients from the Grupo Español de Linfomas/Transplante Autólogo de Médula Osea Spanish Cooperative Group. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*. [Online] 19 (5), 1395–1404. Available from: doi:10.1200/JCO.2001.19.5.1395.

Sureda, A., Canals, C., Arranz, R., Caballero, D., et al. (2012) Allogeneic stem cell transplantation after reduced intensity conditioning in patients with relapsed or refractory Hodgkin's lymphoma. Results of the HDR-ALLO study - a prospective clinical trial by the Grupo Espanol de Linfomas/Trasplante de Medula Osea (GEL/TAMO) and the Lymphoma Working Party of the European Group for Blood and Marrow Transplantation. *Haematologica*. [Online] 97 (2), 310–317. Available from: doi:10.3324/haematol.2011.045757.

Tafuri, G., Stolk, P., Trotta, F., Putzeist, M., et al. (2014) How do the EMA and FDA decide which anticancer drugs make it to the market? A comparative qualitative study on decision makers' views. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*. [Online] 25 (1), 265–269. Available from: doi:10.1093/annonc/mdt512.

Takaki, K. (1906) Three Lectures on the preservation of health amongst the personnel of the Japanese navy and army. *The Lancet*. [Online] 167 (4317), 1451–1455. Available from: doi:10.1016/S0140-6736(01)10951-7.

Tang, H., Foster, N.R., Grothey, A., Ansell, S.M., et al. (2010) Comparison of Error Rates in Single-Arm Versus Randomized Phase II Cancer Clinical Trials. *Journal of Clinical Oncology*. [Online] 28 (11), 1936–1941. Available from: doi:10.1200/JCO.2009.25.5489.

Tappenden, P., Jones, R., Paisley, S. & Carroll, C. (2006) *The use of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer*. [Online]. Available from: https://www.nice.org.uk/guidance/TA118/documents/colorectal-cancer-metastatic-bevacizumab-cetuximab-assessment-report2.

Temple, R. & Ellenberg, S.S. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Annals of Internal Medicine*. 133 (6), 455–463.

Thall, P.F. & Simon, R. (1990) Incorporating historical control data in planning phase II clinical trials. *Statistics in medicine*. 9 (3), 215–228.

Thistlethwaite, D.L. & Campbell, D.T. (1960) Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*. [Online] 51 (6), 309–317. Available from: doi:10.1037/h0044319.

Thompson, S.G. & Higgins, J.P.T. (2002) How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*. [Online] 21 (11), 1559–1573. Available from: doi:10.1002/sim.1187.

Trotta, F., Leufkens, H.G.M., Schellens, J.H.M., Laing, R., et al. (2011) Evaluation of Oncology Drugs at the European Medicines Agency and US Food and Drug Administration: When Differences Have an Impact on Clinical Practice. *Journal of Clinical Oncology*. [Online] 29 (16), 2266–2272. Available from: doi:10.1200/JCO.2010.34.1248.

Tugwell, P. & Knottnerus, J.A. (2015) Is the 'Evidence-Pyramid' now dead? *Journal of Clinical Epidemiology*. [Online] 68 (11), 1247–1250. Available from: doi:10.1016/j.jclinepi.2015.10.001.

Turner, R.M., Spiegelhalter, D.J., Smith, G. & Thompson, S.G. (2009) Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 172 (1), 21–47.

Van Nooten, F., Dewilde, S., Van Belle, S. & Marbaix, S. (2007) *Cost-effectiveness of sunitinib as second line treatment in patients with metastatic renal cancer in belgium*.

VanderWeele, T.J. & Ding, P. (2017) Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*. [Online] Available from: doi:10.7326/M16-2607.

Vickers, A.J., Ballen, V. & Scher, H.I. (2007) Setting the Bar in Phase II Trials: The Use of Historical Data for Determining "Go/No Go" Decision for Definitive Phase III Testing. *Clinical Cancer Research*. [Online] 13 (3), 972–976. Available from: doi:10.1158/1078-0432.CCR-06-0909.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., et al. (2014) Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*. [Online] 13 (1), 41–54. Available from: doi:10.1002/pst.1589.

Vreman, R.A., Bouvy, J.C., Bloem, L.T., Hövels, A.M., et al. (2019) Weighing of Evidence by Health Technology Assessment Bodies: Retrospective Study of Reimbursement Recommendations for Conditionally Approved Drugs. *Clinical Pharmacology & Therapeutics*. [Online] 105 (3), 684–691. Available from: doi:10.1002/cpt.1251.

Wade, G. (2010) *The Centralised Procedure*. In: February 2010 European Medicines Agency, London UK. p.

White Junod, S. (2015) *FDA and Clinical Drug Trials: A Short History*. [Online]. January 2015. Available from: http://www.fda.gov/AboutFDA/WhatWeDo/History/Overviews/ucm304485.htm [Accessed: 13 January 2015].

Wierda, W.G., Kipps, T.J., Mayer, J., Stilgenbauer, S., et al. (2010) Ofatumumab As Single-Agent CD20 Immunotherapy in Fludarabine-Refractory Chronic Lymphocytic Leukemia. *Journal of Clinical Oncology*. [Online] 28 (10), 1749–1755. Available from: doi:10.1200/JCO.2009.25.3187.

Wu, J. & Xiong, X. (2016) Survival trial design and monitoring using historical controls: J. Wu and X. Xiong. *Pharmaceutical Statistics*. [Online] Available from: doi:10.1002/pst.1756 [Accessed: 19 July 2016].

Zagar, A.J., Kadziola, Z., Lipkovich, I. & Faries, D.E. (2017) Evaluating different strategies for estimating treatment effects in observational studies. *Journal of Biopharmaceutical Statistics*. [Online] 0 (0), 1–19. Available from: doi:10.1080/10543406.2017.1289953.

Zeissler, M.-L., Li, V., Parmar, M.K.B. & Carroll, C.B. (2020) Is It Possible to Conduct a Multi-Arm Multi-Stage Platform Trial in Parkinson's Disease: Lessons Learned from Other Neurodegenerative Disorders and Cancer. *Journal of Parkinson's Disease*. [Online] Available from: doi:10.3233/JPD-191856.

Zia, M.I., Siu, L.L., Pond, G.R. & Chen, E.X. (2005) Comparison of Outcomes of Phase II Studies and Subsequent Randomized Control Studies Using Identical Chemotherapeutic Regimens. *Journal of Clinical Oncology*. [Online] 23 (28), 6982–6991. Available from: doi:10.1200/JCO.2005.06.679.

Zinzani, P.L., Bendandi, M., Stefoni, V., Albertini, P., et al. (2000) Value of gemcitabine treatment in heavily pretreated Hodgkin's disease patients. *Haematologica*. 85 (9), 926–929.

APPENDIX A EXCLUSION CRITERIA FOR LITERATURE SEARCH FOR DRUGS APPROVED USING UNCONTROLLED CLINICAL TRIALS

This literature search was performed to identify the treatments licensed on the basis of uncontrolled clinical studies in the EU and US since 1999; the point where economic evaluation began to play a formal role in the use of pharmaceuticals in the UK, with the establishment of NICE.

To identify drugs licensed on this basis, the European Medicines Agency (EMA) website, and the US Food and Drug Agency (FDA) website were searched, for all approvals since 1 January 1999 based on uncontrolled clinical studies, i.e. where all arms of a trial involve the investigational product. The search included both new drug approvals, and new indications granted, provided these were on the basis of uncontrolled studies only, with control arm data available in the patient population.

A number of exclusion criteria were applied to the search – these are listed below and applied in the order given.

- Generic drugs and biosimilar drugs – Generic versions of existing products are licensed at the point of patent expiry. However, they are licensed on the basis of similarity to the existing product, with efficacy shown by the trials that were completed by the original product. As such, they do not fit the criteria of being licensed on the basis of only single arm data. Biosimilar drugs are excluded for the same reason as these are copies of existing biological drugs, which base their efficacy data on that of the original biological product
- Diagnostic technologies and medical devices – Some diagnostic technologies (for example radioimaging) are required to be licensed by regulators. However, as they do not have a therapeutic effect, they do not fit the research question. The same rationale applies to medical devices, which may need to undergo the regulatory process as they include a licensed medicine (for example delivery systems) – as these generate evidence on the effectiveness of the delivery system, and not the efficacy of a drug treatment, they have been excluded from the literature search
- Vaccines – Newly licensed vaccines are based on known and well understood technologies, building on a base of existing evidence of their efficacy. For example, although influenza vaccines must be licensed each year, the underlying technology is the same, only the influenza strains of influenza included in the vaccine vary (which are specified by the World Health Organisation). Whilst there may be little or no direct trial evidence to demonstrate the efficacy of an individual vaccine, the efficacy of the

concept has been shown. As such, they have been excluded from this literature search

- Antimicrobial products – The standard of proof for antimicrobial treatments is different to that of new chemical entities – as stated by the FDA ‘Antimicrobial drugs differ from other classes of drugs in that they are intended to affect microbial, rather than patient physiology. For this reason, the technical reports of in vivo and in vitro effects on microorganisms are critical for establishing effectiveness’ (Center for Drug Evaluation and Research, 1997). The drug approval path for antimicrobials therefore often involves in vitro experiments, calculation of ‘break points’, along with trials in an initial indication to demonstrate efficacy. Additional indications may then be approved based on in vitro work coupled with case series for treating different strains of bacteria. Due to these different regulatory standards (due to a well understood pathophysiology for these products (Powell, 2000)), they have been excluded from the literature review
- Blood products and recombinant blood products – Where blood products require licensing from regulatory bodies for manufacturing standards, if identical to human blood products, the question of mechanism of action and proof of efficacy is not relevant. Consequently the regulatory question is different, in not requiring proof of concept / efficacy, but instead equivalence to the naturally occurring version, with the manufacture of the products also examined. For this reason they have been excluded from this literature search
- Fixed-dose combinations of existing pharmaceuticals – Whilst requiring regulatory approval for sale, fixed-dose combination products are combinations of already approved products whose efficacy has already been demonstrated. Unless additional claims are made of the combination product (for example an additional benefit of combining the products), these have been excluded from this literature review
- Treatments that were refused a marketing authorisation – These do not fit the criteria of having been licensed on the basis of uncontrolled data due to not receiving a license. Treatments that were licensed but subsequently withdrawn, however, are included in the review

All other treatments are included in this review.

APPENDIX B LITERATURE SEARCH FOR ECONOMIC EVALUATIONS OF DRUGS LICENSED ON THE BASIS OF UNCONTROLLED CLINICAL TRIALS

To identify economic models constructed based on uncontrolled clinical trial data, a number of databases were searched, which are listed below:

- Medline
- International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Scientific Presentations Database
- The National Institute for Health and Care Excellence (NICE), Scottish Medicines Consortium (SMC), and All Wales Medicines Strategy Group (AWMSG) health technology assessments

The search strategies taken for each database searched are discussed in turn

B1 MEDLINE (SEARCHED USING PUBMED)

Medline is the main database in the biomedical sphere, with all important research findings indexed.

To find economic models, the search strategy shown below was used for all 66 drugs found to be licensed based on uncontrolled study data.

Figure B-1: PubMed search strategy for economic models of drugs based on uncontrolled study data

1. *Generic drug name*
2. *Drug brand name EU*
3. *Drug brand name US*
4. Or 1-3

5. Cost-Benefit Analysis
6. Cost-utility
7. Cost-effectiveness
8. Pharmacoeconomic*
9. health economic*
10. cea
11. cua
12. markov*
13. “patient level simulation”
14. “discrete event simulation”
15. “monte carlo”
16. “decision tree”
17. “quality adjusted life”
18. qaly*

19. qald*
20. qale
21. “disability adjusted life”
22. hta
23. “health technology assessment”
24. or 5-23

25. 4 AND 24

Search strings 1-3 are to identify each individual drug, and were repeated for each treatment individually.

Search strings 5-23 are to identify economic evaluations, and are the same for all treatments.

Search 24 identifies economic models related to the drugs identified as being licensed on the basis of uncontrolled clinical studies.

Where a generic name included a chemical type (for example bosutinib monohydrate), the chemical type (in this case monohydrate) was omitted to give broader results. If a generic or trade name included a space, the term was included in quotation marks. Radiogardase (Prussian Blue) was also searched for as two separate terms using an OR term – Radiogardase OR “Prussian Blue”.

Due to the number of results for Zinc OR Wilzin, coupled with cost-effectiveness search terms (461 hits) the term “AND Wilson*” was added. Likewise for paclitaxel (361 hits), the term “AND Kaposi*” was added.

The resulting hits were then filtered by abstract review, to find papers describing economic evaluations in the relevant indication. Exclusion criteria were abstracts that were for other indications, solely modelling other drugs, clinical or scientific data, or using subsequently available randomised controlled trial data.

The full text for papers likely to describe models or modelling approaches were then retrieved and reviewed fully. The resulting dates of the searches, number of hits, and number of full-text papers retrieved is shown in Table B-1.

Table B-1: Results of PubMed searches for economic evaluations of drugs licensed on the basis of uncontrolled study data

Generic name	Date of search	Publication hits	Full publications retrieved	Relevant publications
Zinc	05/08/2014	1	0	0
Sodium Phenylbutyrate	08/08/2014	1	0	0

Anagrelide	04/08/2014	4	2	2
Paclitaxel	08/08/2014	1	0	0
Argatroban	04/08/2014	13	3	2
Sodium Ferric Gluconate Complex	08/08/2014	8	3	0
Busulfan	04/08/2014	11	0	0
Temozolomide	08/08/2014	46	1	0
Bexarotene	04/08/2014	1	0	0
Temoporfin	08/08/2014	3	2	2
Gemtuzumab Ozogamicin	05/08/2014	4	1	1
Alemtuzumab	04/08/2014	14	2	1
Nitisinone	08/08/2014	2	0	0
Arsenic Trioxide	04/08/2014	3	0	0
Tositumomab; Iodine I 131 Tositumomab	08/08/2014	14	0	0
Abarelix	04/08/2014	1	0	0
Imatinib Mesylate	05/08/2014	83	3	3
Imatinib Mesylate	05/08/2014	83	4	2
Cetuximab	04/08/2014	92	8	3
Trabectedin	08/08/2014	11	6	5
Cladribine	05/08/2014	7	1	1
Gefitinib	05/08/2014	55	1	0
Bortezomib	04/08/2014	43	1	1
Ferric Hexacyanoferrate(Ii)	05/08/2014	19	0	0
Clofarabine	05/08/2014	2	0	0
Pentetate Calcium Trisodium	08/08/2014	0	0	0
Pentetate Zinc Trisodium	08/08/2014	0	0	0
Nelarabine	05/08/2014	0	0	0
Betaine Anhydrous	04/08/2014	9	0	0
Dexrazoxane Hydrochloride	05/08/2014	8	3	0
Alglucosidase Alfa	04/08/2014	4	2	2
Sunitinib Malate	08/08/2014	61	4	2
Imatinib Mesylate	05/08/2014	83	0	0
Imatinib Mesylate	05/08/2014	83	0	0
Imatinib Mesylate	05/08/2014	83	0	0
Dasatinib	05/08/2014	23	0	0
Dasatinib	05/08/2014	23	0	0
Imatinib Mesylate	05/08/2014	83	0	0
Imatinib Mesylate	05/08/2014	83	0	0
Vorinostat	05/08/2014	6	0	0
Bortezomib	06/08/2014	45	0	0
Hydroxocobalamin	05/08/2014	9	0	0
Nilotinib Hydrochloride Monohydrate	08/08/2014	13	0	0
Ixabepilone	05/08/2014	6	3	0
Tocofersolan	08/08/2014	1	0	0

Bendamustine Hydrochloride	04/08/2014	6	1	0
Romidepsin	08/08/2014	0	0	0
Ofatumumab	08/08/2014	2	1	1
Pralatrexate	08/08/2014	1	0	0
Carglumic Acid	04/08/2014	1	0	0
Cholic Acid	05/08/2014	2	0	0
Omacetaxine Mepesuccinate	08/08/2014	0	0	0
Alipogene Tiparvovec	04/08/2014	0	0	0
Taliglucerase Alfa	08/08/2014	1	0	0
Asparaginase Erwinia Chrysanthemi	04/08/2014	9	2	1
Brentuximab Vedotin	04/08/2014	0	0	0
Brentuximab Vedotin	04/08/2014	0	0	0
Crizotinib	05/08/2014	7	2	0
Defibrotide	05/08/2014	3	0	0
Glucarpidase	05/08/2014	14	0	0
Carfilzomib	04/08/2014	1	0	0
Vismodegib	05/08/2014	0	0	0
Bosutinib	04/08/2014	1	0	0
Pasireotide Diaspartate	08/08/2014	2	0	0
Cholic Acid	05/08/2014	2	0	0
Lomitapide Mesylate	05/08/2014	1	0	0
Pomalidomide	08/08/2014	1	0	0
Raxibacumab	08/08/2014	1	0	0
Ponatinib Hydrochloride	08/08/2014	0	0	0
Ponatinib Hydrochloride	08/08/2014	0	0	0
Metreleptin	05/08/2014	0	0	0
Ibrutinib	05/08/2014	1	0	0
Ibrutinib	05/08/2014	1		
Ceritinib	04/08/2014	0	0	0

B2 INTERNATIONAL SOCIETY FOR PHARMACOECONOMICS AND OUTCOMES RESEARCH (ISPOR) SCIENTIFIC PRESENTATIONS DATABASE

ISPOR is the main organisation for economists working in economic evaluation. The calendar for the organisation revolves around two pivotal meetings – in late May, ISPOR International, a conference held in North America (2013 attendance: 2975), and in November, ISPOR Europe (2013 attendance: 3800) (ISPOR, 2020). As a part of the meetings, a large number of research posters and podiums are presented (in excess of 1000), which are not all subsequently published in peer reviewed journals. Due to the large number of pharmaceutical company personnel in attendance, and specialist audiences, this is also likely to be a venue for the presentation of models. To find any models that were identified, the ISPOR Scientific Presentations Database, a database of all accepted presentations was searched for each product identified in Appendix A, which were then filtered to identify relevant economic models. The results of these searches are shown in Table B-2

Table B-2: Results of searches for economic evaluations of drugs licensed on the basis of uncontrolled study data in the ISPOR Scientific Presentations Database

Generic name	Date of search	Hits	ISPOR hits	ISPOR Abstracts for review	Relevant ISPOR abstracts
Zinc	01/08/2014	0	0	0	0
Sodium Phenylbutyrate	01/08/2014	0	0	0	0
Anagrelide	28/07/2014	1	1	0	0
Paclitaxel	30/07/2014	141	158	0	0
Argatroban	28/07/2014	2	2	1	0
Sodium Ferric Gluconate Complex	01/08/2014	3	0	0	0
Busulfan	30/07/2014	4	4	1	1
Temozolomide	01/08/2014	24	24	0	0
Bexarotene	30/07/2014	1	1	0	0
Temoporfin	01/08/2014	0	0	0	0
Gemtuzumab Ozogamicin	30/07/2014	5	5	0	0
Alemtuzumab	24/07/2014	13	17	0	0
Nitisinone	30/07/2014	2	2	0	0
Arsenic Trioxide	28/07/2014	2	4	2	0
Tositumomab; Iodine I 131 Tositumomab	01/08/2014	1	1	0	0
Abarelix	24/07/2014	0	0	0	0
Imatinib Mesylate	30/07/2014	119	119	1	0
Imatinib Mesylate	30/07/2014	119	119	2	1
Cetuximab	30/07/2014	96	106	1	0

Trabectedin	01/08/2014	6	6	2	1
Cladribine	30/07/2014	2	2	1	0
Gefitinib	30/07/2014	42	48	2	1
Bortezomib	30/07/2014	62	70	0	0
Ferric Hexacyanoferrate(Ii)	30/07/2014	0	0	0	0
Clofarabine	30/07/2014	2	2	1	0
Pentetate Calcium Trisodium	01/08/2014	2	2	0	0
Pentetate Zinc Trisodium	01/08/2014	2	2	0	0
Nelarabine	30/07/2014	0	0	0	0
Betaine Anhydrous	28/07/2014	0	0	0	0
Dexrazoxane Hydrochloride	30/07/2014	0	0	0	0
Alglucosidase Alfa	28/07/2014	2	0	0	0
Sunitinib Malate	01/08/2014	111	113	5	4
Imatinib Mesylate	30/07/2014	119	119	0	0
Imatinib Mesylate	30/07/2014	119	119	0	0
Imatinib Mesylate	30/07/2014	119	119	1	0
Dasatinib	30/07/2014	62	68	6	1
Dasatinib	30/07/2014	62	68	0	0
Imatinib Mesylate	30/07/2014	119	119	0	0
Imatinib Mesylate	30/07/2014	119	119	0	0
Vorinostat	01/08/2014	0	0	0	0
Bortezomib	02/08/2014	1	70	3	1
Hydroxocobalamin	30/07/2014	0	0	0	0
Nilotinib Hydrochloride Monohydrate	30/07/2014	47	54	2	2
Ixabepilone	30/07/2014	4	4	1	0
Tocofersolan	01/08/2014	0	0	0	0
Bendamustine Hydrochloride	28/07/2014	24	29	1	0
Romidepsin	01/08/2014	0	0	0	0
Ofatumumab	30/07/2014	10	15	3	2
Pralatrexate	01/08/2014	0	0	0	0
Carglumic Acid	30/07/2014	1	1	0	0
Cholic Acid	30/07/2014	1	2	0	0
Omacetaxine Mepesuccinate	30/07/2014	1	1	1	0
Alipogene Tiparvovec	28/07/2014	0	0	0	0
Taliglucerase Alfa	01/08/2014	1	1	0	0
Asparaginase Erwinia Chrysanthemi	28/07/2014	1	1	0	0
Brentuximab Vedotin	30/07/2014	5	7	1	1
Brentuximab Vedotin	30/07/2014	3	7	1	0
Crizotinib	30/07/2014	13	19	2	0
Defibrotide	30/07/2014	0	0	0	0
Glucarpidase	30/07/2014	0	0	0	0
Carfilzomib	30/07/2014	1	1	0	0
Vismodegib	01/08/2014	0	1	0	0
Bosutinib	30/07/2014	5	8	1	1
Pasireotide Diaspartate	30/07/2014	3	3	0	0

Cholic Acid	30/07/2014	1	2	0	0
Lomitapide Mesylate	30/07/2014	0	0	0	0
Pomalidomide	01/08/2014	1	3	0	0
Raxibacumab	01/08/2014	0	0	0	0
Ponatinib Hydrochloride	01/08/2014	4	6	1	0
Ponatinib Hydrochloride	01/08/2014	4	6	0	0
Metreleptin	30/07/2014	0	0	0	0
Ibrutinib	30/07/2014	0	0	0	0
Ibrutinib	30/07/2014	0	0	0	0
Ceritinib	30/07/2014	0	0	0	0

B3 NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE), SCOTTISH MEDICINES CONSORTIUM (SMC) AND ALL WALES MEDICINES STRATEGY GROUP (AWMSG) HEALTH TECHNOLOGY APPRAISALS

To identify economic evaluations, the website of each organisation was searched for each of the 66 drugs identified and indication approved on the basis of uncontrolled study data. The results were then filtered for health technology assessments in the relevant indication, with the relevant appraisals downloaded for further review. The results of these searches are shown in Table B-3.

Table B-3: Results of searches for health technology assessments of drugs licensed on the basis of uncontrolled study data

Generic name	Date of searches	NICE Appraisal	SMC Appraisal	AWMSG Appraisal
Zinc	01/08/2014	-	-	-
Sodium Phenylbutyrate	01/08/2014	-	-	Reference 280
Anagrelide	28/07/2014	-	SMC ID 163/05‡	Reference 23
Paclitaxel	30/07/2014	-	-	-
Argatroban	28/07/2014	-	SMC ID 812/12‡	Reference 1405
Sodium Ferric Gluconate Complex	01/08/2014	-	-	-
Busulfan	30/07/2014	-	SMC ID 337/06	-
Temozolomide	01/08/2014	TA23	-	-
Bexarotene	30/07/2014	-	SMC ID 14/02	-
Temoporfin	01/08/2014	-	SMC ID 96/04	-
Gemtuzumab Ozogamicin	30/07/2014	-	-	-
Alemtuzumab	24/07/2014	-	SMC ID 494/08†	Advice 2208
Nitisinone	30/07/2014	-	-	-
Arsenic Trioxide	28/07/2014	-	-	-
Tositumomab; Iodine I 131 Tositumomab	01/08/2014	-	-	-
Abarelix	24/07/2014	-	-	Reference 354*
Imatinib Mesylate	30/07/2014	TA50, TA70, TA251	SMC ID 01/02, SMC ID 46/02	-
Imatinib Mesylate	30/07/2014	TA86, TA209	SMD ID 08/02, SMC ID 584/09§	Reference 1653
Cetuximab	30/07/2014	TA118, TA150*, TA242	SMC ID 155/05, SMC ID 543/09	Reference 81, Reference 400
Trabectedin	01/08/2014	TA185	SMC ID 454/08¶	Reference 318
Cladribine	30/07/2014	-	Reference 537/09	-
Gefitinib	30/07/2014	-	-	-
Bortezomib	30/07/2014	TA129	SMC ID 126/04	Reference 65
Ferric Hexacyanoferrate(Ii)	30/07/2014	-	-	-
Clofarabine	30/07/2014	-	Reference 327/06	Reference 92
Pentetate Calcium Trisodium	01/08/2014	-	-	-
Pentetate Zinc Trisodium	01/08/2014	-	-	-

Nelarabine	30/07/2014	-	SMC ID 454/08	Reference 216
Betaine Anhydrous	28/07/2014	-	SMC ID 407/07§	-
Dexrazoxane Hydrochloride	30/07/2014	-	-	-
Alglucosidase Alfa	28/07/2014	-	SMC ID 352/07	Reference 17*
Sunitinib Malate	01/08/2014	TA169	SMC ID 343/07‡	Reference 294
Imatinib Mesylate	30/07/2014	-	SMC ID 428/07*	-
Imatinib Mesylate	30/07/2014	-	SMC ID 430/07*	-
Imatinib Mesylate	30/07/2014	-	SMC ID 426/07*, SMC ID 427/07*, SMC ID 923/13*	Reference 2014*
Dasatinib	30/07/2014	TA241	SMC ID 370/07	Reference 103
Dasatinib	30/07/2014	-	SMC ID 371/07	Reference 102
Imatinib Mesylate	30/07/2014	-	-	-
Imatinib Mesylate	30/07/2014	-	SMC ID 429/07*	-
Vorinostat	01/08/2014	-	-	-
Bortezomib	02/08/2014	-	-	-
Hydroxocobalamin	30/07/2014	-	-	-
Nilotinib Hydrochloride Monohydrate	30/07/2014	TA241, TA251	SMC ID 440/08	-
Ixabepilone	30/07/2014	-	-	-
Tocofersolan	01/08/2014	-	SMC ID 696/11	Reference 1180*
Bendamustine Hydrochloride	28/07/2014	TA206*	SMC ID 701/11*	Reference 39*
Romidepsin	01/08/2014	-	-	-
Ofatumumab	30/07/2014	TA202	SMC IS 626/10	-
Pralatrexate	01/08/2014	-	-	-
Carglumic Acid	30/07/2014	-	SMC ID 899/13	Reference 2371*
Cholic Acid	30/07/2014	-	-	Reference 929*
Omacetaxine Mepesuccinate	30/07/2014	-	-	-
Alipogene Tiparvovec	28/07/2014	-	-	Reference 645*
Taliglucerase Alfa	01/08/2014	-	-	-
Asparaginase Erwinia Chrysanthemi	28/07/2014	-	-	-
Brentuximab Vedotin	30/07/2014	-	SMC ID 845/12*	Reference 1255*
Brentuximab Vedotin	30/07/2014	-	SMC ID 845/12*	Reference 1255*
Crizotinib	30/07/2014	TA296	SMC ID 865/13‡	-
Defibrotide	30/07/2014	-	SMC ID 867/14	-
Glucarpidase	30/07/2014	-	-	-
Carfilzomib	30/07/2014	-	-	-
Vismodegib	01/08/2014	-	SMC ID 924/13*	Reference 1037*
Bosutinib	30/07/2014	TA299	SMC ID 910/13	-
Pasireotide Diaspartate	30/07/2014	-	SMC ID 815/12*	Reference 642*
Cholic Acid	30/07/2014	-	-	-
Lomitapide Mesylate	30/07/2014	-	SMC ID 956/14*	Reference 1182*
Pomalidomide	01/08/2014	-	SMC ID 972/14	-
Raxibacumab	01/08/2014	-	-	-
Ponatinib Hydrochloride	01/08/2014	-	-	-
Ponatinib Hydrochloride	01/08/2014	-	-	-
Metreleptin	30/07/2014	-	-	-
Ibrutinib	30/07/2014	-	-	-

Ibrutinib	30/07/2014	-	-	-
Ceritinib	30/07/2014	-	-	-

* No submission received from manufacturer, † Unable to retrieve due to product withdrawal, ‡ Two submissions available under the same reference number, § Three submissions available under the same reference number, ¶ Four submissions available under the same reference number

APPENDIX C CODE USED FOR THE IMPLEMENTATION OF MAIC IN THE SIMULATION STUDY

Should others wish to implement the approach to MAIC tested in the simulation study presented in Chapter 4, the code snippet below demonstrates the implementation of MAIC on both first and higher moments. Also included (commented out for faster runtime) are cross-checks that can be performed on the weighting performed – should the code be used for running an individual analysis, it is advised these be presented.

```
#set up the simulation
    # print(paste0("Scenario ", i, " in scenarios ",
simulation.scenarios.start, ":", simulation.scenarios.end, ", Simulation:
", j, "/", simulation.n, ". Last scenario: ", time.scenario))
    set.seed(simulation.seed.table[i,j]) #set the seed for the simulation

    #set up effect of coefficients for individuals
    run.coefficient.effect <-
matrix(coefficient.effect(simulation.theoretical*simulation.C), nrow =
simulation.theoretical, ncol = simulation.C)

    #simulate Population A
    ##set up covariates
    run.intervention.X <- matrix(NA, nrow = simulation.theoretical, ncol =
simulation.C)
    run.intervention.X[,1:simulation.J] <-
popa.betas.J(simulation.theoretical*simulation.J)
    if (simulation.C>simulation.J) {
      run.intervention.X[,(simulation.J+1):simulation.C] <-
popa.betas.C((simulation.C-simulation.J)*simulation.theoretical)
    }
    ##calculate outcomes
    run.intervention.Y <- matrix(NA, nrow = simulation.theoretical, ncol =
2) # 2 columns - 1. untreated outcomes, and 2. treated outcomes
    run.intervention.Y[,1] <- outcome.notreat(simulation.theoretical,
run.intervention.X)
    run.intervention.Y[,2] <- outcome.treat(simulation.theoretical,
run.intervention.X)

    #simulate popb
    ##set up covariates
    run.control.X <- matrix(NA, nrow = simulation.theoretical, ncol =
simulation.C)
    run.control.X[,1:simulation.J] <-
popb.betas.J(simulation.J*simulation.theoretical)
    if (simulation.C > simulation.J) {
        run.control.X[,(simulation.J+1):simulation.C] <-
popb.betas.C((simulation.C-simulation.J)*simulation.theoretical)
    }
    ##calculate outcomes
    run.control.Y <- matrix(NA, nrow = simulation.theoretical, ncol = 2) #
2 columns - 1. untreated outcomes, and 2. treated outcomes
    run.control.Y[,1] <- outcome.notreat(simulation.theoretical,
run.control.X)
    run.control.Y[,2] <- outcome.treat(simulation.theoretical,
run.control.X)

    #sample from popa and popb to create analysis sets
```

```
    run.popa.sample <- sample(1:simulation.theoretical, run.popa.n, replace
= FALSE)
    run.popa.X <- run.intervention.X[run.popa.sample,]
    run.popa.Y <- run.intervention.Y[run.popa.sample,]
    run.popb.sample <- sample(1:simulation.theoretical, run.popb.n, replace
= FALSE)
    run.popb.X <- run.control.X[run.popb.sample,]
    run.popb.Y <- run.control.Y[run.popb.sample,]

    #Save individual mean survival times for future diagnostics
    simulation.diagnostics[i,j,1,1] <- mean(run.control.Y[,1])
    simulation.diagnostics[i,j,1,2] <- mean(run.popb.Y[,1])
    simulation.diagnostics[i,j,1,3] <- mean(run.control.Y[,2])
    simulation.diagnostics[i,j,1,4] <- mean(run.popb.Y[,2])
    simulation.diagnostics[i,j,1,5] <- mean(run.intervention.Y[,2])
    simulation.diagnostics[i,j,1,6] <- mean(run.popa.Y[,2])

    #True difference i.e. from non-sampled data
    run.sampled.true.surv <- Surv(c(run.popb.Y[,1], run.popb.Y[,2]),
c(rep(1, run.popb.n*2)))
    run.sampled.true.cox <- coxph(formula = run.sampled.true.surv ~
c(rep(0, run.popb.n), rep(1, run.popb.n)), robust = TRUE)
    # run.sampled.true.cox2 <- coxph(formula = run.sampled.true.surv ~
c(rep(0, run.popb.n), rep(1, run.popb.n)),
cluster(as.factor(1:nrow(run.sampled.true.surv)))) #exploring clustering
    simulation.result[i,j,1,1] <- exp(run.sampled.true.cox$coefficients)

    #Naive difference i.e. from sampled data, cross comparison
    run.sampled.naive.surv <- Surv(c(run.popb.Y[,1], run.popa.Y[,2]),
c(rep(1, run.popb.n+run.popa.n)))
    run.sampled.naive.cox <- coxph(formula = run.sampled.naive.surv ~
c(rep(0, run.popb.n), rep(1, run.popa.n)), robust = TRUE)
    simulation.result[i,j,1,2] <- exp(run.sampled.naive.cox$coefficients)
    simulation.result[i,j,1,3:4] <- exp(confint(run.sampled.naive.cox))
    ##SE; i,j,*2* is the key
    simulation.result[i,j,2,2] <- sqrt(diag(run.sampled.naive.cox$var))
#summary(run.sampled.naive.cox)$coefficients[,3]

    #Perform MAIC
    if (simulation.options.maic == "Yes") {
      ##create summary data of the popb
      # colMeans(run.popa.X[,1:simulation.J]) #summary of popa if wanted
for comparison
      run.popb.X.summaries <- matrix(NA, nrow = 2, ncol = simulation.J)
      for (k in 1:simulation.J) {
        run.popb.X.summaries[1,k] <- mean(run.popb.X[,k]) #row 1 - mean
        run.popb.X.summaries[2,k] <- sd(run.popb.X[,k]) #row 2 - sd
      }
      ##recentre popa data for matching
      run.maic.zerod <- run.popa.X[,1:simulation.J]
      run.maic.zerod <- sweep(run.maic.zerod, 2, run.popb.X.summaries[1,],
"-")
      # colMeans(run.maic.zerod) #summary of zerod data if wanted for
comparison
      ##set up starting weights - assume 1 unless we know better
      run.maic.startvalues <- rep(1, simulation.J)
      ##set up function to be minimised
      maic.minimise <- function(theta) {
        sum(exp(as.matrix(run.maic.zerod) %*% theta))
      }
      #find optimum
```

```
      run.maic.result <- optim(fn = maic.minimise, par =
run.maic.startvalues, method = "BFGS")
      #extract betas
      run.maic.betas <- run.maic.result$par
      #calculate resulting weights
      run.maic.weights <- exp(as.matrix(run.maic.zerod) %*% run.maic.betas)
      run.maic.weights <- ifelse(run.maic.weights < simulation.weight.min-
6, simulation.weight.min-6, run.maic.weights)
      #check resulting matching
      ##effective sample size
      # run.maic.ess <- (sum(run.maic.weights)^2) / sum(run.maic.weights^2)
      ##weight with original data should be mean of the popb for each
variable
      # run.maic.confirmation <- matrix(NA, nrow = 3, ncol = simulation.J)
      # for (k in 1:simulation.J) {
      #   run.maic.confirmation[1,k] <- wt.mean(run.popa.X[,k],
run.maic.weights) #row 1 - mean
      #   run.maic.confirmation[2,k] <- wt.sd(run.popa.X[,k],
run.maic.weights) #row 2 - sd
      #   run.maic.confirmation[3,k] <- wt.mean(run.maic.zerod[,k],
run.maic.weights) #row 3 - zerod point estimate (to see errors)
      # }
      ##present results for checking:
      ###weighted with original data - should be close
      # colMeans(run.popa.X[,1:simulation.J]) # popa means
      # run.popb.X.summaries # popb
      # run.maic.confirmation[1:2,]# popa weighted to match popb
      ###weighted with zerod data, should be 0 (or close to it)
      # run.maic.confirmation[3,]# zerod popa weighted to match popb
      ##calculate resulting HR and save outputs
      run.sampled.maic.cox <- coxph(formula = run.sampled.naive.surv ~
c(rep(0, run.popb.n), rep(1, run.popa.n)), robust = TRUE, weights =
c(rep(1, run.popb.n), run.maic.weights))
      simulation.result[i,j,1,5] <- exp(run.sampled.maic.cox$coefficients)
      simulation.result[i,j,1,6:7] <- exp(confint(run.sampled.maic.cox))
      ##SE; i,j,*2* is the key
      simulation.result[i,j,2,5] <- sqrt(diag(run.sampled.maic.cox$var))
#summary(run.sampled.naive.cox)$coefficients[,3]

      #Perform MAIC with higher moments
      ## from Signorovitch's paper:
      #"For example, given the baseline mean and standard deviation of age,
it is straightforward to compute the mean of squared age, which can then be
treated as a separate mean baseline characteristic for matching."
      ##create summary data of the popb
      # colMeans(run.popa.X[,1:simulation.J]) #summary of the popa if
wanted for comparison
      run.popb.X.summaries <- matrix(NA, nrow = 2, ncol = simulation.J)
      for (k in 1:simulation.J) {
        run.popb.X.summaries[1,k] <- mean(run.popb.X[,k]) #row 1 - mean
        run.popb.X.summaries[2,k] <- sd(run.popb.X[,k]) #row 2 - sd
      }
      ##calculate mean squared values from popb summaries
      run.popb.X.meansquared <- matrix(NA, nrow = 1, ncol = simulation.J)
      for (k in 1:simulation.J) {
        run.popb.X.meansquared[1,k] <- mean(rnorm(n=1000, mean =
run.popb.X.summaries[1,k], sd = run.popb.X.summaries[2,k])^2)
      }
      ##recentre popa data for matching - including squared values
      run.maichm.values <- cbind(run.popa.X[,1:simulation.J],
run.popa.X[,1:simulation.J]^2)
```

```
      run.maichm.zerod <- sweep(run.maichm.values, 2,
c(run.popb.X.summaries[1,], run.popb.X.meansquared[1,]), "-")
      # colMeans(run.maichm.zerod) #summary of zerod data if wanted for
comparison
      ##set up starting weights - assume 1 unless we know better
      run.maichm.startvalues <- rep(1, (simulation.J*2))
      ##set up function to be minimised
      maichm.minimise <- function(theta) {
        sum(exp(as.matrix(run.maichm.zerod) %*% theta))
      }
      #find optimum
      run.maichm.result <- optim(fn = maichm.minimise, par =
run.maichm.startvalues, method = "BFGS")
      #extract betas
      run.maichm.betas <- run.maichm.result$par
      #calculate resulting weights
      run.maichm.weights <- exp(as.matrix(run.maichm.zerod) %*%
run.maichm.betas)
      run.maichm.weights <- ifelse(run.maichm.weights <
simulation.weight.min, simulation.weight.min, run.maic.weights) #in case
weight is too low, set a min
      #check resulting matching
      ##effective sample size
      # run.maichm.ess <- (sum(run.maichm.weights)^2) /
sum(run.maichm.weights^2)
      ##weight with original data should be mean of the popb for each
variable
      # run.maichm.confirmation <- matrix(NA, nrow = 3, ncol =
simulation.J)
      # for (k in 1:simulation.J) {
      #   run.maichm.confirmation[1,k] <- wt.mean(run.popa.X[,k],
run.maichm.weights) #row 1 - mean
      #   run.maichm.confirmation[2,k] <- wt.sd(run.popa.X[,k],
run.maichm.weights) #row 2 - sd
      #   run.maichm.confirmation[3,k] <- wt.mean(run.maichm.zerod[,k],
run.maichm.weights) #row 3 - zerod data to see errors
      # }
      ##present results for checking:
      ###weighted with original data - should be close
      # colMeans(run.popa.X[,1:simulation.J]) # popa
      # run.popb.X.summaries # popb
      # run.maichm.confirmation[1:2,]# popa weighted to match popb
      ###weighted with zerod data, should be 0 (or close to it)
      # run.maichm.confirmation[3,]# zerod popa weighted to match popb
      ##calculate resulting survival and save outputs
      run.sampled.maichm.cox <- coxph(formula = run.sampled.naive.surv ~
c(rep(0, run.popb.n), rep(1, run.popa.n)), robust = TRUE, weights =
c(rep(1, run.popb.n), run.maichm.weights))
      simulation.result[i,j,1,8] <-
exp(run.sampled.maichm.cox$coefficients)
      simulation.result[i,j,1,9:10] <- exp(confint(run.sampled.maichm.cox))
      ##SE; i,j,*2* is the key
      simulation.result[i,j,2,8] <- sqrt(diag(run.sampled.maichm.cox$var))
#summary(run.sampled.naive.cox)$coefficients[,3]
    }
```

APPENDIX D EXAMPLE R CODE TO IMPLEMENT METHODS CONCEPTUALISED IN SECTION 5.2

The R code presented was written in R version 3.6.1, and uses simulated data to demonstrate the approaches proposed.

```
tic.project <- Sys.time()

# Details ------------------------------------------------------------

# Creation of controls using
# Approach 1. Difference between previous line and current line of TTP
# Approach 2. Difference between PFS and OS from previous line publication

# Install & load packages --------------------------------------------

# if needed install package install.load
#install.packages("install.load")
library("install.load")
# Packages to be installed
packages <- c(
  "MASS", #bivarate normal sampling
  "ggplot2", #plotting density
  "survminer", #survival plotting
  "flexsurv", #survival regression
  "survival" #used for Surv objects
)
install_load(packages)
rm(packages)

# Settings -----------------------------------------------------------

n.patients <- 250 #number of patients to simulate in each trial
n.sims <- 50000 #number of simulations for sampling from
rich6equal = c("#000043", "#0033FF", "#01CCA4", "#BAFF12", "#FFCC00", "#FF3300") #r Colours

#plot heights and widths
graphheight <- 6
graphwidth <- 10

#set seed for replicable results
set.seed(1337)

# Approach 1. Difference between previous line and current line ----------

# Step 1: Data for previous and current line.
#for ease, here we simulate data using Weibull distributions
approach1.previous.line <- rweibull(n = n.patients, shape = 1.2, scale = 9.5)
approach1.current.line <- rweibull(n = n.patients, shape = 1.2, scale = 10)

#create survival objects
approach1.previous.line.surv <- Surv(approach1.previous.line, rep(1, n.patients))
approach1.previous.line.survfit <- survfit(approach1.previous.line.surv ~ 1)
approach1.current.line.surv <- Surv(approach1.current.line, rep(1, n.patients))
approach1.current.line.survfit <- survfit(approach1.current.line.surv ~ 1)

#plot the resulting survival data
ggsurvplot(fit = approach1.previous.line.survfit, data = approach1.previous.line.surv,
           xlab = "Time",
           ylab = "Survival",
           risk.table = TRUE,
           conf.int = TRUE,
           conf.int.style = "step")

ggsurvplot(fit = approach1.current.line.survfit, data = approach1.current.line.surv,
           xlab = "Time",
           ylab = "Survival",
           risk.table = TRUE,
           conf.int = TRUE,
           conf.int.style = "step")

# Step 2: Fit survival curves to the data
```

```
approach1.previous.line.weibull <- flexsurvreg(approach1.previous.line.surv ~ 1,
dist="weibull")
approach1.current.line.weibull <- flexsurvreg(approach1.current.line.surv ~ 1, dist="weibull")

#Step 3: Take samples from the data
#samples
approach1.previous.line.samples <- mvrnorm(n = n.sims, mu =
approach1.previous.line.weibull$coefficients, Sigma = approach1.previous.line.weibull$cov)
approach1.current.line.samples <- mvrnorm(n = n.sims, mu =
approach1.current.line.weibull$coefficients, Sigma = approach1.current.line.weibull$cov)

#exponentiate as needed
approach1.previous.line.samples <- exp(approach1.previous.line.samples)
approach1.current.line.samples <- exp(approach1.current.line.samples)

approach1.previous.line.predicted <- matrix(NA, 1, n.sims)
approach1.current.line.predicted <- matrix(NA, 1, n.sims)
for (i in 1:n.sims) {
  approach1.previous.line.predicted[i] <- mean(rweibull(100, shape =
approach1.previous.line.samples[i,1], scale = approach1.previous.line.samples[i,2]))
  approach1.current.line.predicted[i] <- mean(rweibull(100,
approach1.current.line.samples[i,1], approach1.current.line.samples[i,2]))
}

#Step 4: Calculate which offers better survival
#estimate benefit by taking one from the other
approach1.netbenefit <- as.vector(approach1.current.line.predicted) -
as.vector(approach1.previous.line.predicted)

##Likely benefit, accounting for uncertainty
summary(as.vector(approach1.previous.line.predicted))
summary(as.vector(approach1.current.line.predicted))
summary(approach1.netbenefit)

#plot the density of the difference in survival
plot(density(approach1.netbenefit), col = rich6equal)

#What percent are above zero i.e. how often is the intervention superior?
100*length(as.vector(approach1.netbenefit[approach1.netbenefit>0]))/n.sims



# Approach 2. Difference between PFS and OS from previous line --------

# Step 1: Import data from the external study for OS and PFS, as well as the OS from the
current study.
#Usually this would be digitsed, but to keep the example self contained, we have used
simulated data
approach2.external.PFS <- rweibull(n = n.patients, shape = 1.1, scale = 9.5)
approach2.external.PPS <- rweibull(n = n.patients, shape = 1.1, scale = 4)
approach2.external.OS <- approach2.external.PFS + approach2.external.PPS

#create a data frames of all data
approach2.external.PFS.dataframe <- data.frame(Time = approach2.external.PFS, Event = rep(1,
n.patients), Data = "PFS")
approach2.external.OS.dataframe <- data.frame(Time = approach2.external.PFS, Event = rep(1,
n.patients), Data = "OS")
approach2.external.dataframe <- rbind(approach2.external.PFS.dataframe,
approach2.external.OS.dataframe)

# Step 2: Fit survival curves
#We suggest the approach of Latimer et al. featured as a NICE DSU report, and also in MDM,
though in this case we assume this has been performed, and a weibull is used
approach2.external.PFS.surv <- Surv(approach2.external.PFS, rep(1, n.patients))
approach2.external.PFS.fit <- flexsurvreg(approach2.external.PFS.surv ~ 1, dist="weibull")
approach2.external.OS.surv <- Surv(approach2.external.OS, rep(1, n.patients))
approach2.external.OS.fit <- flexsurvreg(approach2.external.OS.surv ~ 1, dist="weibull")

#Step 3: Simulate data from the fitted curves
#samples
approach2.external.PFS.samples <- mvrnorm(n = n.sims, mu =
approach2.external.PFS.fit$coefficients, Sigma = approach2.external.PFS.fit$cov)
approach2.external.OS.samples <- mvrnorm(n = n.sims, mu =
approach2.external.OS.fit$coefficients, Sigma = approach2.external.OS.fit$cov)

#exponentiate as needed
approach2.external.PFS.samples <- exp(approach2.external.PFS.samples)
```

```
approach2.external.OS.samples <- exp(approach2.external.OS.samples)

#calculate predicted values
approach2.external.PFS.predicted <- matrix(NA, 1, n.sims)
approach2.external.OS.predicted <- matrix(NA, 1, n.sims)
for (i in 1:n.sims) {
  approach2.external.PFS.predicted[i] <- mean(rweibull(100, shape =
approach2.external.PFS.samples[i,1], scale = approach2.external.PFS.samples[i,2]))
  approach2.external.OS.predicted[i] <- mean(rweibull(100, approach2.external.OS.samples[i,1],
approach2.external.OS.samples[i,2]))
}

#Step 4: calculate expected outcomes of the post progression survival
#subtract PFS from OS, calculate summary statistics of the likely PPS time, and plot hisograms
and density plots
approach2.external.PPS.predicted <- as.vector(approach2.external.OS.predicted) -
as.vector(approach2.external.PFS.predicted)
summary(approach2.external.PPS.predicted)
hist(approach2.external.PPS.predicted)
plot(density(approach2.external.PPS.predicted))

#Step 5: Import survival data and fit a curve to the contemporary study
#Here again we use simulated data for ease, again assuming a weibull distribution is known
approach2.contemporary.OS <- rweibull(n = n.patients, shape = 1.1, scale = 7)
approach2.contemporary.OS.surv <- Surv(approach2.contemporary.OS, rep(1, n.patients))
approach2.contemporary.OS.fit <- flexsurvreg(approach2.contemporary.OS.surv ~ 1,
dist="weibull")

#Step 6: Sample from the fitted contemporary survival fitting, and estmate predicted OS for
comparison (to account for uncertainty)
#samples
approach2.contemporary.OS.samples <- mvrnorm(n = n.sims, mu =
approach2.contemporary.OS.fit$coefficients, Sigma = approach2.contemporary.OS.fit$cov)

#exponentiate as needed
approach2.contemporary.OS.samples <- exp(approach2.contemporary.OS.samples)

#calculate predicted values
approach2.contemporary.OS.predicted <- matrix(NA, 1, n.sims)
for (i in 1:n.sims) {
  approach2.contemporary.OS.predicted[i] <- mean(rweibull(100, shape =
approach2.contemporary.OS.samples[i,1], scale = approach2.contemporary.OS.samples[i,2]))
}

#Step 6: Calculate and plot the expected net survival gain of treatment
approach2.netbenefit <- as.vector(approach2.contemporary.OS.predicted) -
as.vector(approach2.external.PPS.predicted)

##Likely benefit, accounting for uncertainty
summary(as.vector(approach2.external.PPS.predicted)) #historical data
summary(as.vector(approach2.contemporary.OS.predicted)) #contemporary data
summary(approach2.netbenefit) #net benefit

#plot the density of the difference in survival
plot(density(approach2.netbenefit), col = rich6equal)

#What percent are above zero i.e. how often is the intervention superior?
100*length(as.vector(approach2.netbenefit[approach2.netbenefit>0]))/n.sims


toc.project <- Sys.time()
timetorun.project <- toc.project - tic.project
timetorun.project
```