Economic evaluation of factorial randomised controlled trials

DPhil in Public Health

Hilary term 2015

Helen Dakin

St Cross College, Oxford

Economic evaluation of factorial randomised controlled trials

Helen Dakin, St Cross College: DPhil in Public Health, Hilary term 2015

Abstract

Factorial randomised controlled trials (RCTs) evaluate two or more interventions simultaneously, enabling assessment of interactions between treatments. This thesis presents literature reviews, methodological reviews, simulation studies and applied case studies that explore methods for assessing cost-effectiveness based on factorial RCTs.

My systematic review suggests that factorial RCTs account for around 3% of trial-based economic evaluations, although there is currently no guidance or methodological work indicating the most appropriate methods. Around 40% of published studies assumed no interaction between treatments and many were poorly-reported.

Various mechanisms are likely to produce large interactions within economic endpoints such as costs, quality-adjusted life-years (QALYs) and net benefits. Failing to take account of interactions can introduce bias and prevent efficient allocation of healthcare resources. I developed the opportunity cost of ignoring interactions as a measure of the implications of this bias. However, allowing for small, chance interactions is inefficient, potentially leading to over-investment in research if trial-based evaluations are used to inform decisions about subsequent research. Nonetheless, analyses on simulated trial data suggest that the opportunity cost of adopting a treatment that will not maximise health gains from the

healthcare budget is minimised by including all interactions regardless of magnitude or statistical significance.

Different approaches for conducting economic evaluations of factorial RCTs (including regression techniques, extrapolation using patient-level simulation, and considering different components of net benefit separately) are evaluated within three applied studies, including both full and partial factorials with 2x2 and 2x2x2 designs.

I demonstrate that within both trial-based and model-based economic evaluation, efficient allocation of healthcare resources requires consideration of interactions between treatments, and joint decisions about interacting treatments based on incremental cost-effectiveness evaluated “inside-the-table” on a natural scale.

I make recommendations for the design, analysis and reporting of factorial trial-based economic evaluations based on the results of this thesis.

Contents

Acknowledgements	vi
Publications and presentations from the thesis	viii
Glossary	x
List of abbreviations	xiii
Mathematical notation	xv
Chapter 1: Introduction	1
Chapter 2: Conceptual review on issues and methods for analysing factorial trials	8
Chapter 3: Issues raised by economic evaluation of factorial trials	43
Chapter 4: Systematic review of economic evaluations of factorial trials	99
Chapter 5: Analysis of simulated data	116
Chapter 6: IVAN – A full 2x2 factorial trial	132
Chapter 7: KAT – A partial factorial trial	158
Chapter 8: AFORRD – A full 2x2x2 factorial trial	179
Chapter 9: Decision-making for healthcare resource allocation: Joint vs. separate decisions on interacting interventions	200
Chapter 10: Conclusions and recommendations	226
Appendix A: Protocol for conducting the conceptual literature review on issues and methods for statistical analysis of factorial trials in general	240
Appendix B: List of references identified in the conceptual literature review of factorial trial literature	243
Appendix C: Types of factorial trial	249
Appendix D: Examples of different types of interaction	252
Appendix E: Proof that joint decisions inside-the-table give efficient treatment-adoption decisions	270
Appendix F: Protocol for the systematic review of economic evaluations of factorial trials	275

Appendix G: Data extraction table for the systematic review of economic evaluations of factorial trials 281

Appendix H: Data inputs, additional results and Stata code for the simulation analyses described in Chapter 5 339

Appendix I: Additional results and methods for the IVAN economic evaluation 360

Appendix J: Results of the inside-the-table subgroup analysis on the KAT trial 393

Appendix K: Further details on the AFORRD economic evaluation 397

References for all chapters and appendices 406

Acknowledgements

This thesis is my own work, but would not have been completed without input from others. In particular, I would like to thank my supervisor, Alastair Gray, for detailed and insightful comments on all chapters.

I received helpful feedback on material from Chapters 2, 3, 6, 7 and 9 when these were presented at Health Economists' Study Group meetings. In particular, I would like to thank Steve Morris, Richard Grieve, Karla Diaz-Ordaz and Mark Strong for their insightful discussant presentations. I also received helpful feedback on Chapters 6 and 9 from members of the Birmingham Health Economics Unit and the Oxford Health Economics Research Centre, respectively, following presentations I gave. Sarah Wordsworth, Giselle Abangma, Chris Rogers and Barney Reeves gave helpful comments on an early draft of Chapter 6, while Graeme MacLennan, Richard Morris and David Murray did the same for a draft of Chapter 7.

I would also like to thank the participants and investigators of the AFORRD, IVAN and KAT trials for their work in designing the trials and collecting and cleaning data. Rosa Legood conducted preliminary work on the costing and analysis of AFORRD, which I updated and refined; the work conducted by both Rosa and myself was supervised by Alastair Gray. I also received help with the UKPDS Outcomes Model from José Leal. Sarah Wordsworth, James Raftery, Giselle Abangma, Rosa Legood, Chris Rogers & Barney Reeves were involved in planning the IVAN economic analysis. Giselle Abangma assisted me with the IVAN costing analyses and Sarah Wordsworth supervised the work of both myself and Giselle. I conducted the KAT economic evaluation under Alastair Gray's supervision.

My tuition fees were funded by the Health Economics Research Centre. Pfizer funded the AFORRD trial and economic evaluation. The IVAN (project number 07/36/01) and KAT (project number 95/10/01) trials and their economic evaluations were funded by the National Institute for health research (NIHR) Health Technology Assessment (HTA) Programme (www.hta.ac.uk) and will be published in full in *Health Technology Assessment*. The views and opinions expressed are mine alone and do not reflect those of the NIHR, Pfizer or the Department of Health. I would like to thank Abacus International for allowing me to use the HTAInSite within Chapter 9 and the developers of the UKPDS Outcomes Model, which was used in Chapter 8.

I would also like to thank my husband Rob Fellows for his support, suggestions (most notably Figure 9.1) and help with proofreading, formatting and mathematical notation.

However, any remaining mistakes are my own.

Publications and conference abstract/papers from the thesis

Publication	Relation to thesis
Dakin H, Gray A. Economic evaluation of factorial randomised controlled trials: Why the method of analysis matters. Podium presentation at the International Health Economics Association (iHEA) 8th World Congress, Toronto, Canada, 10th-13th July 2011 2011.	Introduce some of the ideas and worked examples presented in Chapters 2-4
Dakin HA, Gray A. Economic evaluation of factorial randomised controlled trials: Why the method of analysis matters Presented at the Health Economists' Study Group (HESG) meeting 23-25 June 2010, Cork, Ireland. 2010.	
Dakin H, Wordsworth S, Gray A, Rogers C, Abangma G, Reeves R. Why consider interactions in trial-based economic evaluation? A case study of a factorial trial. Presented at HESG meeting June 2012, Oxford 2012.	Early version of Chapter 6
Chakravarthy U, Harding SP, Rogers CA, Downes S, Lotery AJ, Dakin H, Culliford L, Scott L, Nash RL, Taylor J, Muldrew A, Sahni J, Wordsworth S, Raftery J, Peto T, Mitchell J, Bradley C, Reeves BC for the IVAN Investigators. A randomised controlled trial to assess the effectiveness and cost-effectiveness of alternative treatments to Inhibit VEGF in Age-related choroidal Neovascularisation (IVAN). Health Technology Assessment 2015 (in press).	Describe economic results of the IVAN trial discussed in Chapter 6
Dakin HA, Wordsworth S, Rogers CA, Abangma G, Raftery J, Harding SP, Lotery AJ, Downes SM, Chakravarthy U. Cost-effectiveness of ranibizumab and bevacizumab for age-related macular degeneration: 2-year findings from the IVAN randomised trial. BMJ Open 2014; 4(7):e005094. Shown in Appendix I.	
Rogers C, Dakin H, Wordsworth S, Abangma G, Raftery J, Harding H, Chakravarthy U, Lotery A, Downes S and Reeves B, on behalf of the IVAN Study Investigators. Cost-effectiveness of ranibizumab and bevacizumab for neovascular age-related macular degeneration: One year IVAN results. Presented at the Association for Research in Vision and Ophthalmology (ARVO) conference, May 2013.	
The IVAN Study Investigators, Chakravarthy U, Harding SP, Rogers CA, Downes SM, Lotery AJ, Wordsworth S, Reeves BC. Ranibizumab versus Bevacizumab to Treat Neovascular Age-related Macular Degeneration: One-Year Findings from the IVAN Randomized Trial. Ophthalmology 2012;119(7):1399-1411.	

Publication	Relation to thesis
IVAN study investigators. Alternative treatments to Inhibit VEGF in Age-related choroidal Neovascularization (IVAN): 1 Year Results. Presented at the ARVO conference, May 2012.	
Dakin HA, Gray AM, MacLennan GS, Morris RW, Murray DW. Independence, interactions and inference in partial factorial trials. Presented at HESG meeting, Exeter, January 2013.	Early version of Chapter 7
Murray DW, MacLennan GS, Breeman S, Dakin HA, Johnston L, Campbell MK, Gray AM, Fiddian N, Fitzpatrick R, Morris RW, Grant AM on behalf of the KAT group. (2014) A randomised controlled trial of the clinical effectiveness and cost-effectiveness of different knee prostheses: the Knee Arthroplasty Trial (KAT). Health Technology Assessment, 18 (19).	Describes economic results of the KAT trial discussed in Chapter 7
Breeman S, Campbell MK, Dakin H, Fiddian N, Fitzpatrick R, Grant A, Gray A, Johnston L, Maclennan GS, Morris RW, & Murray DW (2013). Five-year results of a randomised controlled trial comparing mobile and fixed bearings in total knee replacement. Bone Joint J, 95-B(4), 486-492.	
Breeman S, Campbell C, Dakin H, Fiddian N, Fitzpatrick R, Grant A, Gray A, Johnston L, MacLennan G, Morris R and Murray D on behalf of the KAT Trial Group. Patellar resurfacing in total knee replacement: five year clinical and economic results of a large randomized controlled trial. The Journal of Bone and Joint Surgery (American) 2011; 93:1473-1481.	
Dakin H. 2014 Decision-making for healthcare resource allocation: Joint versus separate decisions on interacting interventions. Presented at the HESG meeting, Sheffield, January 2014.	Early version of Chapter 9

Glossary

Additive [effects]	Where there is no ***interaction*** on a natural (i.e. non-transformed) scale: i.e. $\mu_0 + \mu_{ab} - \mu_a - \mu_b = 0$.[1] This means that the effect of A and B in combination is exactly equal to the sum of their individual effects and the effect of A is the same regardless of whether B is also given: i.e. $\mu_a - \mu_0 = \mu_{ab} - \mu_b$. The ***interaction:effect ratio*** equals 0.
At-the-margins [analysis]	A way of analysing factorial trials in which ***cells*** are pooled together in order to compare outcomes for all patients receiving A with all those not receiving A, thereby ignoring all interactions. The term at-the-margins appears to have been coined by McAlister et al (1). ***Main effects*** are estimated by taking the difference between all of the patients who received A (across cells *a* and *ab*) and all those who did not receive A (across cells *b* and *0*). Results are presented for all patients with each ***level*** of each ***factor***, pooling across cells (sometimes referred to as in the factorial manner (2)).
Cell	A unique combination of factors or treatments evaluated in a ***factorial trial***. For example, within a 2x2 ***factorial trial***, patients will be randomised to one of four cells or treatment-combinations: *0*, *a*, *b* and *ab*.
Factor	One of the treatments or comparisons to which patients in a factorial trial are randomised. For example, a 2x2 factorial trial will include two factors, which might include the presence/absence of drug A and the presence/absence of drug B, or different dosing regimens.
Factorial [randomised controlled] trial (RCT)	Randomised trial in which two or more factors (each with two or more levels) are deliberately varied at the same time, with patients being randomly allocated to receive different combinations of levels for each factor in such a way that interactions between two or more factors can be evaluated.
Full factorial	Factorial design in which all patients are randomised to all factors.
Higher-order interaction	An interaction between more than two factors.
Inside-the-table [analysis]	A way of analysing factorial trials in which each cell (or combination of factors) is considered as a separate treatment. Outcomes for each cell in the factorial design are calculated by taking the average across patients randomised to that factor-combination. ***Simple effects*** will be estimated by calculating the difference or ratio between the outcomes for individual cells: e.g. the effect of A is estimated by comparing outcomes in patients who receive A monotherapy (cell *a*) with those who received no treatment (cell *0*). This type of analysis takes account of interactions. This term appears to have been coined by McAlister et al (1).

[1] Mathematical notation used in the thesis is defined on pages xvi-xvii.

Interaction	The extent to which the effect of A differs depending on whether or not B is given. On a natural scale, the interaction equals $(\mu_{ab} - \mu_b) - (\mu_a - \mu_0) = \mu_0 - \mu_a - \mu_b + \mu_{ab}$: i.e. the difference between the effect of using A and B in combination and the sum of the individual effects.
Interaction:effect ratio	A measure of the size of interactions (relative to between-group differences) that demonstrates whether the interaction is ***super-additive***, ***sub-additive***/***mixed*** or ***qualitative***. When all treatments have the same direction of effect (e.g. when both A and B both increase cost, or both decrease cost), the interaction:effect ratio is calculated by dividing the interaction term by whichever of the simple effects has the smaller absolute magnitude. Dividing the interaction by the *smaller* of the two simple effects ensures that qualitative interactions (those changing the ranking of treatments) have interaction:effect ratios below -1. For "mixed" situations, where the simple effects have different signs (e.g. A increases cost, while B decreases cost), I ensure that qualitative interactions had interaction:effect ratios below -1 by dividing the interaction by the simple effect that has the opposite sign to the interaction term. I coined this term and developed this measure within Section 4.2.
Level	The treatments that are compared within any single factor. For any given factor, patients will be randomised to one of ≥2 levels. For example, Factor A may compare: active drug A and placebo of A (two levels); drug A, drug B and drug C (three levels); or placebo of A and 10, 20 and 40 mg drug A (four levels).
Main effect	Within the literature and among statisticians, this term is used in two different contexts. Some papers (e.g. (3-5)) state that within the context of a factorial trial, the main effect strictly equals the average impact of A, averaged across patients with and without B ($0.5(\mu_{ab} + \mu_a - \mu_b - \mu_0)$): i.e. the treatment effect estimated using at-the-margins analysis. However, the term main effect is more commonly used to refer to any estimate of the effect of a single factor, regardless of whether interactions terms are taken into account. For example, within a regression model $y_i = \beta_o + \beta_A A_i + \beta_B B_i + \beta_{AB} A_i B_i + \varepsilon_i$, the term main effect is often used to describe β_A and β_B (6). Within this thesis, I use the term in both contexts: i.e. to differentiate between "simple" (inside-the-table) and "main" (at-the-margins) effects and to differentiate between interaction terms and the effect of individual factors.
Mixed interaction	An interaction that has the same sign as one treatment effect, but the opposite sign from the other. For mixed interactions, the interaction:effect ratio will be between -1 and 0. I coined this term in Section 2.5.2.
Multiplicative [effect]	Where there is no interaction on a logarithmic scale: i.e. $\ln(\mu_0) + \ln(\mu_{ab}) - \ln(\mu_a) - \ln(\mu_b) = 0$. This means that the *relative* effect (e.g. odds ratio) of using A and B in combination is exactly equal to the *product* of their individual effects (e.g. odds ratios) and that the relative or proportional effect of A is the same regardless of whether B is given: i.e. $\mu_a \div \mu_0 = \mu_{ab} \div \mu_b$.

Non-qualitative interaction	Any interaction that is neither equal to 0 nor qualitative: i.e. where the effect of one treatment varies in magnitude (but not sign) depending on whether the other treatment is given. Non-qualitative interactions will have interaction:effect ratios >-1 that are not equal to 0.
Opportunity cost of ignoring interactions	The net benefit lost from assuming additive effects and making separate decisions on the two factors, rather than taking account of interactions and making a joint decision on both factors simultaneously. This may be assessed under current information (i.e. based on the expected net benefits across the study sample), or under perfect information (i.e. based on the net benefits that would be achieved if all uncertainty was eliminated). I coined this term for the measure I developed in Section 3.8.
Partial factorial randomised controlled trial	Trial evaluating multiple treatments simultaneously on the same patient group by randomising a subset of patients to >1 comparison in a factorial manner, while other patients are randomised to just one comparison or to a different combination of factors.
Qualitative interaction	Where the effect of at least one of the treatments under investigation changes *sign* (not just magnitude) depending on whether or not the other therapy is given (7, 8). Qualitative interactions change the ranking of treatments with respect to that outcome and have interaction:effect ratios <-1.
Simple effect	The impact of adding A to no treatment (i.e. $\mu_a - \mu_0$) (4, 5). This represents the treatment effect estimated using inside-the-table analysis and the coefficient on A (β_A) that is estimated in regression analyses that include a full set of interaction terms.
Sub-additive interaction	Where the effect of the combination is *less* than the sum of the parts and the factors have antagonistic effects (1, 8); the interaction term therefore has the *opposite sign* from both the simple effects and has an interaction:effect ratio between -1 and 0.
Super-additive interaction	Where the effect of the combination is *greater* than the sum of the parts and there is synergy between the factors (1, 7, 8). Given super-additive interactions, the interaction term will have the *same sign* as the simple effects and the interaction:effect ratio will be >0.

List of abbreviations

AE	Adverse event
AFORRD	Atorvastatin in Factorial with Omega-3 fatty acids Risk Reduction in Diabetes [trial]
AIC	Akaike information criterion
ANOVA	Analysis of variance
BIC	Bayesian information criterion
BPH	Benign prostatic hypertrophy
CBA	Cost-benefit analysis
CCA	Cost-consequence analysis
CDR	Common Drug Review
CEA	Cost-effectiveness analysis
CEAC	Cost-effectiveness acceptability curve
CI	Confidence interval
CMA	Cost-minimisation analysis
CONSORT	Consolidated Standards of Reporting Trials [statement]
CRD	Centre for Reviews and Dissemination
CUA	Cost-utility analysis
DIC	Deviance information criterion
EQ-5D	EuroQoL five-dimensional quality of life questionnaire (three-level version unless otherwise specified)
EThOS	(British Library) Electronic Theses Online Service
EVPI	Expected value of perfect information
EVPPI	Expected value of partial perfect information
EVSI	Expected value of sample information
GCEA	Generalised cost-effectiveness analysis
GLM	Generalised linear model
GP	General practice
HESG	Health Economists' Study Group
HRQoL	Health-related quality of life
HTA	Health technology assessment
HUI	Health Utilities Index

ICER	Incremental cost-effectiveness ratio
INB	Incremental net [monetary] benefit
IPW	Inverse probability weighting
ITT	Intention-to-treat
IVAN	Inhibit Vascular endothelial growth factor in Age-related choroidal Neovascularisation [trial]
KAT	Knee Arthroplasty Trial
KMSA	Kaplan-Meier sample averaging
LDL	Low-density lipoprotein
MCMC	Markov chain Monte Carlo
MeSH	Medical Subject Heading
MTA	Multiple technology appraisal
NHB	Net health benefit
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NMA	Network meta-analysis
NMB	Net monetary benefit. For consistency, NMB is used to refer to net benefits in general.
OLS	Ordinary least squares
PBAC	Pharmaceutical Benefits Advisory Committee
QALY	Quality-adjusted life-year
RCT	Randomised controlled trial
SAE	Serious adverse event
SD	Standard deviation
SE	Standard error
SMC	Scottish Medicines Consortium
STA	Single technology appraisal
UK	United Kingdom
UKPDS	United Kingdom Prospective Diabetes Study
US	United States
WHO	World Health Organisation

Mathematical notation

Factorial trials

I describe the size of the factorial matrix as the product of the number of levels in one factor multiplied by the number of levels in another factor. For example, trials with two factors, each with two levels are described as 2x2 (or 2^2), while those with one three-level factor and two two-level factors are described as 3x2x2.

Factors and dummy variables used to represent factors within regression analysis are referred to with capital letters: e.g. A, B. For example, within the regression model $y_i = \beta_o + \beta_A A_i + \beta_B B_i + \beta_{AB} A_i B_i + \varepsilon_i$, A_i would equal 1 for those patients randomised to receive active treatment A and 0 for those patients randomised to receive A's comparator (e.g. placebo of A). Regression coefficients showing the effect of factor X are denoted using the factor abbreviation as a subscript: e.g. β_A.

		Factor A: presence/absence of Drug A	
		Level 0: Placebo for A	**Level 1: Active drug A**
Factor B: presence/ absence of Drug B	**Level 0: Placebo for B**	**Placebo for A + Placebo for B** Treatment label: *00* or *0* Mean (SD) outcome = μ_0 (σ_0) Sample size = N_0	**Drug A + Placebo for B** Treatment label: *a0* or *a* Mean (SD) outcome = μ_a (σ_a) Sample size = N_a
	Level 1: Active drug B	**Placebo for A + Drug B** Treatment label: *0b* or *b* Mean (SD) outcome: μ_b (σ_b) Sample size = N_b	**Drug A + Drug B** Treatment label: *ab* Mean (SD) outcome = μ_{ab} (σ_{ab}) Sample size = N_{ab}

Abbreviations: SD, standard deviation.

Combinations of upper case letters are used to represent interactions: e.g. "AB" represents the interaction between factors A and B, while "ABC" represents the three-factor interaction between A, B and C. Regression coefficients showing the interaction effect are denoted using the abbreviation for the interaction in subscript: e.g. β_{AB}.

For designs with >2 levels of one or more factors, italicised lower case letters (e.g. $a_2b_0c_3d_3$, (9, 10)) are typically used to describe the levels of each factor used in each cell or treatment allocation. However, for brevity, I follow the convention (9, 11) for full factorial trials with two levels per factor of using italicised lower-case letters without subscripts (e.g. "*a*") to indicate level 1 of a factor and indicate level 0 as the absence of any symbol for that factor. The cells in a 2x2 design are therefore designated *0*, *a*, *b* and *ab.*

The mean outcomes for each treatment are denoted μ_0, μ_a, μ_b and μ_{ab}; standard deviations within each cell are denoted as σ_0, σ_a, σ_b and σ_{ab}.

Economic evaluation

Rc is used to refer to the ceiling ratio.

The costs and QALYs accrued in cell *ab* are referred to as $Cost_{ab}$ and $QALYs_{ab}$.

For consistency, I use net monetary benefit (NMB) throughout this thesis to quantify cost-effectiveness and to refer to net benefits in general, where $NMB = QALYs \bullet Rc - Cost$. However, I could, equivalently, have used net health benefit (NHB) throughout, which would have given identical results and conclusions: the only difference would have been expression of net benefits in health units (e.g. QALYs) rather than monetary terms.

Chapter 1: Introduction

1.1 Introduction

There is growing interest in economic evaluation of healthcare technologies to ensure that finite healthcare funds achieve the maximum possible benefit for the population. Such economic evaluations may be based on decision-analytical models using secondary data (12-14). However, increasing numbers of evaluations are conducted alongside randomised controlled trials (RCTs) (15-17).

There is an extensive literature on the appropriate methods for trial-based economic evaluation (18, 19), although there is currently little guidance or research on how such methods should be applied to different types of trial. In particular, factorial trials, which account for around 1-5% of RCTs (1, 6, 20-22), raise several additional issues, such as ***interactions*** and comparisons between multiple treatments.

Factorial studies evaluate ≥2 ***factors*** simultaneously. The factors under investigation may comprise the presence or absence of a drug or other healthcare technology (Table 1.1), ≥2 active treatments, or any other aspect of dose or treatment administration for which outcomes can be usefully compared. Each factor will have ≥2 ***levels***. For example, Factor A may compare: active drug A and placebo of A (two levels, Table 1.1); drugs A, B and C (three levels); or placebo against 10, 20 and 40 mg drug A (four levels). ***Factorial RCTs*** randomise patients to ≥2 levels of ≥2 factors, with different groups of subjects being randomly allocated to receive different combinations of levels for each factor. For example, the Atorvastatin in Factorial with Omega-3 fatty acids Risk Reduction in Diabetes (AFORRD) trial evaluated atorvastatin, omega-3, the combination of the two and neither in patients with diabetes (23). Although most clinical factorial trials compare two levels of each of two factors – a 2x2 (or 2^2) design (24) – any number of levels and factors are possible.

Table 1.1. Illustration of a 2x2 factorial trial evaluating the effect of two drugs (A and B).

		Factor A: presence/absence of Drug A	
		Level 0: Placebo for A	**Level 1: Active drug A**
Factor B: presence/ absence of Drug B	**Level 0: Placebo for B**	*0*: Placebo for A + Placebo for B	*a*: Drug A + Placebo for B
	Level 1: Active drug B	*b*: Placebo A + Drug B	*ab*: Drug A + Drug B

Although factorial designs were used in the 19th century (25, 26) (at least as far back as 1843 (26)), such experiments often only used simple comparisons and did not necessarily use random allocation to blocks (26). Fisher was the first to outline the advantages of factorial designs in 1926 (27), although the term "factorial" appears to have been coined in 1935 (26, 28). Factorial designs have since been applied in diverse fields, which include factorial RCTs of medical interventions as early as 1955 (29). Since then, factorial trials have continued to play a small but important role in medical research.

Factorial trials enable multiple questions to be addressed in the same study (1, 27): for example, we can assess the effectiveness/safety of A and B individually and the effect of A+B combination therapy. In particular, factorial trials allow us to investigate interactions between treatments (6, 9, 11, 24, 28, 30-34). Interactions indicate that the effect of A differs depending on whether B is given and mean that the effect of using A and B in combination is not equal to the sum of the individual effects. Interactions have implications for clinical practice: particularly for decisions about whether or not combination therapy is appropriate or which of the available treatment strategies should be adopted. Interactions may arise due to non-compliance (30), the scale of measurement (2, 6, 10, 30, 31, 35, 36) or pharmacokinetic, biological or behavioural mechanisms (2, 20, 30). The interaction term equals the observed outcomes in ***cell*** *ab* (Table 1.1) minus what we would expect if treatment effects were ***additive***. For example, if A monotherapy increased health-related quality of life (HRQoL) by 0.10 on a utility scale and B increased utility by 0.20, there would be zero interaction (i.e. perfectly additive effects) if utility was 0.30 higher in cell *ab* than cell *0*. Alternatively, there could be a ***super-additive***, or synergistic, interaction if utility in cell *ab* was >0.30 higher than

cell *0* (1, 7, 8) and a ***sub-additive***, or antagonistic, interaction if utility were <0.30 higher (1, 8). Interactions are described as ***qualitative*** if the effect of either A or B changes *sign* (not just magnitude) depending on whether the other treatment is given (7, 8): e.g. if utility were 0.15 higher for *ab* than *0* (since adding A alongside B decreases utility, whereas adding A to placebo increases it).

If there are no interactions, factorial designs provide greater power for the same sample size than three-arm studies evaluating the same treatments and (in theory) could provide the same information as two two-arm studies, but at substantially lower cost (20, 30, 37-39). However, as discussed in Chapter 2, the extent to which we can realise these benefits and obtain consistent, unbiased and statistically efficient estimates of treatment effects depends on how results are analysed and whether or not treatments interact.

The main statistical approaches for analysing clinical endpoints from factorial trials are:

- Assume zero interaction: ***At-the-margins*** analysis (comparing outcomes for all patients receiving A with all those not receiving A) and regression without interaction term(s) are generally statistically efficient (1, 30, 40-42), but implicitly assume that there are no interactions; such methods therefore produce biased estimates whenever the true interaction is not zero (3, 4, 20, 30).[2]
- Assume important interactions exist: ***Inside-the-table*** analysis[3] (considering each combination of factors as a separate treatment) and regression analyses including interaction term(s) almost always have lower statistical power than at-the-margins analysis (3) but allow for interactions and are therefore always unbiased (3, 4).

[2] Bias is a systematic tendency to over- or underestimate the parameter of interest. In this case, analyses excluding interactions systematically overestimate the effect of A compared with not-A in the absence of B by a biased equal to 50% of the true interaction term (3).

[3] Within this thesis, I use McAlister's term "inside-the-table" (1) to describe the method of analysing a factorial trial in which each cell or factor-combination is analysed as a separate study arm, with no pooling, but in which regression analysis is not used. This term has been used by several recent factorial trials citing that review (e.g. (43, 44)), although the earlier literature tends to refer only to analyses of this type as those estimating "simple effects" (3-5, 45) or conducting comparisons between cell means without reference to a specific term (5).

Although factorial designs can be used to efficiently evaluate multiple factors (9, 11, 26, 28, 30, 36, 46, 47) or investigate interactions (9, 11, 24, 28, 30-33), it is generally not possible to achieve both objectives simultaneously (33) since interactions reduce the efficiency gains possible with factorial trials (1, 8, 30, 40-42) and bias the most efficient estimators (3, 4, 8, 20, 30), while large samples are needed to detect interactions (26, 31, 33, 35, 37, 47).

As I show in Chapter 3, factorial designs raise additional challenges for economic evaluation over and above those associated with analysis of clinical endpoints. In particular, interactions are likely to be larger and more common for the endpoints used in economic evaluation (e.g. costs and quality-adjusted life-years [QALYs]) than for clinical endpoints. Regression analysis on economic endpoints also raises statistical challenges (17, 48-53) and transformation of costs and QALYs is problematic (17, 49, 54, 55). Additionally, the objectives and framework for economic evaluation differ from those of clinical endpoints: requiring quantification of uncertainty as the probability and opportunity cost of making the wrong decision; focusing on estimation rather than hypothesis testing; and employing different decision rules for independent or mutually-exclusive options. As result, it may be more important to minimise bias in economic evaluation than maximise statistical efficiency; furthermore, interactions may affect how decisions about treatment adoption are framed, as well as the form of analysis.

1.2 Research question

I aim to research the challenges raised by analysing economic data from factorial trials, evaluate the alternative analytical methods and make recommendations for how economic evaluations on such trials should be conducted and reported. Although factorial RCTs are well established in medicine and at least 40 factorial RCTs with trial-based economic

evaluations have now been published (Chapter 4), my literature reviews have identified no previous studies evaluating the methods for conducting economic evaluations of factorial trials (Chapter 3).

I plan to improve the methods for economic evaluations based on factorial trials, develop a general framework for identifying and dealing with interactions between interventions in economic evaluation and gain insights into the design and analysis of factorial trials in general: particularly those with more complex designs and data structures, such as partial factorial trials. Specifically, I plan to address the following sub-questions:

- How are clinical endpoints analysed in factorial trials?
- What challenges are raised by economic evaluation of factorial trials and what are the possible solutions?
- What mechanisms can produce interactions and how commonly do different types of interaction arise?
- What is the opportunity cost of ignoring interactions?
- How often are economic evaluations of factorial trials conducted and how are they analysed?
- How should we decide which interactions should be taken into account in factorial trial-based economic evaluations?
- How do the analytical methods affect the results of economic evaluations based on factorial RCTs?
- How should economic evaluations based on factorial trials be conducted and reported?
- What impact do interactions between interventions have in model-based economic evaluation and in health technology assessment (HTA)?
- What are the implications of my findings for policy and HTA?

I address these questions and implement the methods by reviewing the literature (Chapters 2-4), analysing simulated data (Chapters 3-5) and conducting trial-based economic evaluations of three factorial trials (Chapters 6-8): the Inhibit Vascular endothelial growth factor in Age-related choroidal Neovascularisation (IVAN) trial; the Knee Arthroplasty Trial (KAT); and AFORRD (Table 1.2). The case studies raise additional methodological issues not present within my simulated data, including missing data, censoring, imbalance in baseline characteristics and extrapolation.

Table 1.2. Case studies presented in the thesis

Study	Design	Analyses conducted within this thesis
IVAN (56, 57)	• UK multi-centre non-inferiority full factorial RCT. • Compared Avastin (bevacizumab) vs. Lucentis (ranibizumab) and continuous monthly dosing vs. as-needed treatment in patients with neovascular age-related macular degeneration.	• Within-trial analysis, including multiple imputation and adjustment for censoring and imbalance in baseline utility. • Net benefit was broken down into different components (i.e. QALYs and 3 types of cost), with different models used for different components. • Base case analysis included interactions that were statistically significant or larger than simple effects. • Results compared against: excluding all interactions; allowing for multiplicative effects on drug cost and excluding all other interactions; and including all interactions. • Also conducted analyses with all possible combinations of interactions.
KAT (58-60)	• UK multi-centre partial factorial trial. • Evaluated three different types of prosthesis design for knee replacement.	• Within-trial analysis, including multiple imputation and adjustment for censoring and imbalance in baseline utility. • Compared at-the-margins analysis against analysing those patients randomised to >1 comparison inside-the-table.
AFORRD (23)	• UK multi-centre RCT with 2x2x2 full factorial design. • Patients with diabetes were individually randomised to atorvastatin vs. placebo and to omega-3 vs. placebo and cluster-randomised by general practice to receive a leaflet designed to improve compliance vs. no leaflet.	• Extrapolated trial results using a patient simulation model. • Evaluated interactions for within-trial costs, clinical endpoints, numbers of diabetic events and extrapolated costs and QALYs. • Compared 3 approaches: allowing for all interactions in 2x2x2 design; ignoring leaflet and allowing for atorvastatin*omega-3 interactions; and at-the-margins analysis. • Evaluated the impact of adjusting for between-group imbalance in baseline characteristics.

Within the high-profile IVAN trial we would expect, a priori, to observe a very large, super-additive interaction on drug costs between drug and dosing regimen, since drug cost equals the cost per dose multiplied by the number of doses received. Whereas AFORRD and IVAN

use ***full factorial*** designs, KAT comprised a ***partial factorial trial***, in which only a subset of patients are randomised to >1 comparison (59, 61, 62), enabling interactions to be investigated, but only in a smaller population. KAT therefore enabled me to develop and evaluate analytical methods for partial factorial trials. Within AFORRD I explore methods for extrapolating factorial trials and compare analytical methods in a situation where two factors had minimal effect, making interactions unlikely (4, 30, 35). AFORRD also evaluates a more "typical" set of factors than IVAN or KAT, comparing the presence vs. absence of one pharmacological intervention and the presence vs. absence of non-pharmacological treatments with a less mature evidence base.

Following a review of the literature on factorial trials in general (Chapter 2), I combine the principles of economic evaluation and analysis of factorial trials to discuss the challenges posed by economic evaluation of factorial trials, exploring potential methods using a worked example (Chapter 3). Chapter 4 presents a systematic review of published factorial trial-based economic evaluations to identify the prevalence of interactions and evaluate how factorial trials were analysed before the current programme of work. I then present the results of simulation work, which attempts to identify the best decision rule for determining which interactions should be taken into account in economic evaluation (Chapter 5). Chapters 6-8 apply the methods to the three empirical case studies (IVAN, KAT and AFORRD). Chapter 9 extends the thesis beyond trial-based economic evaluation and discusses the implications of interactions for model-based economic evaluation, HTA and healthcare decision-making. I then develop guidelines for future economic evaluations based on factorial trials and discuss the contributions, strengths and weaknesses of the thesis, as well as the advantages and disadvantages of factorial trials as vehicles for economic evaluation (Chapter 10).

Chapter 2: Conceptual review on issues and methods for analysing factorial trials

2.1 Introduction

This chapter presents the methods and results of a conceptual review of the literature on factorial trials in general, which was conducted to identify the general principles that can be adapted to develop methods for economic evaluation of factorial trials.

After Section 2.2 outlines the main methods and results of my literature review, Section 2.3 reviews published definitions of the term “factorial trial”, which are used to develop the definition used in the remainder of this thesis. Section 2.4 reviews the advantages and disadvantages of factorial trials and the circumstances in which they are particularly useful. Section 2.5 describes the terminology and notation used throughout the thesis to classify interactions and describe how factorial trials are analysed. Section 2.6 describes the main methods used to analyse such trials with reference to the literature and a worked example, before I finish with a brief discussion of best practice guidelines (Section 2.7).

2.2 Methods and results of literature review

2.2.1 Methods

Following initial MEDLINE searches on 14th January 2010, EMBASE and MEDLINE (including Ovid MEDLINE®, Ovid MEDLINE® Daily Update and Ovid OLDMEDLINE®) were searched using Ovid on 29th January 2010. Search terms included “factorial” in combination with the terms "guideline*", "systematic review", "randomised controlled trial" or "review" (Appendix A). No synonyms of the term “factorial” were used in the searches as the term “factorial” has been in widespread use since the 1930s and the review was not intended to

be completely exhaustive. However, for the most useful articles, I reviewed bibliographies and references that cited them to identify other references that would inform the aims of this review.

Guidelines, systematic reviews, textbooks, review articles, methods papers and other publications were included in the review if they met one of the following aims:

A) Develop a clear, unambiguous definition of what comprises a “factorial RCT” and identify any synonyms for use in later literature searches.

B) Identify specific types or variants of factorial studies that have distinct characteristics and review the methods recommended for the design and/or statistical analysis of these study types.

C) Identify the rationale for conducting factorial trials and the advantages, disadvantages and challenges of this study design.

D) Identify the main approaches used to analyse data from factorial trials, the advantages and disadvantages of each approach and any guidelines on how factorial trials should be analysed or reported.

E) Identify trends on how and when factorial RCTs are commonly used.

Since this thesis focuses on economic evaluation of factorial trials, this review focused on factorial RCTs in healthcare, but also included key references on factorial designs in general (e.g. seminal publications by Fisher or Yates). The review also focuses on general principles and is not intended to identify every paper discussing use of factorial trials, still less to identify every applied example or regression model.

2.2.2 Results

Electronic searches yielded 554 hits (including duplicates), of which 82 publications (summarised in Appendix B) were considered relevant to the aims outlined in Section 2.2.1. Sixty-four of these (78%) comprised journal articles, while the remaining 18 were textbooks or book chapters. Six of the books covered experimental methods in general, while 12 specifically focussed on medical statistics or trial methodology. Nineteen papers were review articles, seven were guidelines and 34 were methodology papers describing primary research. A further four papers described systematic reviews of factorial trials, which each identified between 10 and 44 factorial studies.

2.3 Definition

Published definitions of the term “factorial trial” were reviewed to facilitate development of a clear, unambiguous definition that differentiated factorial trials from other study designs and could be used for subsequent literature reviews and to clearly define the scope of this thesis.

I identified 21 published definitions, which all emphasised that factorial trials evaluate at least two factors simultaneously (Table 2.1). However, although there appears to be general agreement about which studies are factorial and which are not, definitions varied in other aspects. Ten definitions also stated that the ability to detect interactions was a defining feature of factorial trials (6, 10, 26-28, 33, 45, 46, 63, 64). Six definitions stated that studies would only be considered factorial if patients or experimental units were deliberately allocated to different treatments by the investigator (6, 10, 47, 65-67) and a further 14 definitions described factors as “treatments” or “interventions” (7, 24, 26-28, 33, 45, 46, 63, 64, 68-71). However, one definition (72) and some other references (36, 73) suggested that classification factors (e.g. patient characteristics not under experimental control) could also be considered as factors in factorial trials. Some definitions also appeared to cover only

studies that evaluate all combinations of factors (i.e. complete factorial trials) (45, 65, 72), those with equal numbers in each cell (i.e. balanced designs) (40, 65) or those evaluating the presence/absence of treatment (41, 42, 47, 69). However, seminal references also describe unbalanced, incomplete or fractional factorial designs as types of factorial trials (10, 26, 27) and indicate that levels of each factor can be defined in any useful way (5, 10, 11, 26-28, 46).

Table 2.1. Published definitions of factorial trials

Publication	Definition of factorial trials	Combinations of ≥2 factors tested	Allow investigation of Interactions	Factors must be treatments under investigator's control (not patient characteristics)	Type of comparisons	Clearly differentiates factorial designs from others
Seminal references lacking concise definition						
Fisher 1926 (27)	Lengthy discussion of features of factorial trials	✓	✓	(✓)	Any	✓
Yates 1935 (26, 40)		✓	✓	(✓)	Any	✓
Fisher 1935-1966 (8, 28) Chapter VI		✓	✓	(✓)	Any	✓
Cox 1958 (5, 10) (Chapter 6)	Lengthy discussion, including: "We shall call each basic treatment a *factor* and the number of possible forms of a factor the number of *levels* for that factor. A particular combination of one level from each factor determines a *treatment*. The experiment as a whole is called a factorial experiment if all, or nearly all, of the factor combinations are of interest."	✓	✓	✓	Any	✓
Specific definitions: factorial experiments in general						
Yates 1937 (46, 74)	"Factorial experiments are experiments which include all combinations of several different sets of treatments or 'factors'. Information is thus simultaneously obtained on the responses to the different factors and also on the effects of changes in the level of each factor on the responses to the others."	✓	✓	(✓)		✓
Finney 1955 (9) (page 85)	"A factorial experiment is usually (but not necessarily) one in which several states of two or more factors are tested in all combinations"	✓				✓
Festing 2001 (70)	"Factorial experiments are ones in which a number of independent variables [... are] altered within a single experiment"	✓		(✓)		✓

Publication	Definition of factorial trials	Combinations of ≥2 factors tested	Allow investigation of Interactions	Factors must be treatments under investigator's control (not patient characteristics)	Type of comparisons	Clearly differentiates factorial designs from others
Muller 2002 (45) (page 340)	"A factorial design includes all combinations of levels of every treatment investigated. This feature allows estimating and testing the interaction of the two factors (any joint effect not determined from considering the predictors separately)."	✓	✓	(✓)		✓
Specific definitions: factorial trials in medicine						
Meinhert 1986 (68)	"A treatment structure in which one study treatment is used in combination with at least one other study treatment in a trial, or where multiples of a defined dose of a specified treatment are used in the same trial"	✓		(✓)	AB, Dose	No
Last 1988 (72)	"A method of setting up an experiment or study to assure that all levels of each intervention or classificatory factor occur with all levels of the others"	✓		✓/X	Any	✓
Jadad 1998 (69) (cited by (75, 76))	"An RCT has a factorial design when two or more experimental interventions are not only evaluated separately, but also in combination and against a control"	✓		(✓)	AB	✓
Redmond 2001 (63)	"Test the effect of more than one treatment (factor) using a design that permits an assessment of interactions between the treatments."	✓	✓	(✓)		✓
Armitage 2002 (6) (page 247)	"The factorial design […] permits the detection of an interaction between the two factors. This cannot be done without the factorial approach." "Strictly, the term 'factorial design' should be reserved for situations in which the factors are all controllable experimental treatments and in which all the combinations of levels are randomly allocated to the experimental units."	(✓)	✓	✓		✓
Piantadosi 2005 (33)	"Experiments that test the effect of more than one treatment using a design that permits an assessment of interactions among the treatments."	✓	✓	(✓)		✓
Couper 2005 (7)	"Has two or more factors, each having two or more levels. In clinical trial settings, a factor is usually a type of treatment, such as a surgical intervention, pharmacological treatment or some other type of intervention. For each intervention, the levels may be whether that intervention is used, different intensities of the intervention or comparisons between mutually exclusive types of intervention. For pharmacological treatments, the levels may be an active agent versus a matching placebo or varying dosages of a particular agent. In a factorial design, each participant is assigned to one level of each of the factors."	✓		(✓)	AB, Dose	✓

Publication	Definition of factorial trials	Combinations of ≥2 factors tested	Allow investigation of Interactions	Factors must be treatments under investigator's control (not patient characteristics)	Type of comparisons	Clearly differentiates factorial designs from others
Green 2006 (65)	"Assigns patients equally to each possible combination of levels of each treatment"	✓		✓		No
Fox 2006 (47)	"In a factorial clinical trial with a 2x2 format, individuals are randomly assigned to two separate interventions (e.g. interventions A and B) and these interventions are compared with their corresponding control(s)"	✓		✓	AB	✓
Lammertse 2007 (64)	"Factorial designs involve the testing of two or more treatments simultaneously for possible synergistic or antagonistic effects (combination treatments)."	✓	✓	(✓)		✓
Torgerson 2008 (71)	"A trial design where two or more different interventions are evaluated using the same participant sample."	✓		(✓)		✓
Cochrane handbook (24)	"Two (or more) intervention comparisons are carried out simultaneously"	✓		(✓)		No
Portney 2009 (67)	"A factorial design incorporates two or more independent variables, with independent groups of subjects randomly assigned to various combinations of levels of the two variables."	✓		✓		✓
Bowling 2009 (66)	"Several experimental groups are studied and the investigator deliberately varies more than one variable"	✓		✓		✓

Types of comparison: Any, acknowledges that the levels of each factor may be of any type; AB, allows only for *0* vs. *a* vs. *b* vs. *ab* comparisons that compare the presence/absence of ≥2 factors; Dose, allows for comparisons between different doses of ≥1 different drug.

Based on these definitions from the literature and the way in which the term "factorial" is used elsewhere in the literature, I define the following as being necessary and sufficient characteristics of factorial RCTs:

- **Two or more factors (each with two or more levels) are deliberately varied at the same time, with participants being randomly allocated to receive different combinations of levels for each factor.**
    - In this thesis, I consider only those studies for which experimental units (e.g. patients or hospitals) are randomly allocated to ≥2 factors that are under the investigator's control to be *factorial RCTs*. As such, my definition excludes trials in

which only one factor is randomly allocated and the other factor comprises a classification or stratification factor (e.g. age, sex or disease severity). Two-arm RCTs with stratified randomisation, subgroup analyses and/or randomised block designs (6, 67, 70) are therefore not considered factorial. Some references (particularly those outside the medical literature) suggest that factors in factorial trials can include classification or stratification factors (36, 72, 73) and/or describe studies with less than two experimentally-allocated factors as factorial (36); however, such designs cannot be considered *completely randomised* or strictly factorial (6), and my definition matches those described previously (6, 10, 47, 65-67)[4] and is consistent with the way in which the term "factorial" is typically used in the medical literature (Chapter 4).

- In particular, I believe that randomisation to ≥2 factors increases the strength of inferences about interactions and reduces the risk of bias compared with subgroup analyses (Section 10.3).[5]

- **Allow assessment of interactions between ≥2 factors**
  - The ability to evaluate interactions is one of the pivotal advantages of factorial trials (6, 9, 11, 31) and is mentioned in ten published definitions (6, 10, 26-28, 33, 45, 46, 63, 64). Factorial RCTs are the only type of trial that allows unbiased evaluation of interactions between factors without recourse to external evidence

[4] Although plant variety was one of the factors investigated in the agricultural experiments in which the factorial design was first developed (26, 28), variety nonetheless comprised a factor under experimental control in this context, since the experimental units comprised plots of land (not plants), which were randomised to different conditions (including variety) that were under the investigator's control (even though the variety of a specific plant is not).

[5] As discussed in Section 10.3, within factorial trials, patients are *randomised* to receive B or not-B as well as being randomised to receive A or not-A and the trial therefore gives an unbiased estimate of the interaction. For subgroup analyses of two-arm trials, however, patients are not randomised to the classification factor: for example, a subgroup analysis may compare the effect of the randomised treatment A vs. not-A in patients who happened to receive B with the treatment effect of A in patients not receiving B. Since allocation to B vs. not B is non-randomised, any interaction observed in this subgroup analysis may be due to patient selection: e.g. patients receiving B may be sicker or have comorbidities not observed in those patients not receiving B. Any observed difference in the effect of A between patients on B and those not on B may therefore be due to differences in patient characteristics, not treatment B. As result, the interaction estimated from a subgroup analysis is prone to selection bias, while that from a factorial trial is unbiased.

(6, 9, 11, 31, 33), although many factorial trials are under-powered to detect interactions (8, 20, 31, 47, 71).

- For 2x2 factorial trials, all four treatment groups must be evaluated to assess the interaction between the two factors (24, 33).[6] As such, any trial with <4 treatment groups cannot be considered factorial – including three-arm trials comparing *a* vs. *b* vs. *ab*. The requirement for being able to assess interactions also means that for trials comparing *0* vs. *a* vs. *b* vs. *ab*, both A and B should be administered at the same dose and in the same way when used in combination as in monotherapy, since dosing changes would otherwise confound assessment of interaction effects.

Literature searches identified no commonly-used terms that are strict synonyms of the word "factorial", although such studies are occasionally referred to as "matrix studies" or “matrix designs” (77, 78). However, some published factorial trials refer to their design as "2x2” or “two by three” without using the term "factorial" (79, 80). Search terms used for previous systematic reviews of factorial trials included "interaction" and "incremental" (1), although these terms are likely to greatly reduce specificity of literature searches since they are used so widely in other contexts: particularly in economic evaluation.

The literature also identifies several types of factorial trials and related study designs (Table 2.2, Appendix C).

[6] This is also self-evident from first principles. The interaction term indicates the extent to which the observed difference between cells *ab* and *b* differs from the comparison between *a* and *0*; all four cells in a 2x2 design are therefore required to assess interactions (24, 33). However, for larger factorial designs, interactions can be assessed while omitting certain cells providing we are willing to make assumptions about higher level interactions. For example, within a 2^3 design, we can omit the triple combination (*abc*) and still assess two-factor interactions (AB, AC and BC) if we assume that the three-factor interaction (ABC) is negligible: for example, interaction AB must equal $((\mu_a - \mu_0)+(\mu_{ac} - \mu_c))/2 - (\mu_{ab} - \mu_b)$. Similarly, for a 2x3 design where the second factor is quantitative and assumed to have linear effects, we can estimate the model $y = \beta_0 + \beta_A A + \beta_B B + \beta_{AB} AB$ even if no patients were allocated to both level 1 for A and the highest of the three quantitative levels for B.

Table 2.2. Types of factorial trial and related designs

Study design	Definition	Factorial?
Types of factorial trial		
Full factorial	All patients are randomised to >2 comparisons, with all comparisons being evaluated in all patients.	Yes: standard type of factorial trial.
Incomplete factorial (sometimes referred to as partial factorial)*	≥1 factor combinations are omitted from the design without the design corresponding to any fractional factorial design (61, 81, 82).	Yes: If the omitted cells are chosen carefully to balance the factors (9, 32) and higher-order interactions are assumed to be negligible, main effects and lower-order interactions can still be estimated. However, all 4 cells in a 2x2 factorial design are required to assess interactions (24, 33), so 3-arm trials cannot be considered to be incomplete factorials.
Fractional factorial, or fractional replication	Factorial experiment in which a defined fraction (e.g. a half or a quarter) of the treatments are omitted (6, 9, 11, 37, 61, 83).	
Partial factorial*	I define partial factorial trials as those in which a subset of patients are randomised to >1 comparison in a factorial manner, while other patients are randomised to just one comparison or to a different combination of factors (Chapter 7). Often used when economic, geographic, or clinical constraints restrict the comparisons to which patients can be randomised (61).	Yes: >2 factors and interactions can be assessed in the subset of patients randomised to ≥2 comparisons.
Unbalanced factorial designs	Different factor combinations have different sample sizes (47, 84).	Yes: Unbalanced designs complicate analysis (Section 2.6), but nonetheless allow assessment of interactions and assess ≥2 factors.
Factorial trials where interventions target different endpoints or diseases	Factorial trials that use different primary endpoints for different factors and/or randomise patients to ≥2 interventions targeting different diseases (85).	Yes: >2 factors and interactions can be assessed, although interactions may be less common in this situation (37).
Large factorial designs	Trials with a factorial matrix larger than 2x2. E.g. 2^k designs with large numbers of factors or studies with >2 levels.	Yes: same properties as full factorial trials, although there will be >1 interaction to estimate.
Study types that are not considered factorial		
Reciprocal control trials, or "balanced incomplete block designs"	Patients are randomised to 2 groups and receive interventions targeted at different endpoints, with each group acting as the control for the other (37, 61).	No: Although interactions may be less common for treatments targeting different endpoints (37), estimates from such designs may nonetheless be confounded by unforeseen interactions that may introduce bias without any opportunity to investigate such interactions.
Standardly tailored design	A heterogeneous patient cohort with a mixture of risk factors is randomised to one of two groups (61). One group receives usual care, while the other receives pre-specified treatments for all of the risk factors that they display.	No: assesses the effect of the entire package of measures in a broad population but cannot assess interactions or the effects of individual treatments.
Trials comprising two sequential	The same participants completing one RCT are re-randomised in a different comparison after treatment and follow-up of the first trial	No: ≥4 cells of a factorial trial must be within the same experimental block (e.g.

Study design	Definition	Factorial?
sub-studies	had ended (e.g. COPE (86)).	evaluated at the same time) to allow estimation of first-order interactions: otherwise the interaction is therefore confounded by the (time) blocks. However, factorial trials in which recruitment and randomisation for A partially overlaps with that of factor B are considered factorial.

* Several authors refer to incomplete factorial designs as "partial factorial designs" (30, 33, 47, 68); however, I follow the terminology of Byar (81) by using the term "incomplete factorial" to refer specifically to those trials omitting certain treatment combinations from the factorial design and reserve the term "partial factorial" to refer specifically to trials such as KAT (59) or Women's Health Initiative (62) in which some but not all patients are randomised to more than one comparison.

2.4 Advantages, disadvantages and uses of factorial trials

I then reviewed the literature on the advantages and disadvantages of factorial trials and the situations where they are particularly useful or inappropriate to understand why and when factorial designs are used, with particular reference to how this may affect analysis and economic evaluation. The review showed that factorial design can *either* be used to increase statistical efficiency/power when assessing the effect of ≥2 factors (Section 2.4.1), or to investigate interactions between factors (Section 2.4.2). A systematic review suggested that 82% of studies used a factorial design for efficiency and 18% were conducted to assess the incremental benefits of combination therapy (1).

2.4.1 Answering multiple questions efficiently

Evaluating multiple factors simultaneously is a defining feature of factorial trials (Section 2.3). Furthermore, when trials are analysed at-the-margins, the entire study population is used to assess the effect of both factors (20, 35, 39, 87). If the factors do not interact, at-the-margins analysis of a 2x2 factorial trial recruiting N participants will have the same statistical power, and estimate the effects of A and B with the same precision, as two separate trials, each with N participants (one comparing A with placebo and the other comparing B with placebo) (10,

20, 24, 30, 35, 37-39, 47, 71, 88-90). In theory, factorial trials can therefore achieve "two trials for the price of one" (2, 24, 33, 40, 85, 87) and become more efficient as the number of factors increases (10).

However, efficiency gains of this magnitude are rarely possible in practice due to interactions. Interactions will cause at-the-margins estimates to over- or underestimate the effect of each intervention in isolation (3, 4, 8, 20, 30) (Section 2.6.4). Although the bias introduced by interactions is sometimes described as a disadvantage of factorial trials (2, 34), it is in fact a drawback of the at-the-margins approach and can be overcome by analysing the trial inside-the-table or by including an interaction term within regression analyses (3, 4) (Section 2.6.3). However, these analyses base treatment effects on only half of the available sample (3, 4, 6, 9, 10, 26, 33, 46), giving a 2x2 factorial trial on N patients no greater power than two two-arm trials with ½N patients each. Sub-additive or qualitative interactions will also reduce the power of factorial trials analysed at-the-margins (1, 5, 8, 30, 40-42, 74) by masking the effects of each factor (5, 11, 30, 40-42) – potentially causing such an analysis to have less power than two two-arm trials with ½N patients each (30, 40, 74).

2.4.2 Interactions and generalisability

Fully-randomised factorial trials are also the only way to obtain an unbiased estimate of interactions between ≥2 factors from a single study (9, 11, 28, 31, 33, 39, 75, 87).[7] Furthermore, a factorial design allows us to test numerous treatment combinations in a single study and therefore identify the best treatment (46) and/or make pair-wise comparisons between specific factor combinations (2, 76).

[7] Interactions cannot be assessed using three-arm trials (Section 2.3). Furthermore, as discussed in Section 10.3, subgroup analysis of a two-arm study by use of concomitant medication is prone to bias as those patients who receive B alongside study medication A are likely to differ from those not taking B. Indirect comparisons or network meta-analyses can be used to synthesise the results of several non-factorial trials to estimate interaction effects (Section 9.6.1); although meta-analyses based on relative effects avoid bias (91), estimating interaction effects in this way will require multiple studies and will be less efficient than a single factorial trial.

Investigating interactions and assessing outcomes under a wide range of different conditions or factor level combinations also increases generalisability and external validity and widens the basis for inference (9-11, 26, 28, 30, 36, 46, 66, 67): e.g. showing whether results also apply to combination therapy or other doses.[8]

Due to variability and measurement error, however, treatment effects are never exactly additive and some apparent interaction will almost always be observed even if the true interaction is zero (6, 10, 47); subsequently, statistical tests are required to assess whether the data are consistent with zero true interaction (10). However, trials are rarely powered to detect a statistically significant interaction, (8, 20, 31, 47, 71), since the variances for two-way interactions are four times larger than main effects (26, 31, 33, 35, 37, 47) and interactions are generally much smaller than main effects (10, 36, 46). Furthermore, tests for interactions indicate only whether the data are consistent with the absence of an interaction, or estimate the maximum interaction size that is consistent with the data, but cannot rule out the possibility of interaction (10, 47). Furthermore, a statistically significant interaction may also be observed by chance: particularly for experiments where many factors are investigated (6).

Some authors have suggested that interactions are a weakness of factorial trials (2, 34). In particular, there is (justifiable) concern that the trial will observe interactions that are large

[8] Yates comments that whereas separate one-factor experiments comparing treatments using one arbitrarily-chosen fertiliser concentration, would not prove whether our results were valid for other concentrations, a factorial trial comparing treatments and concentrations simultaneously could assess interactions and test whether dose-relationships differed between treatments (26). If there was no interaction, we could conclude that the results were valid for all concentrations and maximise statistical power; conversely, identifying a significant interaction would avoid false conclusions or inappropriately generalising our results to other concentrations. Although Couper commented that factorial trials may be less appropriate for increasing external validity in medicine than in other fields, such as agriculture, as it is infeasible to control for all of the diverse factors that differ between participants (7), I feel that investigating interactions in factorial clinical trials can nonetheless widen the inferential basis and avoid inappropriate generalisation. In glaucoma, for example, most RCTs comprise non-factorial trials comparing ≥2 monotherapy agents (60, 92), although many patients receive under-researched combination therapies (93); in this setting, investigation of interactions using factorial trials would show the extent to which results of monotherapy trials are applicable to combination therapy and prevent inappropriate generalisation.

enough to bias at-the-margins analysis, but too small to be statistically significant (30, 94). However, like many others (9, 26, 34, 35, 94), I feel that it is better to be able to measure interaction effects and allow for them in the analysis rather than simply investigate factors one-at-a-time and be unable to say whether or not treatment effects are dependent on concomitant therapy. Even if the study highlights large but non-significant interactions, it will nonetheless provide preliminary information on the likely size of interactions that can be confirmed in later studies (26) and may be sufficient to exclude or identify qualitative interactions. Separate two-arm trials on each treatment simply show the efficacy of each treatment individually, yet are often used to support decisions on combination therapy.

2.4.3 Practical disadvantages

However, the added complexity of factorial trials does raise practical problems, requiring careful organisation to avoid mistakes (46) and monitor compliance (7, 30) and toxicity (30), as well as increasing the complexity of analysis. Recruitment to factorial clinical trials is often slower than other studies (7, 88, 95, 96), since they can only recruit patients who are indicated for, are able to comply with and consent to having all (or any) of the treatments under investigation (7); however, this difficulty may be partially overcome by using a partial factorial design, which allows clinicians or patients to choose which (or how many) comparisons to participate in (59, 62).

Giving two treatments simultaneously may also reduce compliance (7, 30, 40, 47) (which can introduce interactions (30)) or increase the risk of adverse events (AEs) (7, 88), which raises ethical concerns and may limit use of large factorial designs in medicine. Trials under-powered for AEs or interactions also cannot conclusively demonstrate whether an unexpectedly high AE rate in the combination therapy arm is due to A, B or a toxic interaction (7, 30) and difficulties identifying the cause of AEs may delay individual patients' recovery.

Discontinuing one factor on ethical grounds, while continuing the other will also complicate the analysis (37) and may make it difficult or impossible to evaluate the other factor (88). Nonetheless, some authors argue that the ability to start or stop factors at different times is an advantage of factorial trials (34, 37).

2.4.4 Advantages and disadvantages in specific situations

Factorial trials are clearly only of use in situations where we are interested in ≥2 factors (35) that can be administered together without changing the dose or nature of each treatment (2, 20, 33, 35, 68). For full factorial trials comparing *0* vs. *a* vs. *b* vs. *ab*, it must also be ethical to give some patients no treatment and others *ab* combination therapy (33, 61), although incomplete factorial designs relax this requirement for larger designs.

Factorial trials are particularly useful for determining the optimal dose for combination therapy (7, 30, 82, 84, 97-102), simultaneously varying dose and dosing frequency (5, 103), or demonstrating the superiority of fixed-dose combination therapy over individual monotherapies (97, 100).

Factorial trials are also often used to piggy-back a trial of a non-profitable agent (e.g. an off-patent drug or cheap supplement) (47) or testing of an immature hypothesis (85) on top of a trial of a more commercially-viable treatment. This may be why factorial trials often evaluate nutritional supplements alongside drugs (Chapter 4). Factorial trials evaluating multiple treatments may also increase the chance of the trial producing at least one interesting and novel finding that can be published in a top journal, although this strategy may also introduce a more subtle type of publication bias, whereby publications focus on the comparison that reached statistical significance and do not mention the non-significant comparison. Factorial trials may be particularly cost-effective when the fixed costs of a trial greatly exceed the

marginal cost per patient, since the added cost of introducing an additional factor may be small, even if interactions increase the necessary sample size (7).

Factorial trials can also be useful for "unpacking" a complex intervention by comparing its components separately and in combination (71). They may also help maximise the information gained from studies requiring large samples and long follow-up (5, 7, 30, 35, 73, 104). Factorial designs may be particularly suitable for studies that evaluate non-toxic agents (e.g. preventative treatments or nutritional supplements) that can be used in combination (35, 105), are likely to have small effect sizes (104) and for which interactions may be less likely (7).

The challenges raised by interactions mean that factorial trials are often reserved for situations where interactions are unlikely (1, 30, 35, 42, 90), such as when the factors influence different diseases or endpoints (4, 7, 30, 81) or when one factor may have little/no effect (42).

2.5 Notation and terminology

2.5.1 Notation

Within this thesis, I follow the established convention (9-11, 28, 67) of using upper case letters (e.g. A and B) to describe the factors within a factorial trial and the variables used to represent them in regression analyses and use combinations of upper case letters to represent interactions: e.g. "AB" represents the interaction between factors A and B, while "ABC" represents the three-factor interaction between A, B and C. For designs with >2 levels of ≥1 factors, it is conventional to use italicised lower case letters (e.g. $a_2b_0c_3d_3$, (9, 10)) to describe the levels of each factor used in each treatment. However, for brevity, I follow the convention (9, 11) for full factorial trials with two levels per factor of using italicised lower-

case letters without subscripts to indicate level 1 of a factor with two levels (e.g. "*a*") and indicate level 0 as the absence of any symbol for that factor. The cells in a 2x2 design are therefore designated *0*, *a*, *b* and *ab*, while the mean outcomes for each treatment are denoted μ_0, μ_a, μ_b and μ_{ab} (Table 2.3).

Table 2.3. Illustration of a 2x2 factorial trial evaluating the effect of two drugs (A and B), showing the notation used in the thesis.

		Factor A: presence/absence of Drug A	
		Level 0: Placebo for A	**Level 1: Active drug A**
Factor B: presence/ absence of Drug B	**Level 0: Placebo for B**	**Placebo for A + Placebo for B** Treatment label: *00* or *0* Mean (SD) outcome = μ_0 (σ_0) Sample size = N_0	**Drug A + Placebo for B** Treatment label: *a0* or *a* Mean (SD) outcome = μ_a (σ_a) Sample size = N_a
	Level 1: Active drug B	**Placebo for A + Drug B** Treatment label: *0b* or *b* Mean (SD) outcome: μ_b (σ_b) Sample size = N_b	**Drug A + Drug B** Treatment label: *ab* Mean (SD) outcome = μ_{ab} (σ_{ab}) Sample size = N_{ab}

Abbreviations: SD, standard deviation.

The ***simple effect*** of treatment A is defined as the impact of adding A to no treatment (i.e. $\mu_a - \mu_0$). The ***main effect***[9] of treatment A is defined as the average impact of adding A (averaged across patients with and without B: i.e. $0.5(\mu_{ab} + \mu_a - \mu_b - \mu_0)$).

2.5.2 Types of interaction

Two-way interactions can be calculated as $\mu_{ab}-\mu_a-\mu_b+\mu_0$ (3-5, 106): i.e. as the mean in the *ab* group, minus what we would expect given the outcome in the *0* group and the simple effects of A and B: $\mu_{ab}-(\mu_0+(\mu_a-\mu_0)+(\mu_b-\mu_0))$. Interactions can be illustrated on "two-factor interaction

[9] Within the literature and among statisticians, the term "main effect" is used in two different contexts. Some papers (e.g. (3-5)) state that within the context of a factorial trial, the main effect strictly equals the average impact of adding A, averaged across patients with and without B ($0.5(\mu_{ab} + \mu_a - \mu_b - \mu_0)$): i.e. the treatment effect estimated using at-the-margins analysis. However, the term "main effect" is more commonly used to refer to any estimate of the effect of a single factor, regardless of whether interactions terms are taken into account. For example, within a regression model $y_i = \beta_o + \beta_A A_i + \beta_B B_i + \beta_{AB} A_i B_i + \varepsilon_i$, the term "main effect" is often used to describe the terms β_A and β_B (6). Within this thesis, I use the term in both contexts: i.e. to differentiate between "simple" (inside-the-table) and "main" (at-the-margins) effects and to differentiate between interaction terms and the effect of individual factors.

graphs" (11) plotting outcome (e.g. utility) against Factor A for different levels of B, and vice versa (Figure 2.1).

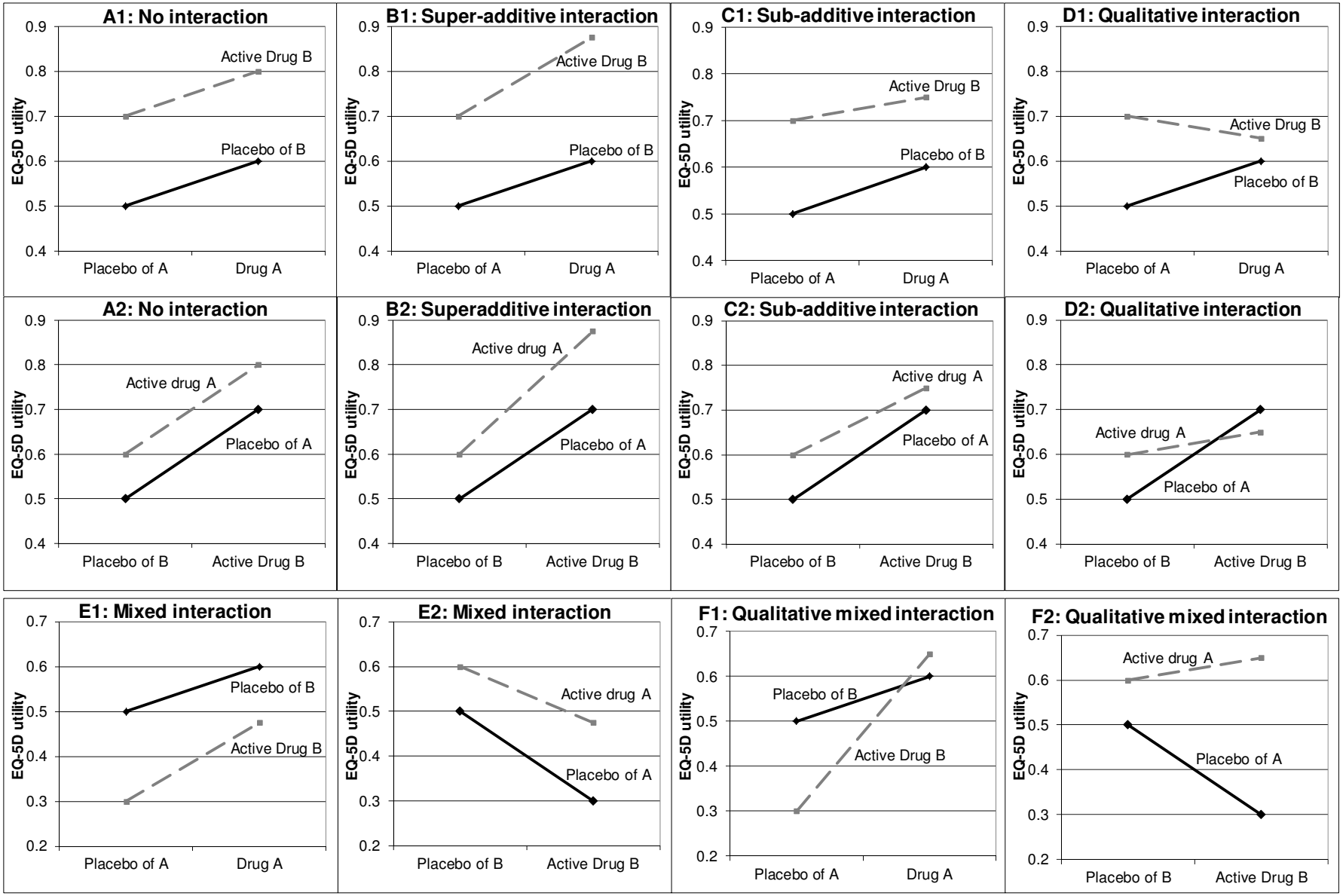
If the effect of *ab* is exactly equal to the sum of the individual effects of the two factors, there is zero interaction and effects are additive. Lines on the two-treatment interaction graph will therefore be parallel, indicating that A has the same effect regardless of whether B is given (Figure 2.1A). For example, if A increases utility by 0.1 and B increases utility by 0.2, there will be zero interaction if μ_{ab} is 0.3 higher than μ_0. Published studies have divided interactions into three categories (6, 30, 34)[10]: super-additive, sub-additive and qualitative.

Super-additive interactions mean that the effect of the combination is *greater* than the sum of the parts and that there is synergy between the factors (1, 7, 8). Such interactions will have the *same sign* as the simple effects (e.g. if μ_{ab} is 0.375 higher than μ_0, interaction=0.075). If there is a super-additive interaction, treatment effect lines will diverge (Figure 2.1B).

Sub-additive interactions indicate that the effect of the combination is *less* than the sum of the parts and that the factors have antagonistic effects (1, 8); the interaction therefore has the *opposite sign* from the simple effects (e.g. if μ_{ab} is 0.25 higher than μ_0, interaction=-0.05). Sub-additive interactions cause the effects of treatment to converge (Figure 2.1C).

[10] Alternative classification systems for interactions are also used in the literature. In particular, many authors classify interactions as "positive" or "negative" (1, 7, 8), although this classification can cause confusion as a sub-additive interaction for an outcome (e.g. utility) that increases with treatment will have a negative sign, while a sub-additive interaction for an outcome (e.g. blood pressure) that decreases with treatment will have a positive sign.

Figure 2.1. Illustration of the effect of interactions within a 2x2 factorial trial.



Qualitative interactions indicate that that the effect of at least one of the treatments under investigation changes *sign* (not just magnitude) depending on whether or not the other therapy is given (7, 8): for example, A might *increase* utility by 0.10 when used alone but *reduce* utility by 0.05 when added to *b*, such that μ_{ab} was only 0.15 higher than μ_0 (interaction=-0.15). When there is a qualitative interaction, the lines on a two-treatment interaction graph will have slopes of opposite signs and will cross on figures drawn with respect to one or both treatments (Figure 2.1D). Qualitative interactions are therefore sometimes referred to as “cross-over” interactions (8, 35, 74, 107). If the effect of A and B individually is in the same direction (e.g. if both increase outcome or both decrease outcome) a qualitative interaction will have the opposite sign but an absolute magnitude that is larger than the smaller of the two simple effects.

The distinction between super-additive, sub-additive and qualitative depends on the size and direction of the simple effects of A and B as well as the size and direction of the interaction term. However, my research demonstrates that interactions can only be neatly classified into super-additive, sub-additive and qualitative if both treatments affect outcome in the same direction (Chapter 4). If one factor decreases outcome while the other increases it, what I describe as a ***mixed interaction*** arises, whereby the interaction has the same sign as one treatment effect, but the opposite sign from the other. For example, if A increases utility by 0.1, but B decreases it by 0.2, any interaction will be mixed, since it will have the same sign as one simple effect and the opposite sign from the other. Mixed interactions will be converging with respect to one treatment (e.g. A), but diverging with respect to the other (e.g. B, Figure 2.1E). Such mixed interactions will be qualitative if the absolute magnitude of the interaction term has larger magnitude and the opposite sign from either of the simple effects (Figure 2.1F). In this example, there will be a mixed qualitative interaction if the interaction term is >0.1 (the simple effect for A), or <-0.2 (simple effect for B).

2.6 Analysis of factorial trials

2.6.1 Summary of analytical methods

Literature searches identified two main ways that factorial trials can be analysed:

- At-the-margins analysis and regression (or analysis of variance, ANOVA) without an interaction term assume that there are no interactions and pool outcomes of ≥2 cells together (Section 2.6.2).
- Inside-the-table analysis, estimation of simple effects and regression (or ANOVA) analyses with interaction terms assume that the factors interact and avoid pooling results across cells (Section 2.6.3).

2.6.1.1 Worked example

I illustrate each method using two sets of simulated results for a clinical outcome measure (e.g. diastolic blood pressure) within a hypothetical 2x2 factorial trial (Table 2.4). Both datasets assume the same true simple effects for drugs A and B, but simulation 1 assumes that the true interaction between A and B equals zero, while simulation 2 assumes a sub-additive interaction. Data were generated in Microsoft Excel 2003 by randomly sampling values for 250 patients in each cell from a normal distribution defined by the "true" cell mean and standard deviation (SD).

Table 2.4. Cell means for the two simulated datasets. For simplicity, these examples are perfectly balanced, with exactly 250 patients assigned to each factor combination.

		Placebo for A	**Drug A**
Mean (SD) outcome for simulation 1: zero interaction between A and B	**Placebo for B**	μ_0=108.4 (σ_0=19.7)	μ_a=100.2 (σ_a=18.4)
	Drug B	μ_b=89.3 (σ_b=20.1)	μ_{ab}=81.4 (σ_{ab}=20.9)
Mean (SD) outcome for simulation 2: sub-additive interaction between A and B	**Placebo for B**	μ_0=110.0 (σ_0=19.8)	μ_a=97.8 (σ_a=20.2)
	Drug B	μ_b=89.4 (σ_b=19.5)	μ_{ab}=83.8 (σ_{ab}=19.5)

2.6.2 Approaches assuming zero interaction

2.6.2.1 At-the-margins approach

The at-the-margins approach treats the factorial trial as though it were two overlapping two-arm RCTs (1) and effectively ignores the factorial design (7). Results are typically presented separately for each factor (e.g. for A vs. not-A and for B vs. not-B), rather than for the individual cells (Table 2.5), which is sometimes referred to as the "factorial manner" of presenting results (2).

Factor means are generally calculated by pooling cells *ab* and *a* (and similarly cells *b* and *0*) together and calculating the mean and SD across the pooled study groups, with treatment effects ($\theta_{A_{AtTheM\,\arg ins}}$) being calculated as the difference between (or ratio of) factor means (1) (Table 2.5)[11]:

$$\theta_{A_{AtTheM\,\arg ins}} = \frac{\sum_{n_a=1}^{N_a} x_{n_a} + \sum_{n_{ab}=1}^{N_{ab}} x_{n_{ab}}}{N_a + N_{ab}} - \frac{\sum_{n_0=1}^{N_0} x_{n_0} + \sum_{n_b=1}^{N_b} x_{n_b}}{N_0 + N_b} = \frac{\mu_a N_a + \mu_{ab} N_{ab}}{N_a + N_{ab}} - \frac{\mu_0 N_0 + \mu_b N_b}{N_0 + N_b} \quad (2.1)$$

Where x_{n_k} is the outcome observed in the nth patient in arm k. If groups have equal patient numbers, $\theta_{A_{AtTheM\,\arg ins}} = 0.5 \cdot [(\mu_a - \mu_0) + (\mu_{ab} - \mu_b)]$.

Similarly, the standard error (SE) around at-the-margins estimates of treatment effect is equal to the square root of the variance across cells *a* and *ab* combined plus the variance across cells *0* and *b* combined. At-the-margins analysis therefore estimates the average treatment effect across the study population, weighted by the number of patients assigned to each treatment-combination.

$$\mathrm{var}(\theta_{A_{AtTheM\ \arg ins}}) = \sqrt{\frac{\sum_{n_a=1}^{N_a}(x_{n_a} - \mu_a)^2 + \sum_{n_{ab}=1}^{N_{ab}}(x_{n_{ab}} - \mu_{ab})^2}{(N_a + N_{ab} - 2)(N_a + N_{ab})} + \frac{\sum_{n_0=1}^{N_0}(x_{n_0} - \mu_0)^2 + \sum_{n_b=1}^{N_b}(x_{n_b} - \mu_b)^2}{(N_0 + N_b - 2)(N_0 + N_b)}} \quad (2.2)$$

[11] McAlister et al give a worked example of at-the-margins analysis of binary variables that are additive on a multiplicative scale (1), which was adapted to derive Equation 2.1.

Table 2.5. Results of the two simulated datasets using the at-the-margins approach

	Treatment A		Treatment B	
	Placebo	**Active drug**	**Placebo**	**Active drug**
Simulation 1: zero interaction between A and B				
Factor mean (SD)	98.8 (22.1)*	90.8 (21.8)*	104.3 (19.5)*	85.4 (20.8)*
(Main) effect of treatment (SE)	-8.02 (1.39)*		-18.96 (1.28)*	
Simulation 2: sub-additive interaction between A and B				
Factor mean (SD)	99.7 (22.1)*	90.8 (21.0)*	103.9 (20.9)*	86.6 (19.6)*
(Main) effect of treatment (SE)	-8.93 (1.37)*		-17.34 (1.28)*	

* Significantly greater than zero ($p<0.05$).

2.6.2.2 Regression and ANOVA

Regression techniques or ANOVA without interaction terms provide an alternative to at-the-margins analysis and also assume no interaction. In principle, any regression technique can be used to analyse factorial trials, although continuous clinical outcome measures are generally analysed using ordinary least squares (OLS) or ANOVA (33, 67, 106), while dichotomous measures (e.g. mortality) are generally analysed using logistic regression, which assumes ***multiplicative*** treatment effects (20). ANOVA methods for analysing factorial designs are widely described (6, 9, 11, 26, 36, 46) and give identical conclusions and p-values to OLS, but different output (108). The literature also describes models for survival endpoints (40, 42, 73, 74, 94) and ordinal data (109). For simplicity, I focus on OLS within this chapter, but consider alternative model specifications elsewhere.

For a 2x2 factorial trial, the simplest regression model includes just two explanatory variables: one dummy (A) indicating whether or not factor A is given and another (B) for factor B (6, 30, 36):[12]

[12] Within this thesis, I use "dummy" or "reference cell" coding for 2^k factorial trials (assigning 1 to indicate the presence of each factor and 0 to indicate its absence). An alternative coding ("effect coding" (45) or "geometric notation" (11)) is often used in the clinical literature whereby the presence and absence of each factor are coded up as 1 and -1, respectively (11, 42, 45). In each case, interaction terms are calculated as the product of the terms of the individual factors. I use dummies as this is routinely used in health economics and facilitates interpretation of coefficients: particularly when interaction terms are included, although the same methods can be applied with either coding. Predicted values and R^2 from regression will be the same for both coding systems, although the value and interpretation of all coefficients differs: for example, the constant term from effect coding indicates

$$y_i = \beta_o + \beta_A A_i + \beta_B B_i + \varepsilon_i \quad (2.3)$$

This model implicitly assumes zero interaction between the treatments. Coefficients β_A and β_B represent the (main) treatment effects for A and B.

If patient numbers are identical in each group (i.e. if groups are perfectly balanced), the point estimates for treatment effects from this model will be identical to those from the at-the-margins method: e.g. the β_A estimates shown in the first and third columns of Table 2.6 are identical to the main effects in Table 2.5. However, at-the-margins analysis (Equation 2.1) calculates the weighted average treatment effect, giving equal weight to all experimental subjects and therefore gives different results from regression whenever groups are not perfectly balanced.[13]

SEs will also differ between at-the-margins analysis and regression unless SDs are also identical across groups, since OLS assumes that groups have equal variance (homoskedasticity), whereas at-the-margins analysis allows SEs to vary between cells depending on sample size and SD. This effect means that β_A and β_B have equal SEs within Tables 2.6, whereas the SEs for main effects in Table 2.5 differ between A and B.

---

the overall mean (across all cells), whereas the constant in reference coding indicates the mean outcome in the control group (μ_0).

[13] An alternative method is often given in older literature (3, 4, 6, 9, 10, 26, 33, 46), which facilitates estimation of main effects for 2^k trials without the need for statistical software; such estimates are equivalent to those obtained from regression providing that either the groups are perfectly balanced or there is zero interaction. Using this method, the regression coefficient β_A for a 2x2 trial with n patients per arm equals $\beta_A = \theta_{A_{AtTheM\arg ins}} = (\mu_{ab} - \mu_b + \mu_a - \mu_0)/2$, while its variance equals $\mathrm{var}(\beta_A) = \mathrm{var}(\theta_{A_{AtTheM\arg ins}}) = \sigma^2/n$. This method estimates the crude mean treatment effect, giving equal weight to all factor combinations regardless of sample size; this method is therefore misleading unless groups are perfectly balanced. By contrast, regression models estimate values for β_A, β_B and β_0 that minimise the sum of squared deviations from model predictions across all observations, giving proportionately more weight (or leverage) to outliers.

Table 2.6. Results of the two simulated datasets using OLS regression with/without an interaction term, calculated using the `regress` command in Stata version 11.0 (StataCorp LP, College Station, Texas).

		Mean (SE) for simulation 1 (zero interaction)		Mean (SE) for simulation 2 (sub-additive interaction)	
		Without interaction term	**With interaction term**	**Without interaction term**	**With interaction term**
Treatment effect for A (β_A)		-8.02 (1.25)*	-8.20 (1.77)*	-8.93 (1.25)*	-12.24 (1.76)*
Treatment effect for B (β_B)		-18.96 (1.25)*	-19.14 (1.77)*	-17.34 (1.25)*	-20.65 (1.76)*
Constant term (β_0)		108.3 (1.08)*	108.4 (1.25)*	108.4 (1.08)*	110.0 (1.25)*
Interaction term (β_{AB})		N/A	0.36 (2.50)	N/A	6.62 (2.49)*
Predictions from regression model†	Placebo	108.3 (1.08)*	108.4 (1.25)*	108.4 (1.08)*	110.0 (1.25)*
	Drug A + Placebo	100.3 (1.08)*	100.2 (1.25)*	99.4 (1.08)*	97.8 (1.25)*
	Placebo + Drug B	89.4 (1.08)*	89.3 (1.25)*	91.0 (1.08)*	89.4 (1.25)*
	Drug A + Drug B	81.3 (1.08)*	81.4 (1.25)*	82.1 (1.08)*	83.8 (1.25)*

* Significantly greater than zero ($p<0.05$).
† The predicted values and associated SEs were estimated using the `predict` post-estimation command within Stata.

Regression analysis is generally considered the most appropriate method since it allows for variations in sample size between groups (20, 36) and adjusts for the effect of the other intervention (20). Regression also allows us to control for covariates (20) (which can increase precision (20, 110, 111)) and provides a flexible range of analytical tools for different types of data. Regression models are particularly useful when one or more factor is quantitative, since they can estimate response curves or surfaces indicating the shape of the relationship between factors (e.g. dose) and outcome that can help predict the best dose combination (10, 11, 26, 46, 97, 98, 100, 112). Regression analysis also facilitates predictions of mean outcomes for each cell in the factorial design.

2.6.3 Approaches assuming interaction(s) exist

2.6.3.1 Inside-the-table approach

Inside-the-table analysis treats each cell within the factorial design as a separate treatment (1) and estimates simple effects showing differences between individual cell means. For a 2x2 factorial design, outcomes are therefore estimated and presented separately for each of the four treatment arms based only on those patients randomised to that combination of

treatments (Table 2.7), allowing the reader to see the effect of interactions directly (2) and avoiding pooling cells.

Table 2.7. Simulated results from a 2x2 factorial trial: inside-the-table approach

		Placebo for A	**Drug A**
Mean outcome (SD) for simulation 1: zero interaction between A and B	Placebo for B	108.4 (19.7)*	100.2 (18.4)*
	Drug B	89.3 (20.1)*	81.4 (20.9)*
(Simple) effect of treatment A (SE)		-8.20 (1.70)*	
(Simple) effect of treatment B (SE)		-19.14 (1.78)*	
Mean outcome (SD) for simulation 2: sub-additive interaction between A and B	Placebo for B	110.0 (19.8)*	97.8 (20.2)*
	Drug B	89.4 (19.5)*	83.8 (19.5)*
(Simple) effect of treatment A (SE)		-12.24 (1.79)*	
(Simple) effect of treatment B (SE)		-20.65 (1.75)*	

* Significantly greater than zero ($p<0.05$).

The inside-the-table outcomes for cell *a* (μ_a and σ_a, Table 2.7) therefore equal the mean and SD across the 250 patients receiving *a* (Table 2.4). The "simple effect" of A ($\theta_{A_{SimpleEffect}}$) is simply the difference between outcomes for those patients randomised to *a* (μ_a) and those randomised to *0* (μ_0) (3-5):

$$\theta_{A_{SimpleEffect}} = \mu_a - \mu_0 \qquad (2.4)$$

Furthermore, given balanced, homoskedastic groups:

$$\mathrm{var}(\theta_{A_{SimpleEffect}}) = 2\sigma^2 / N, \qquad (2.5)$$

where σ^2 equals the common population SD and N equals the number of patients in each cell. The simple effect has been shown to provide an unbiased estimate of the effect of A in the absence of B (3, 4), which is often of specific interest (36): e.g. to compare licensed drug doses.

Interactions between treatments can be estimated as (3-5, 106):

$$Interaction = \mu_{ab} - \mu_a - \mu_b + \mu_0 \qquad (2.6)$$

$$\mathrm{var}(\text{interaction}) = 4\sigma^2 / N \qquad (2.7)$$

2.6.3.2 Regression with interaction term(s)

Identical results are obtained using regression analyses including all interaction terms (Table 2.6). For a 2x2 factorial design, one interaction term (β_{AB}) is required[14]:

$$y_i = \beta_o + \beta_A A_i + \beta_B B_i + \beta_{AB} A_i B_i + \varepsilon_i \quad (2.8)$$

Within this model, the coefficients β_A and β_B represent the treatment effects for A and B, adjusted for the interaction between the two factors. Point estimates for β_A and β_B from OLS regression will be identical to the treatment effects estimated from the inside-the-table approach (even if the groups are unbalanced) and SEs will be similar, but not necessarily identical if there is heteroskedasticity (Tables 2.6-2.7). This regression model also directly estimates the interaction effect (β_{AB}), which equals the difference between the observed mean outcomes for group *ab* and $\beta_0 + \beta_A + \beta_B$.

2.6.4 Advantages and disadvantages of allowing for interactions

Increased power is the key advantage of assuming no interaction and pooling cells (3, 5, 47): pooling ensures that data from all cells is used to calculate treatment effects. By contrast, inside-the-table analysis and regression with interaction terms implicitly assume that the effect of A in the presence of B provides no information on the effect of A alone and avoid pooling data on the effect of A with/without B. Such analyses therefore base treatment effects on only half of the available data and almost always produce less precise estimates than analyses assuming no interaction (3, 5, 20, 35, 39, 47, 87). For example, within simulation 1, the SE around β_A is 1.77 with an interaction and 1.25 without (Table 2.6). Analyses allowing for interactions therefore do not exploit the efficiency advantages of factorial trials (47) and provide no more statistical power than splitting the sample in two and

[14] Unbiased analysis of 2^k designs will require multiple two-factor and ***higher-order interactions*** (9-11, 33, 36, 46), while further two-factor interactions will be needed to allow for all possible interactions in studies with >2 qualitative levels per factors (36).

conducting two separate non-factorial trials (one comparing A vs. placebo and one B vs. placebo).

However, treatment effect estimates from analyses assuming no interaction are biased whenever the interaction does not equal zero (3, 4, 8, 20, 30). The bias inherent within this approach equals half of the true interaction (3, 8) and can lead to incorrect conclusions. By contrast, simple effects from inside-the-table analysis are always an unbiased estimate of the effect of each factor in isolation (regardless of interactions) (3, 4); similarly, regression with interactions gives unbiased results providing that the normal assumptions of homoskedasticity, Gaussian residuals and independent observations hold.

To illustrate this bias, imagine that mean blood pressure with *0* (μ_0) is 100 mmHg, while μ_a=80 mmHg, μ_b=90 mmHg and μ_{ab}=75 mmHg. The interaction therefore equals 5 mmHg ($\mu_{ab}-\mu_a-\mu_b+\mu_0$). At-the-margins analysis would suggest that the effect of A vs. not-A is -17.5 mmHg ($0.5\cdot[(\mu_a-\mu_0)+(\mu_{ab}-\mu_b)]=0.5\cdot[(80-100)+(75-90)]$) and that the effect of B is -7.5 mmHg ($0.5\cdot[(\mu_b-\mu_0)+(\mu_{ab}-\mu_a)]=0.5\cdot[(90-100)+(75-80)]$). The sub-additive interaction therefore means that at-the-margins analysis overestimates μ_a-μ_0 and μ_b-μ_0 by 2.5 mmHg: half of the size of the interaction. By contrast, inside-the-table analysis will correctly estimate that the effect of A (without B) is -20 mmHg ($(\mu_a-\mu_0)=(80-100)$) and that the effect of B (without A) is -10 mmHg ($(\mu_b-\mu_0)=(90-100)$).

Estimates assuming no interaction will systematically overestimate treatment effects when super-additive interactions are present (1, 4, 30) and systematically underestimate response (or dilute treatment effects) if sub-additive interactions exist (1, 4, 30). Within Simulation 2, ignoring the sub-additive interaction decreases β_A from -8.93 to -12.24 (Table 2.6; suggesting bias=3.31). By contrast, allowing for the interaction in Simulation 1 (where the

true interaction is zero) has only a small, chance effect (decreasing β_A from -8.02 to -8.20, Table 2.6). Whereas an inefficient analysis (e.g. inside-the-table analysis in trials with zero true interaction) will approach the true value as the sample size increases towards infinity, results of a biased analysis (e.g. at-the-margins analysis of trials where factors interact) will always systematically differ from the true value regardless of sample size.

Qualitative interactions can completely distort the conclusions of analyses assuming no interaction (2, 6, 20, 42) by suggesting that treatment is effective (or ineffective) overall, when in fact treatment is effective when given alone but harmful in combination (or vice versa). For example, if the simple effects of A and B equal -10, but *ab* has the same outcome as *0* (μ_0=10, μ_a=20, μ_b=20, μ_{ab}=10), analyses assuming no interaction will estimate main effects to equal zero even though both treatments work when used in isolation. Similarly, if treatments are only effective when used together (μ_0=10, μ_a=10, μ_b=10, μ_{ab}=20), analyses assuming no interaction will find both to be effective (main effect=5).

The power of analyses ignoring interactions is also (spuriously) increased by super-additive interactions (3, 5, 30), and reduced by sub-additive interactions (1, 20, 30, 40-42). If large sub-additive or qualitative interactions exist, analyses assuming no interaction can have lower power than inside-the-table analysis (40) or non-factorial trials (5, 30, 40, 74).

Nonetheless, estimates from analyses assuming no interaction may sometimes be informative even if (***non-qualitative***) interactions exist (20, 36). In particular, at-the-margins analysis estimates the mean effect of A in a population in which the proportion of patients receiving B is equal to that seen in the trial (e.g. 50% for a balanced design), which may be useful if this reflects clinical practice (36). However, it is generally agreed that pooling results across cells will be misleading when qualitative interactions exist (20, 33).

2.6.5 Two-stage approach

The methods described in Sections 2.6.2-3 are often used as part of one of three analytical strategies. The “always pool” approach implicitly assumes that there is zero interaction and analyses the data at-the-margins or using regression without an interaction term regardless of the size or statistical significance of the interaction term (4, 5). Conversely the “never pool” approach assumes that an important interaction exists and bases conclusions on inside-the-table analyses or regression with interaction term(s) regardless of the results (3-5). As a compromise between bias and inefficiency, statisticians commonly use a two-stage testing approach (3-8, 10, 36, 65), in which an initial test is done to assess whether statistically significant interactions exist, which determines the type of analysis done in the second step. This is also known as an “adaptive” (5) or “sometimes pool” (3-5) approach.

The initial test for interaction may be done using Equations 2.6-2.7, regression/ANOVA with interaction terms or tests for interactions used in subgroup analysis (6, 107, 113, 114). The significance level used in the interaction test (alpha) is often >0.05 (4, 5, 31, 42) and may be chosen to minimise bias in treatment effect estimates (3, 4) or distortion of the type I error rate (3, 5) (i.e. the risk of rejecting the null hypothesis when it is true).

If the initial test identifies no statistically significant interactions, results are normally analysed using regression without interaction term(s) or at-the-margins analysis (3, 7, 65). Similarly, inside-the-table analysis or regression with interaction terms are typically used when statistically significant interactions are observed (3-6, 10, 65).

The two-stage approach has been shown to increase power compared with the “never pool” approach when interactions are negligible (3). However, this approach has a type I error rate that exceeds alpha (3, 5).

The two-stage approach is hindered by the fact that very few trials are powered to detect interactions (20) and may therefore miss important interactions that could bias estimates from analyses omitting interaction terms (2). The two-stage approach gives biased estimates of the effect of each treatment as monotherapy, although bias is lower than the always-pool approach unless interactions are very small (4). This approach relies upon the arbitrary choice of significance level and may encourage researchers to equate absence of proof of an interaction with proof of absence. However, the test for an interaction should more properly be viewed as a test of whether the data are consistent with the absence of an interaction or linear effects (3, 10), and is sometimes justified on the grounds that it gives the most parsimonious model (36).

The maximum statistic is a variant of this approach, which assumes that interactions cannot exist unless both treatments are effective individually (4). Several authors also recommend first investigating meaningful simple transformations to test whether these render the interaction non-significant in the initial test (6, 10, 36), which can remove non-qualitative interactions (2, 6, 10, 30, 31, 35, 36), but not qualitative ones (6, 35).

2.6.6 Bayesian priors

A Bayesian approach that uses informative priors for the interaction provides a middle-ground between including and excluding the interaction term (42, 115-117). Standard regression without an interaction term (Equation 2.3) is equivalent to Bayesian regression with a prior forcing the interaction to equal zero ($\beta_{AB} \sim N(0,0)$) (42). Similarly, the standard regression model with an interaction (Equation 2.8) is equivalent to using a flat, uninformative prior for the interaction term ($\beta_{AB} \sim N(0,\infty)$), which implicitly assumes that the interaction may take any value between -∞ and +∞ and that all values are equally likely.

In practice, researchers may be sceptical of unexpected interactions or consider small interaction values more likely than large ones. These beliefs can form the basis of sceptical or subjective priors for the interaction, such as $\beta_{AB} \sim N(0, x)$, where *x* is set to 1/1.96 times the upper 95% credible interval (the Bayesian equivalent of a confidence interval [CI]) around main effects to reflect a prior belief that we are 95% certain that the interaction will not exceed the magnitude of main effects (116). Although priors for interactions may have zero mean if sub-additive and super-additive interactions are considered equally likely or if there is little prior knowledge of the interaction, priors with negative means can be used to allow for previous evidence or opinion that sub-additive interactions are more probable (42). For example, Welton et al defined priors that reflected the belief that there is a 95% probability that the interaction will be sub-additive but non-qualitative (115).

2.6.7 Power calculations

Most factorial trials base their sample size on the power to detect significant main effects assuming no interaction between factors (20, 31, 42, 47); such power calculations use the same standard methods as would be used for a two-arm trial comparing A vs. not-A (8). If the sample size required to achieve a certain power and significance level differs between factors, researchers commonly choose whichever is largest out of the two sample sizes (20, 47). However, the true power of at-the-margins analysis depends on the interaction between treatments, as well as treatment effects, SDs and sample size (1, 8, 30, 40-42). Power calculations for at-the-margins analysis can be adjusted to allow interactions (8, 30). Alternatively, power calculations can be conducted inside-the-table for specific pairwise comparisons using the same methods as for a two-arm trial; again, researchers would typically use the largest sample size estimated across all comparisons of interest (31, 47). Alternatively, within a Bayesian framework, sample size can be estimated from posterior variances (42) or the value of information (118). The sample size required to detect

significant interactions can be estimated from the SD and anticipated interaction size; this will generally be four-fold higher than that required for main effects (20, 31). Trials powered for either main effects or simple effects will therefore be under-powered for even qualitative interactions (8, 42). Despite this, the statistical significance of interactions is often used to determine whether the primary analysis considers interactions.

2.6.8 Adjustment for multiple statistical tests

Opinions differ about whether (or to what extent) p-values from factorial trials should be adjusted to allow for multiple comparisons (6, 7, 11, 31, 35, 36, 45, 65, 67, 73). The probability of a type I error increases with the number of tests conducted, which (if alpha=0.05) means that for every 20 statistical tests, one test will falsely find a statistically significant result (6). For this reason, clinical trials pre-specify a single primary endpoint (119-123) and adjust p-values, 95% CI or critical values whenever multiple primary endpoints are analysed for a two-arm trial (120, 123); commonly used adjustments include Bonferroni, Tukey's honestly significant difference and Scheffé's method (6, 11).

For factorial trials, a minority of authors have suggested that it is necessary to adjust at-the-margins estimates for multiple comparisons (7, 31, 65), although this is rarely done in practice. Furthermore, no such adjustments would be done if the same questions were assessed in two parallel two-arm trials (7, 31, 35) and adjustments for multiple comparisons are normally considered inappropriate for tests that flow naturally from the study design (6). However, adjustments for multiple comparisons may be warranted when many different pair-wise comparisons between cell means are done in inside-the-table analysis (1, 36, 45, 67) or when comparing multiple pairs of column means for factors with >2 levels (11, 67).

Performing an initial test on the overall model using ANOVA or F-test on the whole regression model before analysing interactions and main effects may avoid inflating the risk

of type I errors without conducting highly conservative adjustments to p-values (7, 31, 65, 73). The initial test assesses whether there are any differences between any factor combinations, whether all main effects and interactions equal zero or (equivalently) whether the model explains any variability; if no differences are observed, no further tests are conducted.

2.7 Guidelines for factorial trials

My review identified three sets of guidelines making specific recommendations on conduct, analysis and reporting of factorial trials in medicine: the Consolidated Standards of Reporting Trials (CONSORT) guidelines (124); an article specific to factorial trials that is now considered an extension to CONSORT (20); and the Cochrane Handbook for systematic reviews (24).[15] Several review articles also make recommendations (1, 2, 47).

It is recommended that reports of factorial trials, like all RCTs, should follow CONSORT guidelines (124) and that primary analyses address principle research questions (20). CONSORT also recommend that factorial trials should describe their design as such and consider the possibility of interactions (124).

The Cochrane Handbook recommends presenting results at-the-margin as though the trial were two two-arm parallel group trials (like Table 2.5) (24) to facilitate incorporation of results into meta-analyses. However, several review articles strongly recommend presenting treatment means and their SDs (like Table 2.7) (1, 2, 20) as well as or instead of at-the-

[15] Two further papers give guidelines on experimental design outside medicine (32, 70), while one other guideline recommends that factorial designs are considered when designing studies on spinal-cord injuries, without making recommendations on design, analysis or reporting of such studies (64). Guidelines from the Food and Drug Administration were also identified, although these simply comment that "Properly designed factorial studies may be analyzed as a series of pairwise comparisons, representing, within a single study, separate demonstrations of activity of a drug as monotherapy and in combination with another drug", without giving sufficient information on whether the pair-wise comparisons in question should be done at-the-margins or inside-the-table.

margins presentation to allow the reader to assess the magnitude and implications of interactions for themselves (1, 2, 20). In particular, at-the-margins estimates can be calculated from treatment means, but not vice versa (1, 2), although it may be infeasible to present all cell means for studies evaluating many factors (20). Other authors recommend using inside-the-table presentation where interactions exist and at-the-margins presentation in other cases (47). Baseline data may be presented at-the-margins or inside-the-table depending on what is feasible or whether interactions are expected (20). CIs and statistical tests for the interaction effect should be reported even if the trial is powered only for main effects (1, 20).

In practice, many factorial studies ignore the implications of the factorial design (7). Despite the potential for bias (3, 4, 8, 20, 30), most studies assume no interaction or use the two-stage approach (1). Several key papers also recommend these approaches for analysis of primary trial data and for meta-analyses of factorial trials based on observations that very few clinical trials observe important or statistically significant interactions (1, 24).

2.8 Conclusions of literature review

In this chapter, I conducted a conceptual literature review of the existing literature on factorial trials, attempting to identify and summarise previous narrative reviews (e.g. (30, 35)), guidelines (e.g. (20)), methods papers (e.g. (4)) and systematic reviews (e.g. (1)). In particular, I developed a definition of factorial trials that will be used throughout this thesis: namely that ≥2 factors (each with ≥2 levels) are deliberately varied at the same time, with participants being randomly allocated to receive different combinations of levels for each factor, such that interactions between ≥2 of the factors under investigation can be assessed. I also reviewed the literature on the advantages and disadvantages of factorial designs, including efficient assessment of multiple treatments, investigation of interactions, broadened

inductive basis and added complexity. Finally, I reviewed the methods for analysing factorial trials and the different assumptions that they make regarding interactions, and the literature on how factorial trials are conducted in practice. Chapter 3 builds on this clinical literature review to discuss the issues raised by using factorial trials as the basis for economic evaluation.

Chapter 3: Issues raised by economic evaluation of factorial trials

3.1 Outline of chapter

While the rationale and appropriate analysis of clinical endpoints within factorial trials are well-established in an extensive literature (Chapter 2), little has been written or researched on the implications of factorial design for trial-based economic evaluation.

In this chapter, I first introduce the aims and principles underpinning economic evaluation (Section 3.2.1) and contrast the objectives and analytical methods for assessing efficacy and effectiveness in clinical trials with those of trial-based economic evaluation (Section 3.2.2). Section 3.3 then describes a literature review examining any previous research on economic evaluation of factorial trials.

Section 3.4 discusses the mechanisms by which interactions can arise for costs, QALYs and cost-effectiveness and highlights the potential for these interactions to be larger for economic outcomes than for clinical endpoints. Section 3.5 then discusses whether the treatments evaluated in factorial trials should be evaluated using decision rules for “independent” or “mutually-exclusive” alternatives. Section 3.6 discusses the challenges of analysing and interpreting economic data from factorial trials and proposes several solutions. Section 3.7 continues by discussing the tests and decision rules that can be used to identify which interactions need to be taken into account in the analysis. Each of these sections illustrates the challenges and methodology using a worked example. Section 3.8 introduces a novel measure of the ***opportunity cost of ignoring interactions***, which is followed by conclusions in Section 3.9.

This chapter is based on my own experience of conducting trial-based economic evaluation (with reference to standard textbooks and guidelines) and my thoughts on the impact of this for factorial trials. These are supplemented, where cited, by methodological papers on specific issues and the conceptual review described in Chapter 2.

3.2 Introduction to trial-based economic evaluation

3.2.1 Aims and principles of economic evaluation

Economic evaluation comprises a set of tools for assessing whether a technology or intervention represents good value for money. It has been defined as "the comparative analysis of alternative courses of action in terms of both their costs and consequences" (125).[16] In healthcare, most economic evaluations involve cost-effectiveness analysis (CEA) or cost-utility analysis (CUA) (126),[17] in which cost-effectiveness is quantified as an incremental cost-effectiveness ratio (ICER) equal to the difference in cost between two treatment strategies, divided by the difference in health outcomes (Figure 3.1).

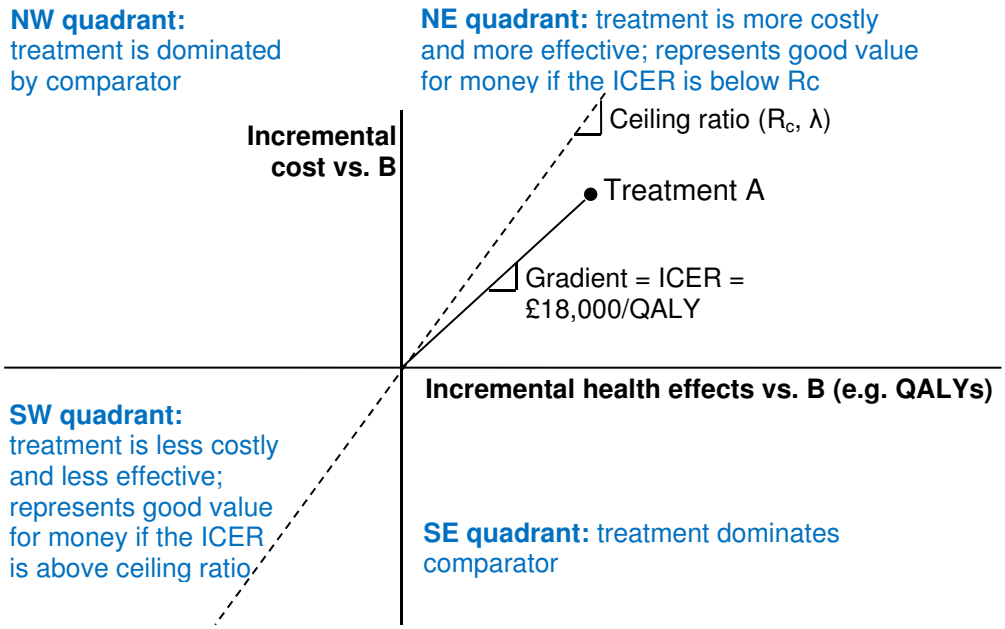
$$ICER = \frac{Cost_A - Cost_B}{QALYs_A - QALYs_B} \quad (3.1)$$

CEA measures health outcomes in natural units (e.g. numbers of patients responding or life-years gained), while CUA measures health outcomes in QALYs, which capture both quality and length of life (125). Interventions are always compared against the next best alternative treatment.

[16] Quote taken from page 9.

[17] The other main forms of economic evaluation (cost-consequence analysis, CCA, and cost-benefit analysis, CBA (125)) are used less commonly (126) and are beyond the scope of this thesis. Cost-minimisation analysis (CMA) involves comparison of alternatives based on cost alone, implicitly assuming that all treatments have equal health effects, but is no longer considered a full form of economic evaluation (125, 126). The terminology used to describe different types of economic evaluations varies between countries. Here, I use the UK terminology, which defines CUA as any economic evaluation measuring health in non-monetary composite units (e.g. QALYs or disability-adjusted life-years) that capture both quality and length of life; conversely, CEA is defined as any evaluation calculating an ICER measured in any other units. In the US literature, however, CEA is often used as a blanket term to cover all economic evaluations calculating ICERs, with CUA being a sub-category of CEA.

Figure 3.1. Cost-effectiveness plane



CEA and CUA aim to inform decisions about how to allocate resources to maximise the amount of health gained from a fixed healthcare budget (125, 127, 128). Since the healthcare budget is largely fixed at any given point in time, committing resources to any purpose has an opportunity cost equal to the health that is forgone from the best alternative use of those resources. For example, CUA may suggest that a new treatment (A) costs £18,000 per QALY gained compared with current practice (B; Figure 3.1). Given a fixed budget, the opportunity cost of adopting A is that we can no longer afford to give treatment C for a different condition. A is therefore good value for money (and will increase the health gained from the budget) only if the ICER for A vs. B is lower than that for C vs. its next best alternative: i.e. if the benefits of adopting A exceed the opportunity cost (not doing C). In practice, however, it is rarely possible to identify which treatment(s) will (or should) be displaced by a new technology.

League tables comprise one theoretical approach to decision-making (16). A league table is constructed by ranking all healthcare interventions by their ICERs (with each ICER measured relative to the next best alternative treatment for that patient group). Treatments are then

adopted in ascending order of ICER until the cumulative cost of implementing them equals the healthcare budget (Section 3.5). In practice, however, it is unlikely to be feasible to assess the cost-effectiveness of all interventions using timely, jurisdiction-specific studies that use robust, consistent methodology.

CUA and CEA results are therefore normally interpreted using a ceiling ratio (referred to algebraically as *Rc*, Figure 3.1) that represents an estimate of the shadow price of a QALY (129). This ceiling ratio may be derived from a league table (Section 3.5), or past healthcare decisions (16, 125, 129), although many countries use a simple "rule of thumb": in the UK, for example, the National Institute for Health and Care Excellence (NICE) considers treatments costing <£20,000-£30,000/QALY gained to be cost-effective (130).

However, ICERs have several statistical properties that make them difficult to use in statistical analyses (Section 3.6.2.2). Cost-effectiveness is therefore often expressed as the net monetary benefit (NMB) or net health benefit (NHB) (51, 52, 131-133):[18]

$$NMB = QALYs \bullet Rc - Cost \qquad NHB = QALYs - Cost / Rc \qquad (3.2)$$

3.2.2 Trial-based economic evaluation

Economic evaluation requires data on costs and health benefits for ≥2 alternatives. Costs and benefits are typically estimated using decision-analytical models and/or RCT(s).

Around 30% of published economic evaluations are based on a single trial (13, 134). RCTs provide unbiased estimates of treatment effect with high internal validity and enable prospective collection of data on healthcare resource use and patients' responses to multi-

[18] For consistency, I use NMB throughout this thesis to quantify cost-effectiveness and to refer to net benefits in general. However, I could, equivalently, have used NHB throughout, which would have given identical results and conclusions: the only difference would have been expression of net benefits in health units (e.g. QALYs) rather than monetary terms.

attribute utility scales, facilitating unbiased comparison of costs and QALYs between study arms (19, 134-136). Economic evaluations often “piggy-back” onto trials that were primarily designed to answer clinical questions (19, 125, 137, 138); this approach provides a practical, cost-effective framework for collecting economic data, but may mean that the trial is under-powered for economic endpoints (18, 125, 139-141), includes limited follow-up (17, 18, 125, 134, 137, 138, 142, 143) or does not reflect routine clinical practice (18, 125, 134, 137, 138, 143, 144). However, many of these problems can be overcome by planning the study with the economic evaluation in mind and conducting a pragmatic, or naturalistic, trial that aims to inform decision-making, mirrors routine clinical practice as closely as possible and assesses real-world effectiveness rather than efficacy under controlled conditions (17-19, 125, 137, 138, 142, 143).

Model-based economic evaluations, which synthesise data from different sources (typically including RCTs) using a decision-analytical model (12-14, 16, 125), provide an alternative method to assess cost-effectiveness but rely upon assumptions and secondary data (12, 125, 137). However, in practice, many model-based economic evaluations incorporate individual-patient data (13, 14, 16, 134, 143) and trial-based economic evaluations often require modelling to extrapolate beyond the end of the trial (12, 16-19, 134, 135, 142, 143, 145) or generalise trial results to the population of interest (16, 143).

Trial-based economic evaluation is now well-established, with numerous guidelines and textbooks (16-19, 49, 142). However, the objectives, statistical approach and conduct of economic evaluation differ markedly from analysis of clinical endpoints (Table 3.1). As I discuss in the remainder of this chapter, these differences between evaluations of efficacy and cost-effectiveness have particular implications for analysis of factorial trials. Table 3.1 therefore gives context to the subsequent discussion by contrasting typical practice within economic evaluation and analysis of clinical endpoints from RCTs.

Table 3.1. Comparison of approaches used in analysis of clinical trials and trial-based economic evaluation

Approach used in analysis of clinical trials	Trial-based economic evaluation
Explanatory trials (which comprise most trials) aim to assess efficacy and tolerability under controlled conditions with high internal validity (137, 138, 146), whereas pragmatic trials aim to inform decisions about which treatment is most effective or best tolerated in routine clinical practice (119, 138, 144, 146).	Aim to inform decisions about resource allocation (134, 137, 138). Trials used to assess cost-effectiveness should ideally use a pragmatic, naturalistic protocol and aim to mirror routine clinical practice in the jurisdiction(s) of interest as closely as possible, include a wide variety of patients and measure "real-world" effectiveness (17-19, 125, 134, 137, 138, 142, 143, 146).
Most trials compare just two treatments (119, 122, 134) and the comparator is often chosen on regulatory or marketing grounds (147). Placebo is often used when efficacy has not yet been demonstrated (122, 147), while the best available, licensed current treatment is often used in Phase III/IV to compare efficacy and safety against current practice (147).	Comparator(s) for economic evaluation should ideally comprise the main (or most cost-effective) alternative(s) used in routine clinical practice in the relevant jurisdiction (17, 125, 134, 137, 138, 143, 145). Multiple comparators are often appropriate (125, 134). Modelling may be required to assess cost-effectiveness against the most relevant comparator (16, 134, 143).
Investigators should pre-specify a single primary endpoint (119-123), which is specific to the condition in question and should ideally be measured objectively (especially for unblinded studies) (121-123). P-values must be adjusted in trials with >1 primary endpoint (120, 123). Surrogate markers are often used, especially for explanatory (rather than pragmatic) trials (146), but require evidence linking them to clinically meaningful endpoints, such as death or onset of disease (119, 120, 123).	Conclusions should be based on the cost/QALY or NMB (18, 19, 145, 148), although event-based approaches enable QALY calculations to be informed by the primary endpoint (16, 49, 149). Intermediate endpoints (e.g. blood pressure) should be avoided unless there is evidence linking them to final outcomes (16-18, 125, 137, 138, 143, 145).
Primary analysis should be conducted at a single pre-specified timepoint (with or without allowance for repeated measures) (120, 121). The benefits of repeated intermediate evaluations of patients may be limited (119, 122), but can increase power (123). Follow-up often ends when the clinical event of interest occurs (18, 125, 137, 138).	Addressing the economic question requires a time horizon long enough to capture all relevant differences between treatments (134, 138, 143, 145): including costs and health outcomes after clinical events (18, 125, 137, 138). Extrapolation beyond the end of the trial is often required to achieve this (16-18, 134, 135, 142, 143, 145). Some guidelines recommend that a purely within-trial analysis (taking the end of the trial or median follow-up as the time horizon) is also conducted (17, 18, 142), although others warn that this could be misleading (134). Future costs and outcomes should be discounted (18, 145).
Randomisation and adequate concealment of randomisation help avoid selection bias and ensure that there are no systematic differences between study arms (87, 119-123); chance differences between study arms may occur in small trials, although between-arm differences in baseline characteristics should approach zero as sample size increases. Stratification can be used to ensure that certain disease categories that are expected to respond differently are evenly distributed between groups (120-123, 147). Unbiased estimates of treatment effects can therefore be calculated as the difference in post-treatment outcomes or change from baseline between arms (122, 123), although some trials adjust for covariates to increase power (119, 120, 122, 123, 146).	Randomisation and allocation concealment are recommended to reduce bias (134-136). Treatment effects (incremental costs and effects) can be calculated as the difference in the arithmetic mean cost divided by the difference in arithmetic mean QALYs with/without adjustment for covariates (17-19, 49, 127, 136, 138). However, adjustment for baseline utility is recommended to avoid bias (150).

Approach used in analysis of clinical trials	Trial-based economic evaluation
Blinding of patients, their physicians and/or those assessing outcomes of treatment avoids bias introduced by the placebo effect, observer bias and differential management based on known treatment allocation (87, 119-123, 147)	Blinding helps to avoid bias. However, use of placebos may limit external validity since placebo treatment may affect resource use or outcome but would not be given in routine clinical practice (125, 137).
ITT analysis (including all patients as-randomised) is recommended to avoid bias that may result from correlations between adherence and prognosis and ensure that results approximate efficacy in practice (87, 119, 120, 122, 123). However, imputation of missing data is rarely conducted and complete-case analysis is commonly used (119, 123): partly as analyses are often based on a single outcome measure, for which complete-case analysis controlling for relevant covariates is valid under the assumption of missing at random.	ITT analysis is recommended (17, 18). Imputation of missing data (particularly multiple imputation) is frequently conducted (151) and recommended (16-19), particularly as multiple variables are aggregated to form the ICER (16, 125).
Sample size is chosen to ensure that the study has adequate (normally 80%) power to detect a specified difference in the primary endpoint, based on assumptions about likely SDs (119-123, 152, 153). In practice, however, the difference in outcomes deemed clinically significant is often back-calculated from the sample size that can realistically be recruited (122, 152).	Sample size is typically determined by clinical endpoints (17-19, 138, 142, 146, 152), although guidelines recommend power calculations on economic endpoints to decide whether economic evaluation is feasible (18). Methods for determining sample size based on frequentist methods (17, 49, 142) or EVSI (49) exist, but are rarely used at present. As result, many trials are under-powered to detect differences in costs and/or QALYs (18, 137-141, 146).
Non-parametric methods are often used when outcomes are not normally distributed (122, 123). Comparisons may be based on relative effects or absolute mean differences (122, 123).	Differences in costs and QALYs should be based on absolute differences in arithmetic means regardless of distribution (17-19, 49, 125, 138, 142, 154-156) to reflect total budgetary costs and to enable maximisation of expected health outcomes from the budget (17, 154).
Almost all trials are analysed using strictly frequentist analyses (119, 120, 123).	Bayesian methods and interpretation are commonly used alongside frequentist methods: particularly CEACs (17, 49, 125, 157) and (less commonly) value of information analysis (49, 157).
Conclusions are driven by statistical significance at alpha=0.05 (120, 122, 123, 153). For superiority trials, the primary aim is to detect whether or not treatment is superior to control, rather than necessarily estimate the difference in outcomes between arms; clinical analyses therefore focus on hypothesis testing rather than estimation (120, 123, 153), although estimating the magnitude of effects remains important (120, 122, 123). If the study finds treatment to be statistically significantly superior to control in the primary endpoint at the 0.05 level, it will typically be concluded that treatment is more effective than control (122, 123, 153); if no significant difference is observed, investigators typically conclude that there is no evidence of a difference or that the study shows that the difference between the treatments is likely to be <X (the upper 95% CI) (122, 123).	Many researchers argue that inference is irrelevant and that the option with highest expected NMB should be adopted regardless of statistical significance, with uncertainty driving only conclusions about further research (125, 157). However, some other authors disagree and conclusions (or the strength with which conclusions are drawn) are often driven by statistical significance. Economic evaluation focuses on estimating the ICER (or NMB) and uncertainty around it (18, 49, 125, 134, 136, 138), including the probability and consequences of making the wrong decision (18, 134, 136).
Non-inferiority or equivalence trials are powered to exclude a particular difference in efficacy by	Analysts should conduct CEA or CUA regardless of whether statistically significant differences in

Approach used in analysis of clinical trials	Trial-based economic evaluation
selecting the sample size such that the 95% CI around the difference in efficacy excludes clinically significant values (119, 120, 122, 123). If the non-inferiority endpoint is met, investigators will typically conclude that treatment is non-inferior to control at the 0.05 level (123).	health outcomes are observed (17, 18, 125, 126, 148, 158).
Relative effects for clinical endpoints can usually be generalised to other countries or settings (17, 125); as result, most meta-analyses assume that relative effects are either constant or exchangeable between trials regardless of where they were conducted (134, 159).	Variations in prices, relevant comparators and clinical practice mean that incremental costs and ICERs cannot normally be generalised to other countries or settings (17, 18, 125, 134, 137, 138, 142, 160). Glick et al recommend that trial-based evaluations collect data on whether cost-effectiveness is affected by the other treatments given and whether cost-effectiveness differs between developed and developing countries (17).

Abbreviations: CEAC, cost-effectiveness acceptability curve; EVSI, expected value of sample information; ITT, intention-to-treat.

3.3 Previous methodological studies on research on economic evaluation of factorial trials

Literature searches were conducted to identify any previous research on economic evaluation of factorial trials. These included review of all hits identified in the literature searches described in Chapters 2 and 4, searches of past papers presented to the Health Economists' Study Group (HESG, http://www.hesg.org.uk) for the term "factorial" and searches of the British Library Electronic Theses Online Service (EThOS, http://ethos.bl.uk) for the terms "factorial" and "cost" in combination (conducted on 7th October 2011). The searches of HESG and EThOS were replicated on 7th August 2013, alongside a search of Medline via PubMed for (factorial[Title]) AND ("economic evaluation" OR "cost-effectiveness"). These searches identified no methodological studies, reviews or guidelines on economic evaluation of factorial trials other than those resulting from this thesis.

However, one previous study on Bayesian methods for trial-based economic evaluation has developed methods relevant to analysis of factorial trials (115, 116, 118). The authors analysed two cluster-randomised factorial trials using a Bayesian logistic regression model

that allowed for the effect of each factor, interactions and clustering by centre. The model was used to estimate the probability of each intervention being cost-effective and the value of information. The authors also assessed the impact of placing an informative prior on the interaction term, which was based on the belief that a priori there was a 95% probability that the interaction was sub-additive but not qualitative (115, 116). The conclusions, value of information and the probability that the optimum treatment was cost-effective were very sensitive to the assumptions around the interaction (informative prior, vague prior or assuming no interaction). Within the discussion of the paper at the Royal Statistical Society (115), it was suggested that there may also be a correlation between main effects and the interaction term if both were affected by the ease by which women can be "persuaded" to attend breast cancer screening. However, the papers did not discuss any other issues relating to the design or analysis of economic evaluations based on factorial trials.

3.4 Causes of interactions in costs, QALYs and NMB

For clinical outcomes (such as blood pressure or the log-odds of events), interactions may arise due to non-compliance (30) or pharmacokinetic, biological or behavioural mechanisms (2, 20, 30). Interactions may also be an artefact of the scale of analysis (2, 6, 10, 30, 31, 35, 36): if two factors have a multiplicative effect, whereby the effect of A and B in combination is equal to the product (not the sum) of the individual effects of A and B, there will be no interaction on a logarithmic scale, but a positive interaction on a natural scale.[19] For example, if 23% of patients receiving no treatment have a cardiovascular event, compared with 20% on *a*, 13% on *b* and 11% on *ab*, there is zero interaction on a log-odds scale but an interaction of 0.01 [0.23-0.20-0.13+0.11] on a natural probability scale (Table 3.2). Such positive interactions are super-additive if simple effects are also positive (since simple effects

[19] The reverse is also true: if there is a complete absence of an interaction between two factors on an additive scale, there necessarily must be an interaction on a multiplicative scale (35).

and the interaction have the same sign), and sub-additive if (as in this case) simple effects are negative.

Table 3.2. Example of interactions resulting from treatments having a multiplicative effect on the odds of clinical events.

	Likelihood of clinical events			Drug cost	Event cost	Total cost	Total QALYs	NMB
	Log-odds	Odds	Probability*					
0	-1.20	0.30	0.23	£0	£277	£277	0.815	£16,023
a	-1.40	0.25	0.20	£200	£237	£437	0.842	£16,403
b	-1.90	0.15	0.13	£350	£157	£507	0.896	£17,413
ab	-2.10	0.12	0.11	£550	£131	£681	0.912	£17,559
Interaction	0.00	0.03	0.01	£0	£15	£15	-0.010	-£234

* Probability = odds/(1+ odds)
The cells of a 2x2 factorial design are referred to as *0*, *a*, *b* and *ab* following the mathematical notation defined in Section 2.5.1.

In addition to those mechanisms producing interactions for clinical endpoints, several additional mechanisms produce interactions for costs, QALYs and NMB, which are classified in the taxonomy shown in Table 3.3. Worked examples of each type are shown in Appendix D.

Table 3.3. A taxonomy of types of interaction

Interactions between treatments given to the same patients	1: Direct pharmacological, behavioural or biological mechanisms [H*]	
	2: Scale effects	2a: Multiplicative effects on the risk/hazard/odds of clinical events or mortality [A*]
		2b: Multiplicative effects between quality and length of life [B*]
		2c: Multiplicative effects on cost [C,I*]
		2d: Multiplicative effects between immediate mortality and remaining life expectancy [K*]
		2e: Non-multiplicative scale effects [Q*]
	3: Diminishing (or increasing) marginal effects on HRQoL [D,E*], including ceiling effects on quality and/or length of life [P*]	
	4: Patient pathway	4a: Earlier intervention affects costs or benefits of later intervention (or vice versa) [F*]
		4b: Interactions between diseases [L*]
		4c: Effect of comorbid conditions on treatment costs [M*]
		4d: Future costs [J*]
5: Interactions between treatments given to different patients treated by the same staff or in the same healthcare facilities [N, O*]		

* Worked examples of each type of interaction are shown in Appendix D.

Interactions are most likely to arise between treatments given to the same patients. Firstly, pharmacological, behavioural and biological mechanisms can introduce interactions in costs and QALYs (2, 20, 30) (category 1, example H). For example, one drug may reduce or

accelerate metabolism of the other, giving a second intervention may affect compliance with the first treatment, or biological mechanisms (e.g. homeostasis) may mean that adding A in patients already receiving B (which targets the same biological target) has less impact than when A is given alone. However, the magnitude of these effects may be small compared with other types of interaction unless giving both drugs together produces severe side-effects.

Secondly, treatments will frequently have multiplicative effects or other scale effects on the incidence of clinical events, QALYs or costs (category 2). In particular, interventions that affect the risk of clinical events (e.g. death or stroke) generally have multiplicative effects on outcome (20, 161), increasing or decreasing risk, odds or hazard by a certain percentage. This means that the absolute effect of treatment is larger for high-risk patients and smaller for patients who are already receiving treatment. If clinical events increase costs and reduce quality or length of life, treatments that reduce the chance of events by X% will have super-additive interactions for costs and sub-additive interactions for QALYs (Table 3.2 and example A). All life-extending treatments are likely to have multiplicative effects on all-cause mortality, even if they are used to treat different diseases in the same patients. In principle, other scale effects are possible (category 2e). For example, small interactions can also arise from scale effects between odds, hazards and probabilities (example Q): even if treatments have additive effects on event rates, there will be a small interaction for event probabilities (and vice versa). However, these effects are likely to be small.

Super-additive interactions for QALYs are also possible if one intervention extends life expectancy, thereby increasing the QALY gains from an intervention improving HRQoL (and vice versa) since QALYs equal utility multiplied by time (category 2b, example B). Many treatments and covariates also have multiplicative effects on cost (162): e.g. interventions may reduce length of stay or numbers of consultations by 50% (category 2c, example I). Large super-additive interactions for costs are particularly likely within trials (e.g. IVAN), that

compare several doses regimens, since drug cost equals cost per dose multiplied by frequency (example C).

Thirdly, giving further healthcare interventions for the same or different diseases to the same patients may give diminishing marginal returns for utilities (163, 164), as people may experience a larger improvement in HRQoL from the first intervention than they will from any subsequent one, leading to a sub-additive interaction (category 3). Both additive and multiplicative models of the effect of comorbid conditions on utility have been shown to systematically underestimate utilities for patients with several comorbid conditions (165-167) and utility for patients with comorbid conditions is often little lower than the worst of the individual conditions (165, 166).

Diminishing marginal returns are also built into many utility measures. The Health Utilities Index (HUI) assumes that attributes have a multiplicative effect on utility (164). The UK EQ-5D tariff includes the N3 interaction term and its coefficient values mean that utility improves more when moving from level 3 to level 2 on any domain than moving from level 2 to level 1 (168) (example D). The tariff can also produce super-additive interactions for QALYs: e.g. if two EQ-5D domains (influenced by two interventions) need to improve to remove the N3 or constant term (example E). Ceiling effects within multi-attribute utility scales or other score-based measures can also produce diminishing marginal returns, since the ceiling (e.g. the maximum utility of 1) will limit the extent to which treatments can improve scores or utilities (example P). Ceiling effects will also arise for life-years within studies with a fixed time horizon, since no patients can accrue more than five life-years within a five-year trial, even if treatments had additive effects on life expectancy over a longer period.

Interactions can also arise from the patient pathway (category 4). However, these interactions are less relevant for factorial trials, and are therefore discussed in Section 9.3. Although interactions are most likely to occur between treatments given to the same patients,

there may also be interactions between treatments given to different patients by the same medical staff or in the same medical facility (category 5, Section 9.3).

These mechanisms frequently produce interactions in economic outcomes even when there is no interaction for clinical endpoints. This has also been observed for interactions between treatment and subgroup or country, where heterogeneity in ICERs is frequently observed in trials with homogeneous odds ratios for clinical endpoints (17). A factorial design is also often chosen in cases where no interaction is expected for the primary clinical endpoint (1, 30, 35, 42, 90), although the impact of trial design on economic endpoints is rarely considered. This selection effect may reduce the number of factorial trials with interactions on clinical endpoints (1), but not the incidence of interactions for economic outcomes, further increasing the chance that there are interactions for economic endpoints in studies with additive clinical effects.

Since NMB equals the product of ceiling ratio and QALYs, minus cost (Equation 3.2), super-additive interactions for costs and sub-additive interactions for QALYs will combine to produce even larger sub-additive interactions for NMB (Table 3.2 and Section 3.4.1). Indeed, all analyses observing non-zero interactions for costs or QALYs and ≥1 treatment lying in the north-east or south-west quadrant will observe qualitative interactions for NMB at some ceiling ratio.[20] By contrast, interactions for clinical outcomes are unlikely to occur unless both factors have an effect (4, 30, 35, 46).

[20] Whenever ≥1 treatment lies in the north-east or south-west quadrant and interactions are not exactly equal to zero, there will be a qualitative interaction for NMB at a ceiling ratio equal to the ICER for this treatment relative to its comparator. This is because when the ceiling ratio equals the ICER, the incremental NMB for this treatment relative to its comparator will equal zero and will therefore have an absolute magnitude smaller than any non-zero interaction term. This means that at a ceiling ratio of either ICER+δ or ICER-δ (where δ equals a very small number), the interaction for NMB will always be qualitative: i.e. have a magnitude that is larger than one or more simple effect that has the opposite sign. This qualitative interaction will change the ranking of treatments: e.g. causing *a* to be cost-effective [i.e. increases NMB] compared with *0* even if *ab* is not cost-effective [reduces NMB] compared with *a*. Conversely, if both *a* and *b* dominate (or are dominated by) *0*, qualitative interactions in NMB could only arise if there were qualitative interactions for both costs and QALYs.

3.4.1 Worked example

Data on costs and QALYs for a hypothetical 2x2 factorial trial were simulated to illustrate the impact that clinical events and additive intervention costs can have on the interactions for economic endpoints. The same example is used in Sections 3.6-8.

A simple model based on the following assumptions and data inputs (Table 3.4) was used to simulate data similar to what might be obtained from an RCT of hypertension drugs A and B. In the absence of events and treatment, patients accrued no costs and accrued an average of 25 QALYs, although 0.23% of untreated patients were assumed to experience clinical events, such as stroke, each year (odds=0.30). Both drugs reduce the odds of events, with zero interaction between drugs on a log-odds scale. Each event increases healthcare costs and reduces QALYs by an amount independent of treatment. Treatment side-effects also reduce patients' HRQoL. Microsoft Excel 2003 was used to generate 250 hypothetical patients within each treatment arm.

Table 3.4. Data inputs used in the worked example

Data input		**Mean (SD)**	**Distribution**
Drug cost per year of life	Drug A	£900 (£0)	None (fixed)
	Drug B	£1,500 (£0)	None (fixed)
Cost without events or treatment		£0 (£0)	None (fixed)
Life expectancy without events or treatment (years)		30 (3)	Gamma distribution
Utility without events or treatment		0.835 (0.200)	Beta distribution
Odds of clinical events with no treatment		0.30 (0)	*
Odds ratio vs. no treatment	Drug A	0.82 (0)	
	Drug B	0.70 (0)	
Increase in healthcare costs for each event		£12,000 (£2,000)	Gamma distribution
Reduction in QALYs for each event		1.0 (0.2)	Gamma distribution
Percentage reduction in HRQoL due to side-effects	Drug A	1.0% (0.1%)	Normal distribution
	Drug B	2.0% (0.2%)	Normal distribution

* The number of events experienced by each patient was simulated by comparing the annual probability of events for each patient with a uniform distribution and repeating the process over each year of life up to a patient-specific time horizon defined by the sampled life expectancy value.

Since treatments have a multiplicative effect on the odds of having an event, we observe a sub-additive interaction on the number of events (Table 3.5), with a positive interaction term

and negative main effects. Since each event increases costs but reduces QALYs, we see a positive interaction for costs and a negative interaction for QALYs. After allowing for additive effects (i.e. drug costs and the disutility of side-effects), main effects are positive for costs and QALYs (i.e. both A and B increase costs and improve health). Overall, we therefore see a large super-additive interaction for costs and a sub-additive interaction for QALYs (Table 3.5). Combining costs and QALYs produces a qualitative interaction in NMB at a £30,000/QALY ceiling ratio, since A increases NMB (i.e. is cost-effective) when used alone, but reduces NMB when added to *b*. However, as is commonly observed in trial-based economic evaluations, there is substantial variability and uncertainty in costs, QALYs and NMB that reduces the power to draw inferences about either main effects or interactions.

Table 3.5. Worked example of economic evaluation of 2x2 factorial design: group means

	***0*: Placebo for A + placebo for B (n=250)**	***a*: Drug A + placebo for B (n=250)**	***b*: Drug B + placebo for A (n=250)**	***ab*: Drug A + Drug B (n=250)**	**Interaction**
Mean (SD) no. events per patient	7.2 (2.3)	5.9 (2.4)	5.3 (2.1)	4.5 (2.1)	0.5 (sub-additive)
Mean (SD) cost per patient	£87,804 (£33,508)	£98,324 (£32,408)	£109,109 (£28,851)	£125,015 (£29,958)	£5,386 (super-additive)
Mean (SD) QALYs per patient	18.1 (6.6)	18.9 (5.9)	19.6 (6.0)	19.8 (5.6)	-0.7 (sub-additive)
Mean (SD) total NMB per patient at £30,000/QALY ceiling ratio	£455,010 (£211,746)	£470,155 (£186,507)	£479,504 (£188,881)	£468,985 (£174,617)	-£25,664 (qualitative)

3.5 Research questions for factorial trials and impact on healthcare decisions

Given that economic evaluations aim to directly inform healthcare decision-making, it is essential that factorial trial-based economic evaluations address the most appropriate research question(s) and use decision rule(s) that enable us to adopt treatments that maximise the health gains from the finite healthcare budget. Within economic evaluation in general, two different decision rules are used to allocate resources between multiple

competing uses (16, 125, 128, 169-177), depending on whether treatments are considered to be independent or mutually-exclusive.[21]

When the alternatives are considered to be *independent* (a "shopping spree" problem (171, 172, 174)), the ICERs for the two treatments (each relative to their next best non-dominated alternative) can simply be compared side-by-side (16, 169-175, 177, 182).

If we have a league table showing the cost-effectiveness and budget impact of all possible independent healthcare interventions, we can maximise the amount of health gain from the budget by adopting treatments in ascending order of ICER (16, 169-171, 173, 177, 182).[22] For example, in Table 3.6, we will maximise health given a £5 billion budget by adopting treatments A-E; if treatment F is perfectly divisible, we could also gain a small amount of additional health by spending the remaining £50,000,000 on F. If our league table includes all possible healthcare interventions, we can also identify the shadow price of a QALY, which is equal to the ICER of the last treatment adopted (which would also be the first to be replaced by any new treatment): in this case £19,000/QALY. If we already have an accurate estimate of the shadow price of a QALY, we can use this ceiling ratio to make independent decisions on subsequent treatments, by simply adopting all interventions with ICERs below the ceiling ratio.[23]

[21] Integer and linear programming methods, which avoid assuming divisibility and constant returns to scale, have also been proposed (178-181); these make similar distinctions between independent and mutually-exclusive treatments (178), but are not discussed in detail in this thesis.

[22] League tables maximise health benefits from the budget because we adopt those treatments that generate the maximum health per pound spent (i.e. the minimum cost per unit of health gained). It is generally stated that health benefits are maximised only if the last treatment adopted is divisible and has constant returns to scale. Although these assumptions are not realistic, this is unlikely to matter for two reasons. Firstly, within a complete league table of all healthcare interventions, the cost of each individual intervention will be a very small fraction of the total budget: particularly if the cost-effectiveness of individual treatments is stratified, with the ICER for treatment in different patient groups appearing separately in the league table. Secondly, healthcare systems typically have a soft budget, such that any money left over from one year can be spent in the next (183).

[23] Our ceiling ratio or shadow price of a QALY is equal to the opportunity cost of spending money on a new treatment. The opportunity cost of adopting a new treatment is disinvestment in the least cost-effective treatment that we currently use, which involves losing that number of QALYs per pound spent. We can therefore get more QALYs from the same budget by adopting any treatment that has an ICER lower than the shadow price of a QALY and disinvesting in the least cost-effective treatment

Table 3.6. Illustration of treatment adoption decision rules for independent alternatives

Treatment	ICER (per QALY gained)†	Total cost for population‡	Budget spent‡	Adopt?*
A	£2,950	£500,000	£500,000	Adopt
B	£7,300	£2,000,000	£2,500,000	Adopt
C	£12,000	£750,000	£3,250,000	Adopt
D	£15,000	£1,100,000	£4,350,000	Adopt
E	£19,000	£600,000	£4,950,000	Adopt
F	£25,000	£500,000	£5,450,000	Partially?
G	£30,500	£90,000	£5,540,000	No
H	£50,000	£3,000,000	£8,540,000	No
I	£100,000	£190,000	£8,730,000	No

† Relative to the next best non-dominated alternative.
‡ In thousands.
* At a £20,000/QALY ceiling ratio, or given a £5 billion budget.

Conversely, only one treatment from a set of *mutually-exclusive* alternatives will be adopted at a time (a "competing-choice" problem (171, 172, 174)). Here, we need to calculate the ICER for each option compared with its next best non-dominated alternative (16, 125, 128, 169-177, 184). This requires us to identify and exclude from consideration treatments that are strongly or weakly dominated by others, to identify the efficiency or cost-effectiveness frontier (16, 172) or expansion path (185). Alternatively, if a ceiling ratio has already been established, we can simply adopt the most effective treatment that lies on the cost-effectiveness frontier *and* has an ICER below our ceiling ratio (128, 170, 171). This method is simplified further using NMB, which enables us to simply calculate the NMB of all alternatives at the appropriate ceiling ratio and adopt the treatment with highest NMB (16, 176).[24]

For example, of the mutually-exclusive alternatives in Table 3.7, 20 mg Drug A is extendedly dominated (and therefore does not lie on the frontier), while 40 mg Drug A has highest NMB and generates more QALYs than any other non-dominated alternative with an ICER below the £20,000/QALY ceiling ratio.

---

that we are currently doing. Adopting a treatment that has an ICER lower than our ceiling ratio is equivalent to adopting any treatment that has a higher NMB than its comparator (since $INB = \Delta QALY \cdot R_c - \Delta Cost = NMB_A - NMB_B = (QALY_A \cdot R_c - Cost_A) - (QALY_B \cdot R_c - Cost_B)$).

[24] Just as we maximise the health gained from the budget by adopting the treatment with higher expected NMB when we have two mutually-exclusive alternatives, we can do the same for multiple mutually-exclusive alternatives by adopting the treatment with highest NMB. By multiplying QALYs by their shadow price and subtracting costs, NMB give a measure of the relative value of different treatments that takes account of the opportunity cost of spending on new treatments. In practice, risk-averse decision-makers may consider the uncertainty around ICERs (186), although I take the view in this thesis that inference is irrelevant to treatment adoption decisions (157).

Table 3.7. Illustration of calculation of ICERs and treatment adoption decision rules for mutually-exclusive alternatives

Treatment	Cost	QALYs	Total NMB*	Δ cost†	Δ QALYs†	ICER (per QALY gained)†	Adopt?*
No treatment	£1,000	2.00	£39,000	-	-	-	No
10 mg Drug A	£1,500	2.20	£42,500	£500	0.20	£2,500	No
20 mg Drug A	£1,700	2.25	£43,300	£200	0.05	Extendedly dominated by 10 and 40 mg	No
40 mg Drug A	£1,800	2.30	£44,200	£300	0.10	£3,000	Adopt
80 mg Drug A	£3,600	2.34	£43,200	£1,800	0.04	£45,000	No

* At a £20,000/QALY ceiling ratio
† Relative to the next best non-dominated alternative

We can then include the ICER for each of the non-dominated treatment options in our league table (125, 128, 169-173, 175-177) (Table 3.8). However, since the different doses are mutually-exclusive, the table presents the incremental cost-effectiveness of each option compared with the next best non-dominated alternative: for example, besides the least expensive (10 mg A), our league table would include ICERs for substituting 40 mg in place of 10 mg and substituting 80 mg for 40 mg.

Table 3.8. Illustration of treatment adoption decision rules for independent alternatives, with the addition of one cluster of three mutually-exclusive treatments from Table 3.7.

Treatment	ICER (per QALY gained)†	Total cost for population‡	Budget spent‡	Adopt?*
10 mg Drug A (vs. no treatment)	£2,500	£15,000	£15,000	Adopt
A	£2,950	£500,000	£515,000	Adopt
40 mg Drug A (vs. 10 mg)	£3,000	£27,000	£542,000	Adopt
B	£7,300	£2,000,000	£2,542,000	Adopt
C	£12,000	£750,000	£3,292,000	Adopt
D	£15,000	£1,100,000	£4,392,000	Adopt
E	£19,000	£600,000	£4,992,000	Adopt
F	£25,000	£500,000	£5,492,000	Partially?
G	£30,500	£90,000	£5,582,000	No
80 mg Drug A (vs. 40 mg)	£45,000	£54,000	£5,636,000	No
H	£50,000	£3,000,000	£8,636,000	No
I	£100,000	£190,000	£8,826,000	No

† Relative to the next best non-dominated alternative
‡ In thousands
* At a £20,000/QALY ceiling ratio, or given a £5 billion budget

I conducted a literature review of how textbooks and reviews define "mutually-exclusive" and "independent" when distinguishing between these decision rules (Section 9.2). Although some definitions define mutually-exclusive treatments as those used to treat the same

patient group (16, 125, 129, 169, 170, 187, 188),[25] nearly all definitions make specific reference to non-additive effects as the defining feature that determines whether treatments are considered to be mutually-exclusive or independent (125, 128, 129, 169, 170, 172, 185, 187, 189, 190) (Section 9.2). For example, Johannesson (170)[26] states that:

> "Two programmes, A and B, are mutually-exclusive if implementing programme A (B) means that programme B (A) cannot also be implemented. If the programmes A and B can both be implemented physically, but implementing A (B) means that the costs and effectiveness of B (A) change, then the programmes should be defined as mutually-exclusive. In this case, we have three alternatives: carrying out only A, carrying out only B, or carrying out both A and B. Mutually-exclusive programmes can be viewed as programmes for the same population, e.g. two alternative drugs for ulcer patients."

Johannesson's definition, and others, suggests that whenever treatments interact, we should treat the combinations of treatments as mutually-exclusive alternatives, comparing the combinations incrementally and selecting the single strategy that maximises NMB for that population. Conversely, this implies that if there is no interaction, we should treat the interventions as independent alternatives and compare their ICERs against the ceiling ratio separately, adopting those treatments with ICERs below our ceiling ratio.

For factorial trials, the treatments under investigation are necessarily being evaluated and considered for use in the same patient group and interactions are highly likely for economic endpoints. This would imply that economic evaluations of factorial trials should aim to identify which of the mutually-exclusive factor-combinations (e.g. *0*, *a*, *b* and *ab*) is most cost-effective, rather than making independent decisions about each factor in isolation.

Furthermore, as well as making different assumptions about interactions, the different methods for analysing factorial trials lend themselves to different decision rules for economic

[25] Although definitions based on patient population appear to contradict the definitions based on non-additive effects, the two are largely equivalent because two treatments given to the same patient population are more likely to show non-additive effects on costs and QALYs for the reasons discussed in the previous section. Since interactions are substantially more likely between treatments given to the same patient group than treatments given to different populations, patient group has proven a useful rule of thumb to identify treatments that should be considered mutually-exclusive.
[26] Quote taken from page 135.

evaluation. At-the-margins analysis implicitly treats the factors as independent options, calculating separate ICERs for A vs. not-A and for B vs. not-B that are independently compared against the ceiling ratio and used to make separate decisions on the two treatments. However, even if decisions are made independently on each factor, the decision to adopt both A and B in the same patients implicitly means that *ab* is recommended. By contrast, inside-the-table analysis treats the four cells as mutually-exclusive options, estimating costs, QALYs and NMB for each of the four options and allowing us to identify the cost-effectiveness frontier and select the option that maximises NMB at our ceiling ratio.

Within the example shown in Table 3.5, if we considered the results inside-the-table and made a joint decision between the four cells, we would adopt the treatment that has highest expected NMB at the ceiling ratio of interest (£30,000/QALY): in this case *b*, which has an expected NMB of £479,504 per patient. Alternatively, if we conducted at-the-margins analysis and made independent decisions about A and B, we would estimate that the incremental net [monetary] benefit (INB) for A vs. not-A was £2,313 (0.5*(£470,155+£468,985-£455,010-£479,504)) and estimate that the INB for B vs. not-B was £11,662 (0.5*(£479,504+£468,985-£455,010-£470,155)). We would therefore adopt both A and B for this patient group. Since adopting both treatments simultaneously implies adoption of *ab*, the expected NMB is £468,985. Within this example, interpreting the economic results at-the-margin and making independent decisions about A and B gives misleading information and would lead to us adopting a treatment-combination that has a NMB £10,519 per patient lower than that achieved by allowing for the interaction and making a joint decision on the two treatments.

Indeed, as shown in Appendix E, treating the cells of the factorial trial as mutually-exclusive alternatives and considering interactions between factors will always enable us to maximise the health gains from the budget. Conversely, for two treatments with perfectly additive effects on both costs and QALYs, inside-the-table analysis informing a joint decision between

mutually-exclusive alternatives and at-the-margins analysis informing independent decisions would give identical results and both approaches would maximise expected NMB. As discussed above, we can maximise the health gained from the budget given the information currently available by adopting the option with highest expected NMB, ignoring statistical inference (157). Analyses allowing for interactions give unbiased estimates of the difference between *0* and *a* (3, 4) and therefore unbiased estimates of the expected NMB of each treatment. By implication, this means that allowing for interactions and considering the four cells as mutually-exclusive alternatives must give the correct ranking of cells by NMB and therefore correctly identify the treatment with highest NMB.

By contrast, excluding interactions will introduce bias into NMB estimates whenever treatment effects are not genuinely additive (3, 4, 8, 20, 30). While non-qualitative interactions in NMB will change the magnitude of differences between treatments, qualitative interactions will change the ranking of treatments. Furthermore, qualitative interactions that change which treatment has highest expected NMB will change the conclusions of our economic evaluation and ignoring such interactions will cause us to adopt a treatment that does not actually have highest NMB. If there are no qualitative interactions for NMB, ignoring interactions will give the same conclusions as allowing for interactions and making a joint decision between the four mutually-exclusive alternatives. Crucially, at-the-margins analysis and making independent decisions on interacting factors can lead us to calculate INB and ICERs based on differences between the wrong pair of treatments.

When there is evidence that interactions are negligible, we may choose to omit some or all interaction terms when analysing trial outcomes to maximise power and avoid the risk that a small, chance interaction drives the conclusions. However, it may be appropriate to draw conclusions inside-the-table and make a joint treatment-adoption decision that treats the cells as mutually-exclusive alternatives even when interactions are not taken into account in the analysis. Firstly, drawing conclusions inside-the-table does not introduce any bias

regardless of the presence or size of interactions. Secondly, presenting the costs and QALYs for the full set of mutually-exclusive alternatives and identifying the cost-effectiveness frontier reminds decision-makers of the importance of making a joint decision. In particular, presenting *0*, *a*, *b* and *ab* as mutually-exclusive alternatives forces decision-makers to consider whether *a* is better than *b* or whether *ab* is appropriate, whereas considering A and B separately may lead to recommendations for both treatments without explicit guidance of whether they should be used together.

3.6 Challenges facing analysis of economic endpoints from factorial trials

3.6.1 Difficulties associated with transforming costs and QALYs

Since interactions depend on the scale of measurement, transformation is often used to eliminate or reduce interactions in clinical endpoints of factorial trials (2, 6, 10, 30, 31, 35, 36). For example, if treatment effects are multiplicative (i.e. interaction=0 on a logarithmic scale), logistic regression or log-transformation are often used to analyse results and draw statistical inferences. Transformation can also give data symmetrical, Gaussian distributions (191) and reduce heteroskedasticity (156). However, analysis of transformed data raises additional problems for economic evaluation that are not raised by analysis of clinical endpoints:

1. The correct transformation may not be known (or knowable) (17).

2. Frequently, many trial participants have zero cost (17, 49, 191). However, box-cox transformations (including logarithmic) can only be done on non-zero values (17, 191); arbitrary constants are often added to all zero values, although the magnitude of such constants can affect results (17).

3. No transformation will eliminate qualitative interactions (6, 35), which occur commonly for NMB (Section 3.4). The main reason for this is that qualitative interactions change the direction, not just the magnitude, of treatment effects: e.g. μ_{ab}-μ_b may be positive while μ_a-μ_0 is negative. Monotonic transformations (e.g. taking the logarithm, square-root or power of all values) can increase or decrease μ_{ab}-μ_b and μ_a-μ_0 but will not cause them to have the same sign. In principle, non-monotonic transformations (e.g. trigonometric functions) could be used, but these are unlikely to reflect realistic data-generating mechanisms and would only eliminate qualitative interactions in certain very specific circumstances.

4. Total costs, total QALYs and NMB are the weighted sum of several components (e.g. the cost of drugs, clinical events and side-effects (192)) and treatments may have additive effects on some components and multiplicative effects on others. For example, total cost may be made up of additive drug costs and multiplicative event costs. Analysing total costs on a natural scale will therefore give an interaction arising from the multiplicative effect of events, while analysing total costs on a logarithmic scale will give an interaction arising from additive drug costs. In some cases it may be possible to separate the components that are multiplicative from those that are additive (e.g. to analyse log-transformed event cost with one model and non-transformed drug costs with another), as explored in Section 3.6.2.5 and Chapter 6. However, in many cases, components amenable to different types of transformation may not be separable: e.g. treatments may have both additive and multiplicative effects on HRQoL.

5. Resource allocation decisions must be based on costs and QALYs measured on a natural scale (49). The principle of maximising health gains from a finite budget using a league table or ceiling ratio necessarily means that decisions between competing treatments must be made on an additive basis using untransformed costs and QALYs. As

described in Section 3.5, health gains from a finite healthcare budget would be maximised by adopting those treatments with ICERs below a ceiling ratio that represents the opportunity cost of the option that would be displaced by the new treatment. Setting the ceiling ratio using a league table (16, 125) necessarily requires *adding* the total cost of each treatment to the total already spent. The ICER must therefore be calculated as the *absolute* difference in cost divided by the *absolute* difference in effect. By contrast, decision-making on a log-scale, based on a measure of relative difference in cost divided by relative difference in effect would fail to meet this objective. If analyses are conducted on a non-natural scale, it is therefore necessary to back-transform results so that conclusions can be based on inside-the-table means on a natural scale.

6. Transforming costs and QALYs prior to analysis means that coefficients, statistical inference and predictions are on the transformed scale and are not applicable to the natural scale (17, 49, 162). Furthermore, estimates based on data subjected to non-linear transformations cannot be returned to a natural scale by simply inverting the transformation (191). Instead, more complex back-transformation methods are needed, such as smearing estimators (17, 49) or Taylor series approximation (49), which both require assumptions about the distribution or heteroskedasticity (17, 49). These issues can be avoided using generalised linear models (GLM), which predict a function of the dependent variable (17, 49, 162, 191): for example GLM with log-link could estimate linear effects of A and B on the natural logarithm of mean cost (in contrast to modelling the effect of each treatment on the mean of log-cost) (17). As result, GLM enables analysis of zero values (17), inferences that are valid on a natural scale (17) and simple back-transformation of the output (17, 49). GLM models can also deal with heteroskedasticity and non-Gaussian residuals by separately specifying a distributional “family” that allows the variance to be an increasing function of the mean (17, 162, 191). However, GLM models are biased if the link function is mis-specified (191) and there is no simple way to identify the appropriate link function (17, 162).

Although the issues surrounding the appropriateness of transformation are not unique to factorial trials, this study design does raise particular problems in relation to transformation: for most studies, transformations are used primarily to overcome distributional problems, whereas in factorial studies, suitable transformations may remove interactions and ensure that treatment effects are additive.

3.6.2 Issues raised by regression analysis of factorial trials

Regression analysis is generally the most convenient way to correctly analyse factorial trials and estimate the magnitude and statistical significance of interaction terms: particularly for unbalanced designs (20). Regression also facilitates consideration of covariates, which increases statistical power (17, 49, 51) and may help ensure Gaussian residuals (49). In the context of economic evaluation, regression also facilitates exploration of heterogeneity (15, 17, 48, 49, 193), allowance for between-centre effects/clustering (15, 17, 49, 193-195) and overcoming the bias associated with between-group imbalance in baseline utility (49, 150). However, economic evaluation raises several challenges for regression analysis, including the statistical properties of the ICER, the distribution of costs and QALYs, correlations between variables and characterising uncertainty in a way useful for decision-makers. These issues have been discussed at length previously (16, 17, 49, 191, 196, 197); this chapter therefore draws on these secondary sources, rather than replicating the previous reviews.

3.6.2.1 General challenges facing analysis of costs and QALYs

Costs are generally positively skewed (16, 17, 49, 125, 156, 162, 191) with excess zeros (16, 17, 49, 191) and a long, heavy right tail (16, 17) and are frequently heteroskedastic (156, 162). Survival data (e.g. life-years), are also positively skewed with a heavy right tail (198),

while EQ-5D utilities can take only certain values and are frequently negatively skewed with many values equal to one and other peaks centred at 0.2 and 0.7 (199, 200). The properties of these distributions can, in principle, be allowed for using GLM (17, 49, 50, 191, 200), two-part models (17, 49, 191, 200) and/or adjusted limited dependent variable mixture models (199).

Since economic evaluation aims to inform resource allocation decisions, it is also necessary to estimate the uncertainty around the results in ways that focus on the decision problem, such as cost-effectiveness acceptability curves (CEACs) and value of information analysis (16, 18).

CEACs plot the probability that each treatment is cost-effective (i.e. has highest NMB) at a range of ceiling ratios (201, 202). For comparisons between two treatments, CEACs generally comprise one line showing the probability that treatment is cost-effective compared with its comparator (16, 201, 202); this can be interpreted (and/or calculated) as the one-sided probability that treatment has positive INB (16, 17, 48). However, for decisions between >2 mutually-exclusive alternatives, CEACs show multiple lines plotting the probability that each treatment is the most cost-effective (i.e. has highest NMB) (202). Such CEACs also show a frontier giving the probability that the treatment with highest *expected* NMB is cost-effective; this equals one minus the probability of making the wrong decision (202).[27]

In analyses considering just two treatments, CEACs can be calculated: from one-sided p-values (48); directly from the mean and variance for INB (assuming a normal distribution) (49, 51, 197); or from an empirical distribution of costs and QALYs estimated using

[27] Showing the frontier is necessary as the treatment with the highest probability of having NMB at any given ceiling ratio will not necessarily have highest expected NMB when there are >2 alternatives.

bootstrapping, probabilistic sensitivity analysis or Bayesian Markov chain Monte Carlo (MCMC) methods (16, 49, 53, 116).

Value of information analysis estimates the value of collecting additional information to reduce uncertainty and quantifies the consequences, as well as the probability, of making the wrong decision (13, 49, 157). The expected value of perfect information (EVPI) and the expected value of sample information (EVSI)[28] can be estimated from the mean INB and its SE by assuming a normal distribution (13, 203, 204), or using decision-analytical modelling (13).

However, methods to estimate CEACs or value of information from *regression-based* cost-effectiveness on >2 alternatives using methods other than MCMC (116) are less well developed. In Section 3.6.4 and Chapters 6-7, I develop methods for estimating these measures using bootstrapping.

3.6.2.2 Challenges posed by ratio statistics

Although ICERs comprise the most commonly-used measure of cost-effectiveness, several of their properties make them unsuitable for statistical analyses, such as regression (48, 51-53, 205). Firstly, the ratio of two random variables follows no known distribution and has undefined variance (48, 52, 53, 196, 205). Secondly, the ICER will approach infinity (or negative infinity) as the difference in effectiveness approaches zero (51, 53), which means that the mean of ICERs across samples (e.g. bootstrap replicates) is biased (52) and that covariates are unlikely to have linear effects on the ICER. Thirdly, the ICER has opposite interpretations depending on whether the difference in effectiveness is positive or negative (Figure 3.1), with low ICERs indicating good value for money if treatment is more effective

[28] The EVPI represents the NMB gained from eliminating all uncertainty, while EVSI represents the NMB gained from reducing uncertainty by conducting a trial of a specific size and cost (13, 157).

than its comparator and poor value for money if treatment is less effective (16, 48, 51, 52, 196). The decision rule for the ICER is therefore:

$$\begin{aligned} &AdoptAIf : ICER < Rc \;\&\; QALYs_A - QALYs_B > 0 \\ &OrIf : ICER > Rc \;\&\; QALYs_A - QALYs_B < 0 \end{aligned} \quad (3.3)$$

Finally, ICERs are always estimated relative to a single, specified comparator and can only inform decision-making if they are compared against the next best non-dominated alternative (16). Furthermore, ICERs are not transitive, in that ICERs between one pair of treatments cannot be converted into ICERs between another pair (52): for example, the ICER for A vs. B cannot be estimated from the ICER for A vs. comparator and that from B vs. comparator.

$$ICER_{AvsB} = \frac{Cost_A - Cost_B}{QALYs_A - QALYs_B} \neq \frac{Cost_A - Cost_C}{QALYs_A - QALYs_C} - \frac{Cost_B - Cost_C}{QALYs_B - QALYs_C} \quad (3.4)$$

As a result, ICERs are particularly unsuitable for regression analyses evaluating the effect of >2 treatments.

3.6.2.3 NMB

As described above, cost-effectiveness is often expressed as NMB ($NMB = QALYs \bullet Rc - Cost$) (131) or NHB (52) due to the statistical properties of ICERs. Unlike ICERs, INB simply equals the difference in total NMB (52) and treatments with larger NMB are unambiguously better than those with small values (16, 48, 51, 52).

$$\begin{aligned} &INB_{AvsB} = NMB_A - NMB_B = (QALYs_A \bullet Rc - Cost_A) - (QALYs_B \bullet Rc - Cost_B) \\ &AdoptAIf : INB_{AvsB} > 0 \end{aligned} \quad (3.5)$$

NMB can therefore be readily used as the dependent variable in regression analyses (48). In particular, NMB has defined SEs (16, 48, 52, 53) and the sampling distribution for INB asymptotes towards normal (16, 49, 51, 52).

However, NMB depends on the ceiling ratio. Since the correct ceiling ratio is unknown, researchers generally present results at multiple ceiling ratios (16, 48, 49, 52, 53, 206). However, conducting regression analyses separately at each ceiling ratio is problematic as the distribution of NMB, the covariates that affect it and the appropriate scale of analysis may vary with ceiling ratio (48, 206). This is of particular importance for factorial trials, since the importance of interactions and the appropriate scale of analysis may differ between costs and QALYs and therefore with ceiling ratio: for example, if there is a genuine interaction for QALYs, but treatments have additive effects on costs, omitting the interaction term at Rc=£0 (which is equivalent to a regression on cost) may be appropriate, but interactions will become increasingly important at higher ceiling ratios. Furthermore, if treatments have multiplicative effects on QALYs, but additive effects on cost, the appropriate functional form could be OLS at Rc=£0, but GLM with log-link at an infinite ceiling ratio (which is equivalent to a regression on QALYs).

3.6.2.4 Bivariate regression

Conducting one regression analysis on costs and another on health outcomes avoids the need to replicate models at multiple ceiling ratios and enables model specification to differ between costs and effects (49, 197, 206). However, it is essential to allow for correlations between costs and effects (193, 196, 197): for example, healthier patients may require fewer healthcare resources, while patients dying early in the trial will accrue costs for a shorter time period, as well as accruing fewer QALYs or life-years.

Several methods can allow for correlations between costs and effects, which are illustrated using the worked example in Section 3.6.4. Seemingly-unrelated regression enables two separate regression analyses to be run on the same or overlapping datasets and adjusts SEs to allow for correlations between the two models (49, 197, 206-208). It is more efficient than running separate OLS models when covariates differ between models (49, 197). In

Stata, seemingly-unrelated regression can be done using two commands: `suest` (208) and `sureg` (207). `sureg` implements Zellner's seemingly-unrelated regression using generalised least squares, whereas `suest` uses a modification of the sandwich estimator to estimate the (co)variance of both models simultaneously and allows for regression models of any specification.

Two or more models can also be estimated simultaneously using bootstrapping or MCMC, which both approximate an empirical sampling distribution using repeated iterations. Bootstrapping is commonly used within economic evaluation to allow for the non-Gaussian distribution of costs and QALYs (16). It allows for correlations between costs and QALYs by re-sampling patients with replacement and running analyses on the costs and QALYs for the *same* set of patients within each bootstrap replicate and repeating such analyses thousands of times (16, 196, 205). This approach is extremely flexible and can be used with any analytical methods; in particular, it facilitates subsequent analyses, such as scatter graphs, value of information, CEACs and allowing for censoring or missing data (16). MCMC can also allow for correlations by estimating several parametric models that are all estimated on the same set of iterations (193). Both bootstrapping and MCMC can be adapted to explicitly introduce correlations between costs and QALYs by adding the residual from the first model as an explanatory variable in the second model (193).

3.6.2.5 Multivariate regression

Bivariate regression on costs and QALYs can easily be extended to consider multiple components of NMB (206). For example, costs and QALYs accrued in different time periods can be estimated separately, allowing for correlations between different years and variables. The resulting predictions of costs and QALYs in each time period can be used for interval-based methods adjusting for censoring, such as Kaplan-Meier sample averaging (KMSA)

(16). For factorial trials, it may also be appropriate to analyse different types of cost in different models: for example, interactions may be important for AE costs, but not drug costs, or (alternatively) drug costs may be best modelled on a logarithmic scale, while other costs are additive. I develop and evaluate this type of model in Chapter 6.

Several studies have also taken this principle further, by developing an explicit structure for how treatment(s) and covariates are assumed to affect costs and QALYs (49, 149), which is often described as event-based CEA. For example, if costs and QALYs are assumed to be driven largely by clinical events, trial data can be analysed by first estimating a regression model predicting the odds or hazard of clinical events conditional on treatment indicators and covariates (49). Further models can then estimate the effect of clinical events and covariates on non-drug costs and clinical events; such models may assume that treatment has no effect on costs and QALYs after events. Similar approaches have been used to explicitly allow for multiplicative effects on clinical events and additive effects on drug costs and side-effects (149, 209), facilitate subgroup analyses (149) and extrapolate results (210, 211). Regression results can also be incorporated into Markov models (149) and/or be combined with external data (209). Interestingly, two of the main studies using event-based CEA were factorial (149, 210, 211), but focused on a single factor, without explicitly allowing for the other factor(s) or interactions.

My approach of simply analysing different identifiable types of cost using different models imposes no additional assumptions, but does not facilitate extrapolation. By contrast, event-based CEA naturally extrapolates results beyond the trial time horizon, but relies heavily on additional assumptions about how treatment affects costs and health effects and will only give valid, unbiased results if the model is accurate (49, 206).

3.6.3 Bayesian analysis

As discussed previously, Bayesian analyses with informative priors provide a compromise between including or excluding interactions (42, 115, 116). Although this has rarely been done for clinical endpoints, it shows particular promise for economic evaluation, which is frequently done in a Bayesian framework, or with Bayesian interpretations,[29] and where consideration of external evidence alongside the current trial is recommended (134). Bayesian methods using informative priors not only enable us to down-weight unexpected interactions without excluding them, but also enable inclusion of all prior evidence.

Such methods can be implemented using MCMC (which Welton et al (115, 116) used to consider informative priors on interactions). However, Bayesian bootstrapping (213, 214) provides an alternative that makes fewer assumptions about the distribution of the trial data. Here, greater weight is attached to bootstraps that closely reflect the prior distribution than those that differ greatly from the prior. Such weights can be applied by taking a weighted (rather than crude) average across bootstrap replicates (213, 214), or by "vetting" bootstrap replicates as they are generated, such that bootstraps that closely resemble the prior are more likely to be included in the analysis (213). The vetted bootstrap is particularly convenient for economic evaluation, since it outputs bootstrap replicates that can be given equal weight when calculating CEACs or plotting scatter graphs.

[29] The Bayesian framework can directly focus on informing specific real-world decisions, which are the purpose of economic evaluation (212). In particular, core questions within economic evaluation, such as 'what is the probability that this specific treatment is cost-effective?' are meaningless within a frequentist paradigm (212). Several of the key health economic methods, such as probabilistic sensitivity analysis, multiple imputation and value of information analysis, are inherently Bayesian and CEACs are generally given a Bayesian interpretation (e.g. the "probability that treatment is cost-effective", rather than as a one-sided p-value) (212).

3.6.4 Worked example

The simulated trial data generated in Section 3.4.1 were analysed in Stata versions 11 and 12 (StataCorp LP, College Station, Texas) using at-the-margins, inside-the-table and regression analysis.

At-the-margins analysis suggested that both A and B are cost-effective when they are considered independently: prescribing A (with/without B) costs £25,530/QALY gained vs. not-A, while B costs £20,189/QALY gained vs. not-B (Table 3.9). Based on a £30,000/QALY ceiling ratio, we might therefore conclude that both A and B are cost-effective treatments and have positive INB at a £30,000/QALY ceiling ratio and that we should adopt both A and B (i.e. use *ab*). However, this inference would be incorrect due to the qualitative interaction for NMB and the bias inherent within at-the-margins analysis.

Table 3.9. Worked example of economic evaluation of 2x2 factorial trial: at-the-margins approach

	Treatment A		**Treatment B**	
	Placebo (n=500)	**Active drug A (n=500)**	**Placebo (n=500)**	**Active drug B (n=500)**
Mean cost per patient (SD)	£98,456 (£33,005)*	£111,669 (£33,917)*	£93,064 (£33,348)*	£117,062 (£30,439)*
Difference in cost (SE)	£13,213 (£2,116)*		£23,998 (£2,019)*	
Mean QALYs per patient (SD)	18.9 (6.4)*	19.4 (5.8)*	18.5 (6.3)*	19.7 (5.8)*
Difference in QALYs (SE)	0.52 (0.38)		1.19 (0.38)*	
Mean total NMB per patient at Rc=£30,000 (SD)	£467,257 (£200,813)*	£469,570 (£180,480)*	£462,582 (£199,470)*	£474,245 (£181,783)*
INB at Rc=£30,000 (SE)	£2,313 (£12,075)		£11,662 (£12,069)	
Incremental cost/QALY	£25,530		£20,189	

* Significantly greater than zero ($p<0.05$).

Costs and QALYs were then analysed using seemingly-unrelated regression using the `suest` post-estimation command (208). OLS was used for simplicity although costs, QALYs and NMB were positively skewed ($p \leq 0.001$) and variances were lower for patients receiving B; the effect of using GLM is explored below.

Seemingly-unrelated OLS without interaction terms replicated the mean at-the-margins treatment effects (Table 3.10), although SEs differed. However, prediction of group means

and their SEs (Table 3.10, Figure 3.2A) is easier after regression than at-the-margins analysis. These predictions can be used to consider the cost-effectiveness of the four cells as mutually-exclusive options, draw conclusions about dominance and identify which of the four options maximises NMB, conditional on the model used. In this example, the predicted values from regression suggest that *a* is extendedly dominated by a combination of *b* and *0* (Figure 3.2A) and explicitly predict that *ab* is the optimal treatment at a £30,000/QALY ceiling ratio, costing £25,530/QALY gained compared with *b*. However, these predictions rely on the same assumptions as at-the-margins analysis and will be biased if interactions are present.

Table 3.10. Worked example of economic evaluation of 2x2 factorial trial: OLS regression without interaction term

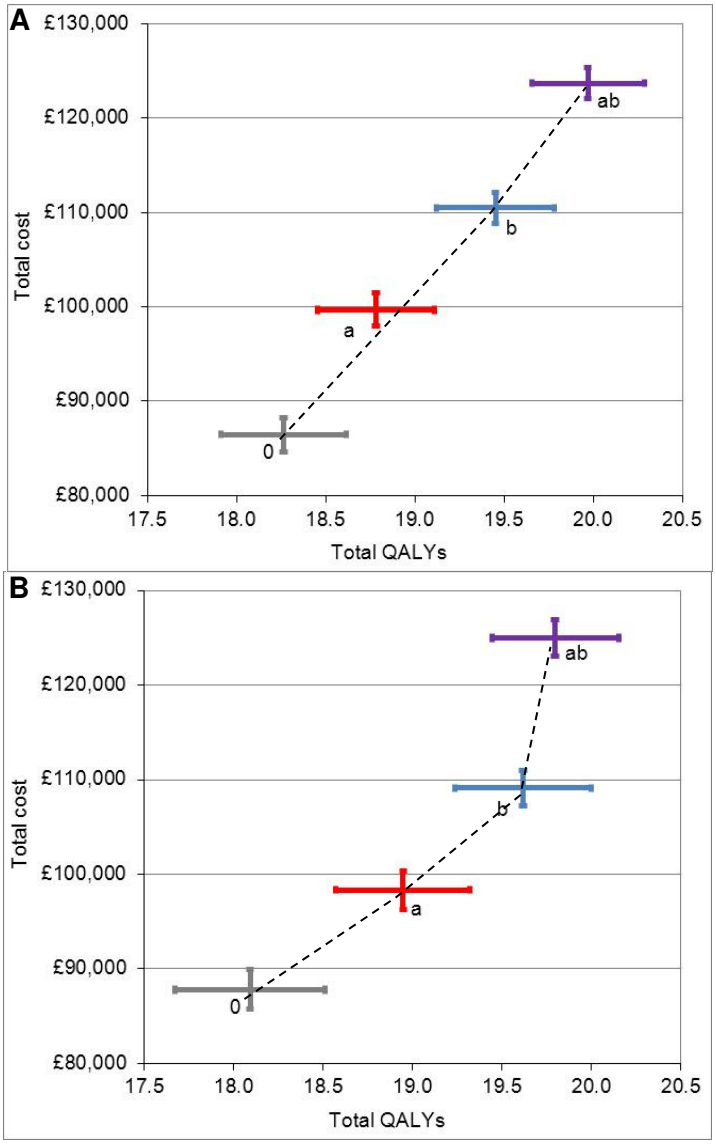
		Total cost/ patient†	Total QALYs/ patient†	NMB/ patient‡	Cost/QALY		
					vs. *0*	vs. *a*	vs. *b*
Treatment effect for A (SE)		£13,213 (£1,974)*	0.52 (0.38)	£2,313 (£12,075)	-	-	-
Treatment effect for B (SE)		£23,998 (£1,974)*	1.19 (0.38)*	£11,662 (£12,075)	-	-	-
Constant term (SE)		£86,457 (£1,794)*	18.26 (0.35)*	£461,426 (£10,457)*	-	-	-
Predicted mean outcome (SE)	*0*: Placebo for A + placebo for B (n=250)	£86,457 (£1,794)*	18.3 (0.35)*	£461,426 (£10,457)*	-	-	-
	a: Drug A + placebo for B (n=250)	£99,670 (£1,753)*	18.8 (0.33)*	£463,739 (£10,457)*	£25,530	-	-
	b: Drug B + placebo for A (n=250)	£110,456 (£1,624)*	19.5 (0.33)*	£473,088 (£10,457)*	£20,189	£16,070	-
	ab: Drug A + Drug B (n=250)	£123,669 (£1,664)*	20.0 (0.31)*	£475,402 (£10,457)*	£21,809	£20,189	£25,530

* Significantly greater than zero (p<0.05).
† Results are based on seemingly-unrelated OLS, which estimates robust SEs.
‡ Based on a ceiling ratio of £30,000/QALY.

However, the predictions from regression models are valid only if the functional form is correct. In this example, we observe large interactions for costs (p=0.172) and QALYs (p=0.377) and a qualitative interaction for NMB (p=0.288, Table 3.11). None of the interaction effects were statistically significant; given that most two-arm trials are not powered to detect significant differences in mean costs or NMB (18, 139) and the variance around the

interaction term is four-fold higher than that for main effects (3, 37), this finding is likely to be common among economic evaluations of factorial trials. However, these interactions nonetheless introduce substantial bias into at-the-margins estimates and those from regression without an interaction term.

Figure 3.2. Cost-effectiveness plane showing predicted costs and outcomes from regression analyses on the same simulated data. **A**: Regression without interaction term. **B**: Regression with interaction term. Error bars show SEs from regression analysis. The dashed black line represents the cost-effectiveness frontier.



Including an interaction term completely changes the conclusions for this example. When we consider each of the four cells as mutually-exclusive alternatives using seemingly-unrelated OLS regression with an interaction term (Table 3.11) or inside-the-table analysis (Table

3.12), we see that although *a* is cost-effective vs. *0* (costing £12,297/QALY gained) and *b* is cost-effective vs. *a* (costing £16,070/QALY gained), *ab* costs £88,573/QALY gained vs. *b* (Figure 3.2B). On that basis, *b* would maximise NMB at a £30,000/QALY ceiling ratio (not *ab* as at-the-margins analysis suggested). As discussed above, the correct conclusions are those from analyses allowing for interactions, which avoid bias from omitted interaction terms.

Table 3.11. Worked example of economic evaluation of 2x2 factorial design: OLS regression with an interaction term

		Total cost/ patient†	**Total QALYs/ patient†**	**NMB/ patient‡**	**Cost/QALY**		
					vs. *0*	**vs. *a***	**vs. *b***
Treatment effect for A (SE)		£10,520 (£2,944)*	0.86 (0.56)	£15,145 (£17,076)	-	-	-
Treatment effect for B (SE)		£21,305 (£2,792)*	1.53 (0.57)*	£24,494 (£17,076)	-	-	-
Interaction (SE)		£5,386 (£3,945)	-0.68 (0.77)	£-25,664 (£24,149)	-	-	-
Constant term (SE)		£87,804 (£2,116)*	18.09 (0.42)*	£455,010 (£12,074)*	-	-	-
Predicted mean outcome (SE)	*0*: Placebo for A + placebo for B (n=250)	£87,804 (£2,116)*	18.1 (0.42)*	£455,010 (£12,074)*	-	-	-
	a: Drug A + placebo for B (n=250)	£98,324 (£2,047)*	18.9 (0.37)*	£470,155 (£12,074)*	£12,297	-	-
	b: Drug B + placebo for A (n=250)	£109,109 (£1,822)*	19.6 (0.38)*	£479,504 (£12,074)*	£13,956	£16,070	-
	ab: Drug A + Drug B (n=250)	£125,015 (£1,892)*	19.8 (0.35)*	£468,985 (£12,074)*	£21,809	£31,375	£88,573

* Significantly greater than zero ($p<0.05$).
† Results are based on seemingly-unrelated regression, which estimates robust SEs.
‡ Based on a ceiling ratio of £30,000/QALY.

The different analytical methods may also produce conflicting conclusions in different ways from this example. In particular, qualitative interactions for costs and/or QALYs could cause *a* or *b* to dominate *ab* even if at-the-margins analysis finds both A and B to be cost-effective. At-the-margins analysis could also find A to cost <£30,000/QALY when inside-the-table analysis shows that the strategy maximising NMB at this ceiling ratio is to give *0*; conversely, *a* could maximise NMB without A being cost-effective in at-the-margins analyses.

Table 3.12. Worked example of economic evaluation of 2x2 factorial design: inside-the-table approach

	***0*: Placebo for A + placebo for B (n=250)**	***a*: Drug A + placebo for B (n=250)**	***b*: Drug B + placebo for A (n=250)**	***ab*: Drug A + Drug B (n=250)**
Mean cost per patient (SD)	£87,804 (£33,508)*	£98,324 (£32,408)*	£109,109 (£28,851)*	£125,015 (£29,958)*
Incremental cost per patient vs. *0* (SE)	N/A	£10,520 (£2,948)*	£21,305 (£2,797)*	£37,211 (£2,843)*
Interaction: cost (SE)	£5,386 (£3,951)			
Mean QALYs per patient (SD)	18.1 (6.6)*	18.9 (5.9)*	19.6 (6.0)*	19.8 (5.6)*
Incremental total QALYs per patient vs. *0* (SE)	N/A	0.86 (0.56)	1.53 (0.57)*	1.71 (0.55)*
Interaction: QALYs (SE)	-0.68 (0.77)			
Mean total NMB per patient at Rc=£30,000 (SD)	£455,010 (£211,746)*	£470,155 (£186,507)*	£479,504 (£188,881)*	£468,985 (£174,617)*
INB vs. *0* at Rc=£30,000 (SE)	N/A	£15,145 (£17,846)	£24,494 (£17,946)	£13,976 (£17,358)
Interaction: NMB at Rc=£30,000 (SE)	-£25,664 (£24,149)			
Cost/QALY vs. *0*	-	£12,297	£13,956	£21,809
Cost/QALY vs. *a*	-	-	£16,070	£31,375
Cost/QALY vs. *b*	-	-	-	£88,574

* Significantly greater than zero (p<0.05).

However, including the interaction term also substantially increased SEs around treatment effects and predicted costs and benefits for each group. This affects statistical inference, CEACs and EVPI. For example, using `suest` the two-sided p-value for the difference in QALYs between *b* and *0* was 0.007 with interactions and 0.002 without.[30]

An additional analysis used bootstrapping to generate an empirical distribution of costs and QALYs, by sampling with replacement 1,000 times (stratified by treatment arm) and running OLS models predicting cost and QALYs with and without interactions on each replicate. CEACs showing the proportion of bootstrap replicates where each of the four treatment-combinations had highest NMB differed markedly between analyses with and without interactions (Figure 3.3). In particular, the probability of any given cell being cost-effective was always closer to 25% (even-odds across the four treatments) when interactions were

[30] Although including interactions increases SEs, p-values may either increase or decrease, depending on the magnitude and direction of changes in the point estimate. For example, the two-sided p-value for the difference in QALYs between *a* and *0* was 0.128 with interactions and 0.177 without, since allowing for interactions also increased the difference between *a* and *0* from 0.518 (SE: 0.383) to 0.856 (SE: 0.563).

considered, except at ceiling ratios where the treatment maximising NMB differs between analyses; this appears to be a general finding. Furthermore, the way in which CEACs are presented will differ depending on whether we interpret the trial as a comparison between four mutually-exclusive cells (Figure 3.3B), or as addressing two independent questions (one on each factor; Figure 3.4).

Figure 3.3. CEACs for multiple comparisons using **A** regression without interaction term and **B** regression with an interaction term.

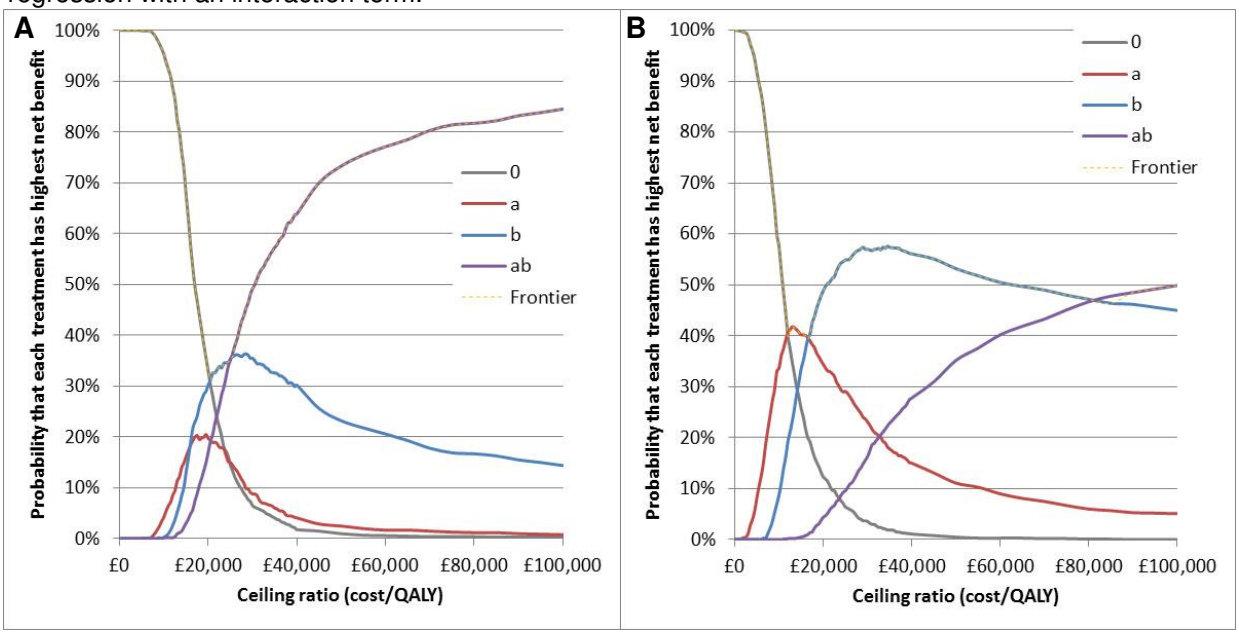
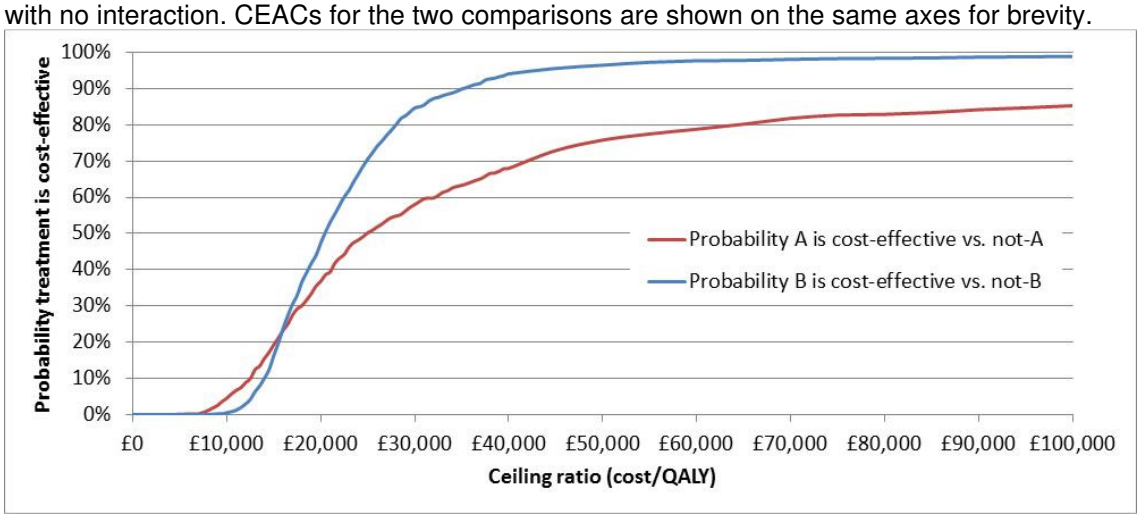


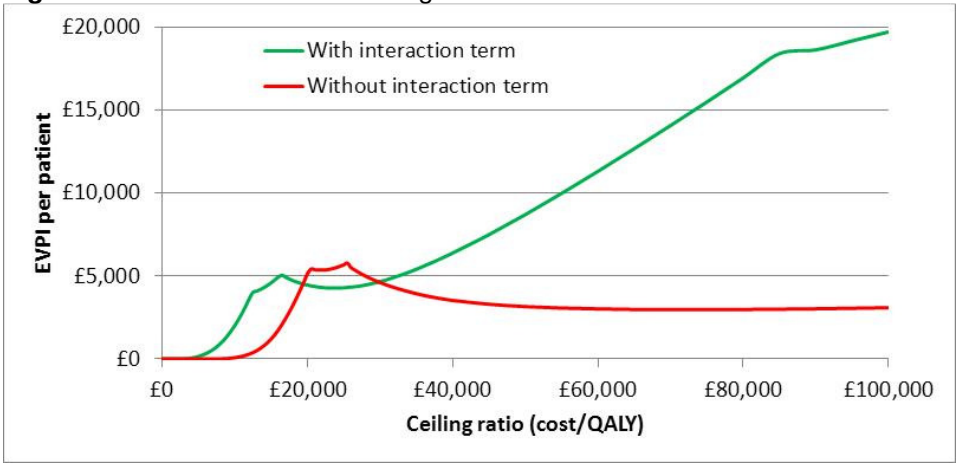
Figure 3.4. CEACs for two pair-wise comparisons, treating the two factors as independent treatments, with no interaction. CEACs for the two comparisons are shown on the same axes for brevity.



Bootstrapping analyses with and without interactions were also used to calculate the EVPI per patient. This was conducted by identifying the treatment with highest NMB in each

bootstrap replicate, averaging across these maximum NMB and subtracting the average NMB for the treatment that maximises expected NMB. The analysis with interactions generally produced much higher EVPI estimates than excluding interactions, except for the range of ceiling ratios at which the treatment having highest expected NMB differs between two analyses (Figure 3.5).

Figure 3.5. EVPI estimates from regression with and without interaction terms.



I also conducted OLS on NMB at different ceiling ratios, allowing for correlations between models using `suest`. This showed that the interaction increased linearly with ceiling ratio, while its SE increased at a faster rate, such that the p-value on the interaction increased from 0.147 to 0.306 as the ceiling ratio increased from £5,000 to £40,000/QALY. The amount of skewness and kurtosis also increased with ceiling ratio.

GLM with log-link and gamma family was also explored (also using `suest`) to evaluate whether this eliminated the interaction. The interaction for cost was smaller in this model than in OLS, with *ab* having 2% (£2,833) higher costs than would be expected from the effect of A and B individually, vs. £5,386 in OLS. However, GLM for QALYs gave a -4% interaction, which is equivalent to -0.748 on a natural scale: somewhat larger than the -0.676 interaction in OLS.

The simulated data were also used to compare different methods for allowing for the correlation between costs and QALYs. All methods other than MCMC gave identical point estimates (Table 3.13). However, all models gave slightly different SEs. In particular, `suest` and bootstrapping generated somewhat larger SEs than running two independent OLS models. Furthermore, allowing for correlations meant that SEs differed between the two factors, even for a trial with perfectly balanced patient numbers. By contrast, using either `sureg` or a mixed model with unstructured covariance generated results that were more similar to the independent OLS models.

Point estimates from MCMC models were numerically slightly different from the other models (potentially due to random fluctuation), but gave very similar SEs to the mixed model. The bootstrapping results shown did not include the residual on cost as a predictor of QALYs; however, identical results (to seven significant figures) were generated when this residual was included. The effect of including the cost residual as a predictor of QALYs within MCMC models was also negligible and comparable to the size of the MCMC error.

The Bayesian MCMC model was also replicated using informative priors for the interaction terms, alongside the same non-informative priors used previously. For simplicity, I followed Welton et al (115), by assuming that there is a 95% probability that the interaction for QALYs is sub-additive but non-qualitative (i.e. that its 95% credible interval goes from $-f$ [zero minus the smallest of the simple effects] to 0). Assuming a normal distribution, this equated to a prior with $mean = -f/2$ and $SD = \left[f/(1.96 \cdot 2)\right]^2$. It was necessary to adapt the methodology for cost, since this interaction was super-additive; for simplicity, I assumed that we can be 95% confident that the interaction for cost is positive and no larger than the simple effect of A. Assuming a normal distribution, this equated to a normal distribution with $mean = f/2$ and $SD = \left[f/(1.96 \cdot 2)\right]^2$.

Table 3.13. Effect of using different methods to allow for the correlation between costs and QALYs. Values represent mean (SE).

	Independent OLS	`suest` OLS	`sureg` OLS or mixed model with unstructured covariance†	Bootstrapping¥	MCMC§		
					With explicit correlation‡	Without explicit correlation‡	With informative prior & no explicit correlation‖
Cost							
A	10,520 (2,794)*	10,520 (2,944)*	10,520 (2,788)*	10,520 (2,935)*	10,530 (2,791)*	10,540 (2,789)*	10,520 (1,818)*
B	21,305 (2,794)*	21,305 (2,792)*	21,305 (2,788)*	21,305 (2,781)*	21,330 (2,793)*	21,300 (2,796)*	21,450 (2,211)*
Interaction	5,386 (3,951)	5,386 (3,945)	5,386 (3,943)	5,386 (3,948)	5,372 (3,959)	5,383 (3,941)	5,088 (1,989)*
Constant	87,804 (1,976)*	87,804 (2,116)*	87,804 (1,972)*	87,804 (2,113)*	87,790 (1,978)*	87,790 (1,971)*	87,800 (1,667)*
QALYs							
A	0.856 (0.542)	0.856 (0.563)	0.856 (0.541)	0.856 (0.563)	0.853 (0.542)	0.858 (0.543)	0.758 (0.490)
B	1.527 (0.542)*	1.527 (0.566)*	1.527 (0.541)*	1.527 (0.561)*	1.524 (0.542)*	1.531 (0.542)*	1.398 (0.433)*
Interaction	-0.676 (0.767)	-0.676 (0.766)	-0.676 (0.765)	-0.676 (0.763)	-0.674 (0.768)	-0.679 (0.769)	-0.421 (0.365)
Constant	18.094 (0.383)*	18.094 (0.420)*	18.094 (0.383)*	18.094 (0.417)*	18.090 (0.381)*	18.090 (0.384)*	18.140 (0.367)*

* p<0.05

† Results estimated in this column were estimated using the `sureg` command in Stata. Identical results were generated using a mixed model estimated (following methods of (215)) by reshaping data on paired costs and QALYs into long format, such that there were two rows per patient (one for costs and one for QALYs), with both cost and QALY data in the same column (Y). To ensure that costs and QALYs had separate coefficients, the model was estimated using interactions between treatment indicators and outcome type (C=1 for cost): $Y_{ij} = \beta_{0Cost} \cdot C_{ij} + \beta_{ACost} C_{ij} \cdot A_{ij} + \beta_{BCost} C_{ij} \cdot B_{ij} + \beta_{IntCost} C_{ij} \cdot A_{ij} \cdot B_{ij}$ $+ \beta_{0QALY} \cdot (1 - C_{ij}) + \beta_{AQALY} (1 - C_{ij}) \cdot A_{ij} + \beta_{BQALY} (1 - C_{ij}) \cdot B_{ij} + \beta_{IntQALY} (1 - C_{ij}) \cdot A_{ij} \cdot B_{ij} + \mu_i + \varepsilon_{ij}$. The model allowed for random effects by patient (μ_i), suppressed the constant term and used unstructured residuals to allow for the fact that costs have larger variances than QALYs.

¥ Means were based on regression analyses on the original sample, while SEs were calculated as the SD across the 1,000 bootstraps.

‡ Two chains with different initial values were run for 10,000 iterations burn-in and 20,000 sample iterations on each chain. Visual inspection of the history and Brooks-Gelman-Rubin diagnostic graphs suggested that good convergence was achieved and identical results (to four significant figures) were estimated when results were rerun using different randomly-generated initial values for all parameters other than the precision. Correlation was introduced by including cost residual (observed cost for patient i minus predicted cost) as predictor of QALYs.

§ Estimated using MCMC in WinBUGS (216). This analysis used linear models predicting costs and QALYs as a function of A, B and A*B and assigned non-informative Gaussian priors to all eight regression coefficients.

‖ Priors for the interaction terms were defined as shown in the text; non-informative priors were used for the constant term and simple effects. Two chains with different initial values were run and convergence was assessed using visual inspection of the history and Brooks-Gelman-Rubin diagnostic graphs. This model converged slowly, although tolerable convergence was achieved after a burn-in of 1 million iterations and 500,000 sampled iterations.

Including this informative prior reduced the absolute magnitude of the interactions for both costs and QALYs and also changed the size of all other coefficients and reduced all SEs (Table 3.13). Concerningly, this subjective prior halved the size of SEs around the interaction term, which may suggest that the type of prior proposed by Welton et al is highly informative and is given a large weight in the analysis.[31]

In this example, it is difficult to deal with interactions by separating costs or outcomes into different components. While it would have been possible to separate costs into drug costs, AE costs and clinical event costs and analyse these three types of events separately, it is difficult to do the same for QALYs. In particular, mean utilities will include the effect of clinical events and side-effects. However, in principle the analysis could have been conducted using event-based CEA (16, 149).

3.7 Criteria for including interactions

Unless we are willing to adopt Bayesian informative priors to down-weight interactions based on subjective prior beliefs or existing evidence (Section 3.6.3), all of the methods discussed in Section 3.6 require judgements about which interactions are included and which are excluded. In particular, the performance of all analyses will be improved by including important interactions that could bias the conclusions, while excluding unnecessary interactions that would reduce statistical power and over-complicate analysis and interpretation.

Researchers therefore need to pre-specify a rule or criterion that they can apply to their single sample (plus any prior information) to determine which interactions are included in the base case analysis. Various rules could be used, including: statistical inference; including all

[31] Welton et al also found their trial results to be sensitive to the priors used for the interaction term (115, 116).

interactions; including all large or qualitative interactions; goodness of fit; and prior evidence or beliefs.

Analysts will frequently need to apply such criteria to decide about the inclusion/exclusion of several interactions. If separate models are estimated for different NMB components (e.g. separate models for costs and QALYs) or at different ceiling ratios, judgements about interactions will need to be made for each model. Furthermore, studies with designs larger than 2x2 will have >1 interaction term per model.

3.7.1 Statistical inference

A two-stage approach to hypothesis testing is typically used to analyse clinical endpoints from factorial trials, whereby interaction terms are only included if they reach a pre-specified level of significance (1, 3, 4, 20, 42). However, there are substantial practical and theoretical arguments against using inference to determine the methods of economic evaluation.

From a practical point of view, clinical trials are likely to be under-powered to detect interactions, since many trials are under-powered to detect between-treatment differences in cost and effect (18, 137-141, 146), which are likely to have variances a quarter of the size of those around interaction terms (26, 35, 37). Furthermore, unlike clinical endpoints, where interactions are generally small (1) and rarely qualitative (34), large or qualitative interactions for QALYs, costs and NMB are likely to be relatively common (Section 3.4). Since the bias introduced by omitting the interaction term is half the size of the true interaction (3), ignoring interactions may introduce negligible bias for clinical endpoints, but cause substantially larger biases for costs, QALYs and NMB that may change the conclusions.

More fundamentally, economic evaluation unavoidably focuses on estimation since we must trade expected health gains against expected cost and compare our ICER against the ceiling

ratio. In particular, it has been argued (most prominently by Claxton (157)) that statistical inference is irrelevant to clinical decision-making and that maximising health gains from the budget requires us to choose the strategy with the highest (posterior) mean NMB regardless of statistical significance. In particular, the conventional significance level is arbitrary[32] and a decision has to be made now to adopt one of the mutually-exclusive alternatives. Continuing to use an existing treatment with lower expected NMB in preference to a newer alternative that has non-significantly higher NMB has an opportunity cost (less health is gained from the available budget) and means that decisions are unduly affected by the order in which treatments come to market. Claxton therefore argued that uncertainty around costs and benefits should only influence decisions about further research to reduce decision uncertainty, rather than directly driving decisions about which treatment to adopt today.[33]

Many (but not all[34]) health economists hold this view of statistical inference (125): for example, many applied and methodological studies start from the premise of maximising expected NMB (e.g. (126, 148, 218)) and NICE commonly recommend treatments with a ≥5% chance of accruing lower NMB than their competitors.[35]

If we consider inference to be irrelevant to healthcare decision-making, it is unclear why the statistical significance of an interaction term should determine the form of analysis. The two-

[32] Notably, the origin of the 0.05 significance level appears to coincide with the promotion of factorial trials by Fisher, who comments that (27): "if the experimenter could say that in twenty years experience with uniform treatment the difference in favour of the acre treated with manure had never before touched [the figure observed with the current intervention], the evidence would have reached a point which may be called the verge of significance."

[33] In practice, decisions about further research cannot be entirely separated from decisions about treatment adoption, since early adoption may affect the feasibility of research (217), although this wider framework for healthcare and research decision-making is beyond the scope of this thesis.

[34] Other health economists are concerned that ignoring statistical inference and uncertainty in healthcare decision-making risks adopting treatments that are actually less effective but appear to maximise NMB in the trial sample due to chance.

[35] However, NICE also commonly recommend several alternative medications for the same group of patients unless there is clear (or statistically significant?) evidence that outcomes differ. For example, NICE recommend both tenofovir and entecavir for hepatitis B (219, 220). This strategy also carries an opportunity cost, since a proportion of patients will continue to receive the treatment with lower expected NMB (entecavir), when the same budget would allow more patients to be treated with tenofovir with similar (or even superior) outcomes (221).

stage approach also equates absence of proof with proof of absence and assumes that the interaction is exactly 0 (with no uncertainty) unless it reaches an arbitrary level of statistical significance.[36]

If we used a two-stage approach in the worked example, we would base our conclusions on regression without an interaction term since interactions were not statistically significant. Omitting the non-significant interactions would have introduced substantial bias into the analysis that would change the conclusions of our economic evaluation and prevent us from allocating healthcare resources efficiently.

3.7.2 Including all interactions

In situations where we are only interested in maximising expected NMB subject to current information to inform the immediate treatment-adoption decision, there is a strong case for including all interactions. Including all interactions, regardless of their size or statistical significance, ensures that adoption decisions are based on an unbiased comparison of expected NMB that avoids additivity assumptions that can introduce systematic bias. This approach is statistically consistent, in that we would always adopt the treatment that truly had the highest NMB if the sample size were infinite. By contrast, a decision rule excluding all interactions would cause us to adopt a suboptimal treatment whenever there was a qualitative interaction in NMB that change which treatment had highest NMB, even with an infinite sample size.[37]

[36] Similar issues are raised by CMA, with many researchers historically assuming that two treatments have equal efficacy based on a lack of statistical significance (126). This implicit assumption within CMA has since been shown to bias estimates of uncertainty in addition to giving misleading conclusions (126, 158).

[37] This can be seen with the example shown in Section 3.6.4: a trial with infinite sample size might give identical point estimates to those shown in Table 3.12, but have SEs equal to 0. Excluding interactions and assuming that the two treatments have additive effects on costs and QALYs would lead us to adopt *ab*, whereas *b* is actually the treatment with highest expected NMB.

In practice however, decisions must be made with imperfect information and finite sample sizes. Although always including interactions eliminates the risk that systematic bias influences our conclusions, our results may be more sensitive to chance. In particular, we may, by chance, observe a large interaction in our sample that drives our conclusions, even if the true population interaction is actually zero. If this is the case, inside-the-table analysis may, by chance, lead us to adopt a treatment that does not genuinely have highest NMB.

Furthermore, even if we consider inference to be irrelevant to the current adoption decision, uncertainty around our results is nonetheless critical to the decision about whether or not to collect further information. As shown in Figure 3.5, including all interactions generally gives higher EVPI estimates than assuming additive effects. If the true interaction is zero, including all interactions may therefore cause us to spend more on research than is socially optimal.

3.7.3 Including "large" interactions

A compromise between inefficiency and bias, or between systematic and chance error, may be to include only interactions above a certain size, or only those that change the conclusions.

Hung et al previously proposed a trade-off between bias and inefficiency, whereby the p-value used to test the statistical significance of interactions was chosen to keep bias within acceptable limits (4). This approach is based on the principle that omitting the interaction term biases estimates of simple effects (μ_a-μ_0) by an amount equal to half the size of the true interaction.

Setting an absolute limit for the size of interaction that can be ignored based on the amount of bias that we are willing to tolerate is therefore another approach that could be used in economic evaluation. However, this opens the question of how large this limit should be.

Minimally important differences for EQ-5D utilities between 0.036 and 0.074 have been proposed (222-224); similarly the smallest difference in utility that could be identified within the EQ-5D valuation study is 0.025 (three months out of 10 years) (225, 226). However, QALY interactions <0.025 may be meaningful and/or larger than main effects for some interventions: e.g. a cheap prevention program that prevents one death per million people. Some guidelines also recommend rounding ICERs to two significant figures (227). Furthermore, the impact of any given amount of bias on decision-making will depend on how close the ICER is to the ceiling ratio.

Alternatively, it may be more appropriate to set limits based on the size of the interaction term *relative* to the simple effects. In particular, allowing for interactions will have no effect on the conclusions about which treatment is adopted unless the interactions are qualitative on a NMB scale (i.e. change the ranking of treatments) at the ceiling ratio(s) of interest.[38] However, since the true shadow price of a QALY is unknown, decision-makers may be interested in results at a range of ceiling ratios. Including all interactions that are qualitative at any ceiling ratio would generally result in inclusion of all interactions, since any quantitative interaction in either costs or QALYs will produce a qualitative interaction in NMB at some ceiling ratio whenever the treatment lies in the north-east or south-west quadrants (Section 3.4). In principle, we could include all interactions that are qualitative at any ceiling ratio within a pre-specified range (e.g. £20,000-£30,000/QALY), although this requires arbitrary judgements about the ceiling ratio range and in practice this will mean inclusion of all interactions whenever the ICER point estimate lies within this range.[39] Furthermore, as discussed in Chapter 6, the impact of allowing for one interaction can vary depending on

[38] Given a set of mutually-exclusive alternatives, we maximise NMB from the budget by adopting the treatment with highest expected NMB. Only interactions that change the ranking of treatments by NMB can therefore change our conclusions. Furthermore, the only type of interactions that can change the ranking of treatments are qualitative interactions, since these, by definition, change the *sign* of the difference between two options.

[39] Since interactions for NMB will be qualitative at a ceiling ratio equal to the ICER (Section 3.4), we will always observe qualitative interactions within the range of ceiling ratios we specify whenever any of the ICERs lie within this range.

which other interactions are considered in the analysis, which makes any decision rule of this type difficult to implement.

Due to these complexities, a pragmatic alternative may be to only include those interactions that are qualitative on the scale of costs or QALYs (rather than NMB). This decision rule is independent of the ceiling ratio and is substantially simpler to implement than those based on interactions for NMB, as separate decisions about interactions can be made for each component of NMB. Including all qualitative interactions for cost and QALYs ensures that the ranking of treatments by NMB is unbiased at ceiling ratios of zero and infinity, although qualitative interactions for NMB will remain at intermediate ceiling ratios.

A variant on this decision rule is to include all interactions that have an absolute magnitude larger than the smallest of the two simple effects, which ensures that the maximum amount of bias tolerated is half the size of the smallest simple effect. This rule ensures that very large super-additive interactions are included, as well as all qualitative interactions, which may help to avoid the risk of adopting a suboptimal treatment at intermediate ceiling ratios.

3.7.4 Other criteria

Model fit provides another rule for trading efficiency against bias: particularly Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the deviance information criterion (DIC). These likelihood-based measures take account of sample size and are based on the amount of information lost by simplifying reality using a model (228). Clinical analyses of some published factorial experiments outside healthcare (e.g. (229)) have used AIC to decide whether to include interactions. However, although these measures do trade efficiency against bias by imposing penalties for model complexity (228), the trade-off is based on information theory, rather than decision analysis or the principles of economic

evaluation. As such, AIC may not minimise the risk or consequences of adopting the wrong treatment.

Empirically, these measures often favour exclusion of very large interactions that could change the conclusions. For example, within the worked example described previously, AIC and BIC both suggest that OLS without an interaction is the best model (Table 3.14), although this biases the analysis in favour of *ab*. GLM models also had higher AIC than OLS. Furthermore, these measures appear to be sensitive to the scale of analysis: since QALYs are substantially smaller than costs, they have smaller likelihoods, which means that the two-point penalty imposed for each model parameter has proportionately more effect on AIC for QALYs than cost. Furthermore, AIC often does not distinguish the fit of different GLM link functions unless numerous highly-prognostic covariates are included (162).

Table 3.14. Effect of interactions on model fit for the worked example. AIC and BIC were estimated using the `estat ic` post-estimation command in Stata.

Model		**Cost**		**QALYs**	
		AIC	**BIC**	**AIC**	**BIC**
OLS	With interaction	23,541	23,560	6,446	6,465
	Without interaction	**23,540**	**23,555**	**6,445**	**6,459**
GLM with log-link and gamma family with interaction	With interaction	25,116	25,135	7,908	7,927
	Without interaction	25,114	25,128	7,906	7,921

Another approach would be to include only those interactions that are hypothesised a priori and exclude all unexpected interactions. This is equivalent to a Bayesian analysis taking an extremely strong subjective prior that assumes that we are 100% certain that an unexpected (or unexplainable) interaction equals zero.

Furthermore, the decision rules discussed above could be used in combination. For example, we could include all interactions that are either statistically significant or larger than either simple effect. In particular, it may also be appropriate to consider the biological plausibility of any interactions observed, in addition to applying statistical rules. For example, in the United Kingdom Prospective Diabetes Study (UKPDS), the authors included only variables that

were both statistically significant and supported by prior epidemiological evidence (210). Similarly, in a factorial trial, it may be appropriate to exclude interactions that have no biological explanation even if they are statistically significant or larger than simple effects, or include an interaction that was hypothesised a priori regardless of its magnitude or significance.

3.8 Opportunity cost of ignoring interactions

Many of the methods used to identify subgroups and estimate treatment-subgroup interactions (e.g. tests for interactions (107)) are also relevant to estimation of treatment-treatment interactions within factorial trials. The expected value of stratified decision-making has been proposed to quantify the benefits of allowing for heterogeneity in economic evaluation and making separate decisions for different subgroups (218, 230). The value of stratification is defined as the NMB gained from adopting treatment in only those subgroups for which it has positive expected INB, minus the NMB of adopting it in the whole population (218, 230).

The static value of stratification ($\Delta_s NMB$) reflects the benefits of stratification under current information (218, 230):

$$\Delta_s NMB = NMB_s - NMB = \sum_{k=1}^{s} w_k \left(\max_j NMB_k\right) - \max_j NMB \qquad (3.6)$$

Where $\max_j NMB$ equals the NMB of the treatment that would be adopted for the total population (without stratification), $\max_j NMB_k$ equals the NMB of adopting the best treatment for subgroup k and w_k equals the weight or proportion of the population who are in subgroup k. Similarly, the dynamic value of stratification reflects the NMB gained under

perfect information and equals the EVPI with stratification minus the EVPI without stratification.

In principle, either a 2x2 factorial trial and a subgroup analysis could evaluate A vs. not-A in patients receiving B and in patients not receiving B and explore interactions between A and B. However, since the factorial trial has *randomised* allocation to B, it can give stronger causal inferences (Section 10.3) and inform decisions about use of B as well as A. When interpreting subgroup analyses, we consider the stratifying factor to be outside of our control and stratification means that we make recommendations about whether A should be given to older people or to those who also need treatment B, as well as independently recommending whether A should be given to younger people or those not needing treatment B. By contrast, when interpreting the results of a factorial trial, both factors are under the decision-maker's control: he/she can make recommendations on whether or not B should be given as well as making recommendations on A. For factorial trials, stratification means that we consider the cells as separate, mutually-exclusive treatments and adopt the treatment-combination that maximises expected NMB, rather than considering factors independently.

I propose the ***opportunity cost of ignoring interactions*** as a measure an analogous to the value of stratification, which shows the NMB lost from assuming additive effects and making separate decisions on the two factors, rather than taking account of interactions and making a joint decision on both (or all) factors simultaneously.[40] The opportunity cost of ignoring interactions under current information ($OpCostIgnoring\,\mathrm{int}_{\mathrm{Current}}$) is therefore equal to the expected NMB for the treatment-combination that would be adopted if we take account of all interactions, minus the NMB for the treatment combination that would be adopted based on an analysis ignoring some or all interactions. This equals zero whenever this pair of analyses give the same conclusion. To obtain a fair, consistent, unbiased estimate of the opportunity

[40] As per the convention used within this thesis, I define the opportunity cost of ignoring interactions in terms of NMB. However, it could equivalently be expressed in health units (e.g. QALYs) using NHB.

cost, both NMB estimates should be based on analyses allowing for interactions (i.e. inside-the-table).[41]

$$OpCostIgnoring\,\mathrm{int}_{Current} = \max(NMB_{inside-the-table}) - NMB_{inside-the-table}(C) \qquad (3.7)$$

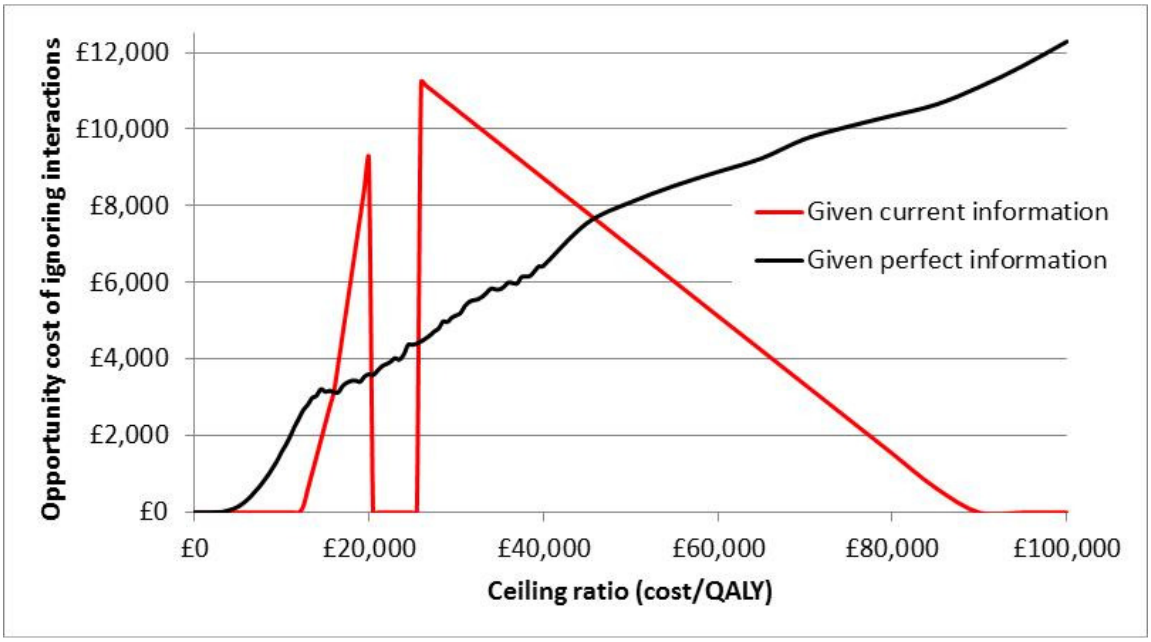
Where *C* represents the treatment that is predicted to have highest NMB ignoring interactions (i.e. at-the-margins) and $NMB_{inside-the-table}(C)$ equals the inside-the-table estimate of NMB for *C*.

For example, at a £30,000/QALY ceiling ratio, $OpCostIgnoring\,\mathrm{int}_{\mathrm{Current}}$ within the worked example equals £10,518/patient: the inside-the-table estimate of the NMB for the treatment that would be adopted based on inside-the-table analysis (*b*, NMB=£479,504; Table 3.12), minus the inside-the-table estimate of the NMB for the treatment that would be adopted based on at-the-margins analysis (*ab*, NMB=£468,985). In other words, the value of allowing for interactions is £10,518/patient. This means that ignoring interactions and basing decisions on at-the-margins analysis wastes £10,518 of National Health Service (NHS) funds per patient treated, which is equivalent to losing 0.53 QALYs/patient. At a population level, spending £1 billion on *ab* will allow us to treat 8,000 patients and accrue 16,000 QALYs, whereas allowing for interactions and adopting *b* will allow us to treat 9,165 patients and accrue 18,000 QALYs: 2,000 more than we would achieve using the biased at-the-margins estimates and making separate decisions on A and B.

$OpCostIgnoring\,\mathrm{int}_{\mathrm{Current}}$ was zero in three ceiling ratio ranges (£0-£12,000, £20,500-£25,500 and ≥£90,000; Figure 3.6), which correspond to the ratios at which analyses with and without interactions give the same treatment adoption decision.

[41] If the opportunity cost is calculated using NMB estimates from two different analyses (e.g. if the NMB for the treatment combination that would be adopted based on at-the-margins analysis is calculated as the maximum NMB predicted in regression without interaction terms), the opportunity cost may, misleadingly, appear to be zero (or even negative) if the bias inherent within at-the-margins analysis causes NMB for that treatment to be overestimated.

Figure 3.6. Opportunity cost of ignoring interactions given both current and perfect information.



The opportunity cost of ignoring interactions under perfect information ($OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$) can be estimated by replicating equation 3.7 across the posterior distribution of NMB for each treatment (e.g. by repeating it for each of our B bootstrap replicates) and averaging across the distribution (or across all bootstraps).

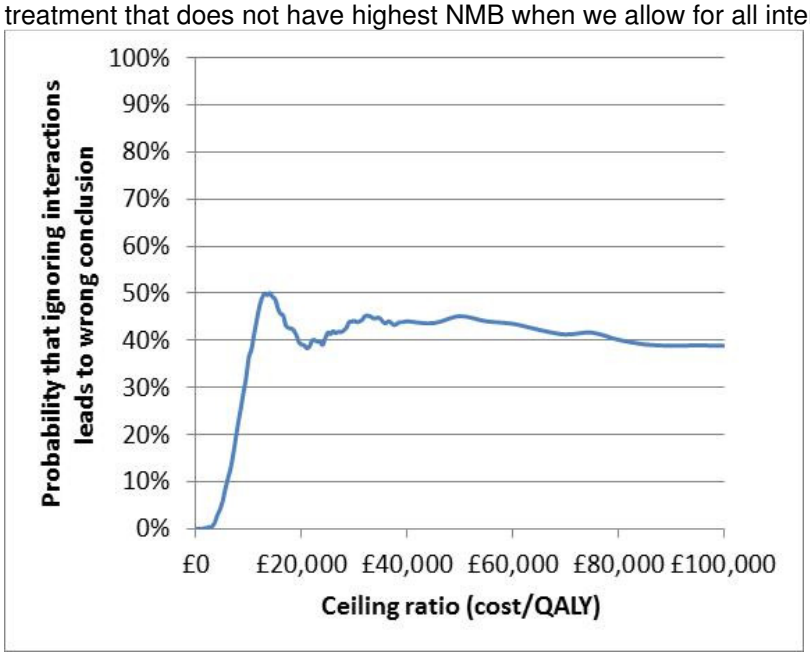
$$OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}} = \frac{1}{B}\sum_{b=1}^{B}\left(\max(NMB_{inside-the-table,b}) - NMB_{inside-the-table,b}(C_b)\right) \quad (3.8)$$

For the worked example described above, this was implemented by identifying the treatment that would be adopted based on inside-the-table and at-the-margins analyses for each bootstrap replicate. The opportunity cost of ignoring interactions was then calculated for each bootstrap replicate as the difference in the inside-the-table NMB estimates for these two treatments; these values were then averaged across all 1,000 bootstrap replicates.

Bootstrapping also enables us to calculate the probability that ignoring interactions would give a misleading conclusion (i.e. adoption of a treatment that does not have highest NMB when we allow for interactions), which is equal to the proportion of bootstrap replicates in

which the opportunity cost is >0. Here, the probability is zero at ceiling ratios <£1,000 (since there is negligible chance that any treatment other than *0* has highest NMB regardless of whether interactions were considered) and peaks at £13,000/QALY (Figure 3.7), just above the ICER for *a* vs. *0*.

Figure 3.7. Probability that ignoring interactions leads to the wrong conclusion (i.e. adoption of a treatment that does not have highest NMB when we allow for all interactions).



$OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$ takes account of the implications of adopting the wrong treatment, as well as the probability that the conclusions are sensitive to interactions; it therefore follows a similar trend to the probability, but continues to rise with ceiling ratio as the value placed on maximising QALYs increases (Figure 3.6). Unlike $OpCostIgnoring\,\mathrm{int}_{\mathrm{Current}}$, which equals zero when the point estimates of the analyses give the same conclusion, $OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$ will not equal zero unless there is no decision uncertainty. Situations where $OpCostIgnoring\,\mathrm{int}_{\mathrm{Current}}$ equals zero, but $OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$ is >0 are analogous to those with dynamic value of heterogeneity (218), suggesting that conducting further research could make the conclusions sensitive to interactions. $OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$ will generally be approximately equal to the difference in EVPI between the two analyses, but will

rarely be exactly equal since the EVPI estimates from at-the-margins analysis are biased due to interactions being ignored, while those from inside-the-table analysis may be overestimated due to inefficiency. However, it remains unclear precisely how analysts and decision-makers should interpret $OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$ and what the implications of this figure are for decision-making.

Both $OpCostIgnoring\,\mathrm{int}_{\mathrm{Current}}$ and $OpCostIgnoring\,\mathrm{int}_{\mathrm{Perfect}}$ are based on comparisons between two specific analyses; the analyses being compared need to be clearly stated in all applications. One limitation of these measures is that for applied examples, the inside-the-table analysis used to estimate NMB and obtain unbiased conclusions is only a “silver standard” and could give the “wrong” treatment adoption decision by chance.

3.9 Conclusions

In economic evaluations based on factorial RCTs, it is essential to give appropriate consideration to how the factorial design affects the analysis. I have shown that large or qualitative interactions are particularly likely to arise for economic endpoints (Section 3.4). Furthermore, to inform efficient allocation of scarce healthcare resources, it is essential to consider both factors simultaneously and treat the cells as mutually-exclusive alternatives (Section 3.5). However, although the interpretation of results must be based on inside-the-table presentation of mean costs and outcomes, it may nonetheless be useful to exclude interactions from some of the analyses used to calculate mean costs and outcomes.

Regression analysis of economic endpoints from factorial trials raises several unique issues, such as how interactions for NMB vary with ceiling ratio (Section 3.6). The worked example also demonstrates how the choice of analytical method could affect the conclusions of the

economic evaluation and resource allocation decisions based upon it (Section 3.6.4). Chapters 4-8 explore the extent to which this happens in real trials.

Various criteria can be used to identify which interactions are important and should be included in the base case analysis, including statistical significance, goodness of fit and absolute or relative size (Section 3.7); these criteria are evaluated in Chapter 5. In addition to empirical performance, the choice of criteria may depend on the more fundamental question of how important it is to have unbiased point estimates relative to efficient statistical analysis. For clinical endpoints, it may be preferable to accept a small amount of bias to gain statistical power, since conclusions are driven by statistical significance, not point estimates. By contrast, in order to obtain unbiased (if imprecise) point estimates for economic evaluation, it is more appropriate to include interaction terms unless interactions are shown to be negligible, rather than excluding them unless interactions are shown to be important. This conclusion differs from those of seminal reviews on clinical analyses of factorial RCTs (1, 20), which recommended that at-the-margins analysis best addresses the principal research question and is appropriate for a trial design that is typically used in situations where there is no interaction for the primary endpoint. This difference largely results from the different endpoints, objectives and interpretive framework that are appropriate for economic evaluation, which mean that we need to choose between treatment combinations based on incremental costs and QALYs, rather than focusing on the statistical significance of differences in clinical endpoints on any convenient analytical scale. However, although unbiased, analyses allowing for interactions may nonetheless give the wrong answer by chance (especially for small samples) and would systematically overestimate the value of information and bias CEACs in cases where there is genuinely no interaction.

Chapter 4: Systematic review of economic evaluations of factorial trials

4.1 Introduction

Before commencing economic evaluations on my case studies, I conducted a systematic review of economic evaluations based on factorial RCTs to review the features and methodology used in previous studies and assess how economic evaluations of factorial trials were typically conducted before the programme of work described in this thesis.

The aims of this review were:

- To identify all examples of full economic evaluations conducted on patient-level data from factorial design RCTs that were published before 31st December 2009.
- To review the characteristics, study design and statistical methods used in these studies, identify any issues raised by their conduct or results and assess the strengths and weaknesses of each approach.
- To estimate the incidence of different types of interactions within trials and assess the impact of interactions on the conclusions.

After describing the methods of the systematic review, I summarise the characteristics of previous economic evaluations based on factorial RCTs and the methods they used to analyse and present the data.

4.2 Methods for systematic review

MEDLINE (including daily update and old MEDLINE), EMBASE, Econlit and Journals@Ovid were searched through Ovid on 9th February 2010. I also searched www.bmj.com, Tufts CEA

registry (https://research.tufts-nemc.org/cear/Default.aspx), Wiley Interscience, National Institute for Health Research (NIHR) HTA publications list (http://www.hta.ac.uk) and Centre for Reviews and Dissemination (CRD, http://www.crd.york.ac.uk/crdweb)[42] on the same date. Search terms to identify factorial trials (e.g. "factorial", "2 x 2", "2 by 2", "two by two", or "2 x 3") were combined with search terms to identify economic evaluations ("cost-effect*" or "economic evaluation"; Appendix F). Since some papers on factorial trial-based economic evaluations do not describe the design as factorial, clinical papers on factorial trials that mentioned plans for an economic evaluation or collection of cost data were also identified during the main literature searches. Additional targeted literature searches were then conducted to identify papers reporting economic evaluations of these specific factorial trials.

I examined titles and abstracts to assess whether they met all of the following inclusion criteria:

- Described the methods and/or results of a full economic evaluation quantifying the costs *and* benefits of interventions designed to improve health or affect healthcare systems. CEA, CUA, cost-consequence analyses (CCA) and cost-benefit analyses (CBA) were classed as full economic evaluations, although cost-minimisation analyses (CMAs) were excluded.
- Used patient- or cluster-level data from a factorial RCT. Factorial RCTs were defined as studies in which ≥2 factors (each with ≥2 levels) are deliberately varied at the same time, with participants and/or centres being randomly allocated to receive different combinations of levels for each factor in such a way that interactions between ≥2 factors can be evaluated (Section 2.3). Studies in which the interaction between factors was, to a greater or lesser extent, confounded by treating the group assigned to no treatment or combination therapy differently were included only if the authors described their study as factorial (Section 4.3). Studies with individual-level randomisation were only included if

[42] CRD includes the Database of Abstracts of Reviews of Effects (DARE), the NHS Economic Evaluation Database (EED) and the HTA database.

they used patient-level data. However, I also included cluster-randomised trials that collected only one set of data on costs and/or benefits for each *cluster* (e.g. measuring cost per general practice (GP), rather than per patient).

- Published at least brief details of the methods and/or results of the trial-based economic evaluation before 31st December 2009. Studies were not excluded from the review based on language, providing that at least some text (e.g. an abstract) was available in English. Protocols published as journal articles by 31st December 2009 were also included.

I extracted data on study characteristics, study design, statistical methods and key findings based on the main economic evaluation publication and any associated protocols, HTA reports or clinical publications cited in the economic evaluation paper (Appendix G).

Mean costs and mean health benefits within each cell of the factorial design and their SDs were extracted if reported. These data were used to estimate the magnitude, influence and (where possible) statistical significance of interactions. The magnitude of interactions relative to simple effects was calculated using the ***interaction:effect ratio***.[43] I developed this ratio as a single measure of the size of interactions (relative to between-group differences) that demonstrates whether the interaction is super-additive, sub-additive/mixed or qualitative. When all treatments have the same direction of effect (e.g. when A and B both increase cost, or both decrease cost), the interaction:effect ratio is calculated by dividing the interaction

[43] The interaction:effect ratio differs from the "interaction ratio" used by McAlister et al (1). The interaction ratio is simply the ratio of relative effect (e.g. odds ratio) of A vs. not-A for patients also receiving B, divided by the relative effect of A vs. not-A for patients not receiving B (interaction ratio = $\left(odds_{ab} / odds_{b}\right) \div \left(odds_{a} / odds_{0}\right)$) and therefore equals the interaction on a logarithmic scale (interaction on log-scale = $\exp[(\ln[odds_{ab}] - \ln[odds_{b}]) - (\ln[odds_{a}] - \ln[odds_{0}])]$). Unlike the interaction:effect ratio, the interaction ratio is appropriate only for data interpreted on a multiplicative scale and does not distinguish between qualitative and non-qualitative interactions. At least one previous paper has used the ratio of the interaction divided by simple effect to describe the ranges of interaction magnitude in which different analytical approaches performed best (4). However, this study did not include any adjustment for mixed interactions, did not link the ranges of ratio values with different types of interaction and did not propose this ratio as a method for describing interactions in general.

term by whichever of the simple effects[44] has the smaller absolute magnitude. Dividing the interaction by the *smaller* of the two simple effects ensures that qualitative interactions (those changing the ranking of treatments) have interaction:effect ratios <-1. For "mixed" situations, where the simple effects have different signs (e.g. when A increases cost, while B decreases cost), I ensured that qualitative interactions had interaction:effect ratios <-1 by dividing the interaction by the simple effect that had the opposite sign from the interaction term.

The interaction:effect ratio (IER_{AB}) therefore equals the ratio of the interaction (I_{AB}) divided by the simple effect of A ($\delta_A = \mu_a - \mu_0$), where the factor defined as A is chosen such that for non-mixed interactions $|\mu_a - \mu_0| < |\mu_b - \mu_0|$ and for mixed interactions δ_A has the opposite sign to I_{AB}:

$$IER_{AB} = I_{AB} / \delta_A = (\mu_0 - \mu_a - \mu_b + \mu_{ab}) / (\mu_a - \mu_0) \quad (4.1)$$

In all cases, interaction:effect ratios <-1 indicate qualitative interactions, ratios between -1 and 0 indicate sub-additive or mixed interactions, ratios equal to 0 indicate additive effects, while interaction:effect ratios >0 indicate super-additive interactions.

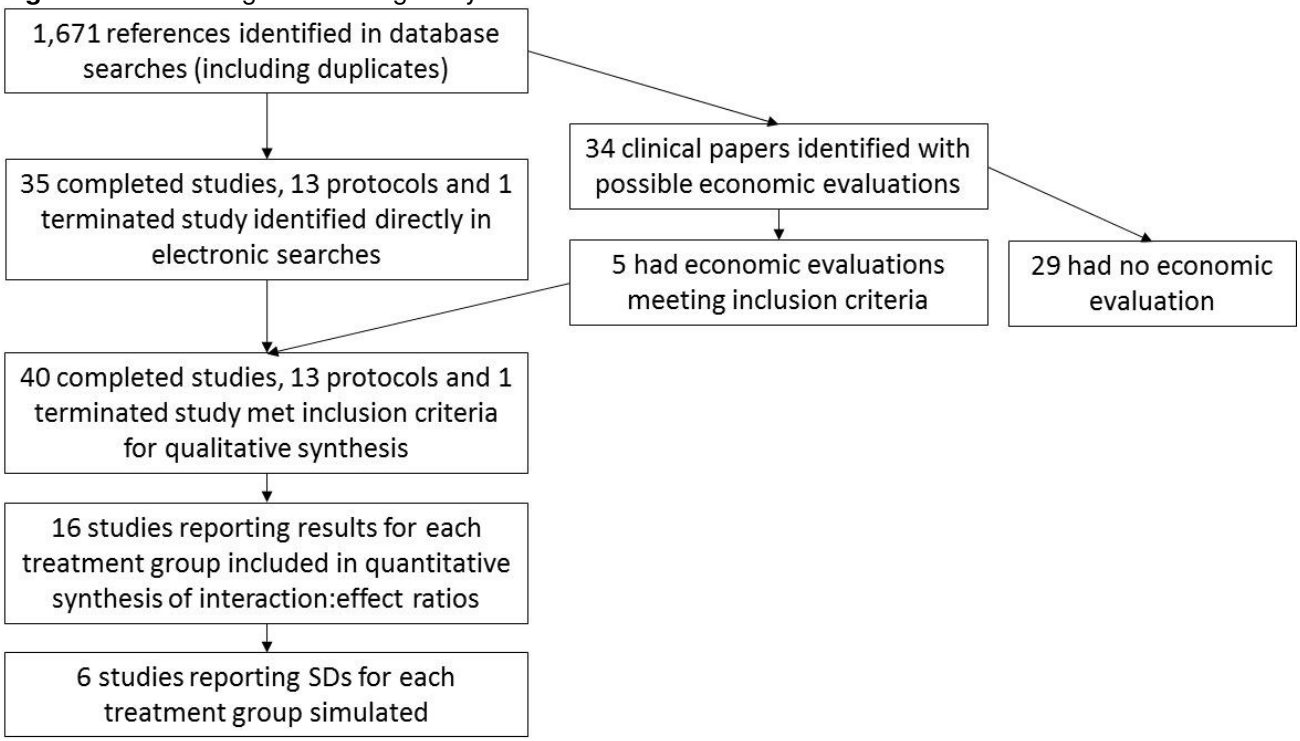
4.3 Results of the literature review

Across all searches, 1,671 references were identified (including duplicates; Figure 4.1, Appendix F). Of these, 40 complete studies presenting economic evaluation results, 13 published protocols and one study that terminated early due to poor recruitment but was published as an HTA monograph without analysis of economic results met the inclusion

[44] The interaction equals $I_{ab} = (\mu_0 - \mu_a - \mu_b + \mu_{ab})$, while the simple effect of A is defined as $\delta_A = \mu_a - \mu_0$: i.e. the difference between outcomes in the group receiving A alone and the group receiving no treatment.

criteria. Five further protocols were identified from the HTA website, but were excluded since they have not been published in journals. Appendix G gives details of all included studies.

Figure 4.1. Flow diagram showing study identification



The number of factorial trials with economic evaluations appears to be increasing, with 19% (11/40) of completed studies being published in the last two years of the review period (Table 4.1).

Table 4.1. Characteristics of the studies meeting inclusion criteria

		Completed studies	Protocols and terminated studies	Overall
Publication year*	2009	18% (7/40)	50% (7/14)	26% (14/54)
	2008	10% (4/40)	14% (2/14)	11% (6/54)
	2007	5% (2/40)	29% (4/14)	11% (6/54)
	2006	10% (4/40)	0% (0/14)	7% (4/54)
	2005	15% (6/40)	0% (0/14)	11% (6/54)
	2000-2004	28% (11/40)	7% (1/14)	6% (3/54)
	Pre-2000	15% (6/40)	0% (0/14)	9% (5/54)
Mainly/entirely publicly funded		75% (30/40)	100% (14/14)	81% (44/54)
Country where research done	UK	28% (11/40)	43% (6/14)	31% (17/54)
	US	38% (15/40)	7% (1/14)	30% (16/54)
	Other individual country	23% (9/40)	29% (4/14)	24% (13/54)
	Multinational	13% (5/40)	21% (3/14)	15% (8/54)

		Completed studies	Protocols and terminated studies	Overall
Disease area	Cardiovascular disease	45% (18/40)	0% (0/14)	33% (18/54)
	Musculoskeletal	13% (5/40)	14% (2/14)	13% (7/54)
	Cancer	15% (6/40)	0% (0/40)	11% (6/54)
	Smoking, drugs or alcohol	13% (5/40)	29% (4/14)	17% (9/54)
	Other	25% (10/40)	57% (8/14)	33% (18/54)
Factors target different diseases		3% (1/40)	14% (2/14)	6% (3/54)
Type of intervention‡	Pharmaceutical	45% (18/40)	36% (5/14)	43% (23/54)
	Physiotherapy, exercise and related	15% (6/40)	29% (4/14)	19% (10/54)
	Nutrition	8% (3/40)	14% (2/14)	9% (5/54)
	Surgery	13% (5/40)	7% (1/14)	11% (6/54)
	Training, counselling, incentives and logistical interventions	50% (20/40)	43% (6/14)	48% (26/54)
Sample size†	Mean (range)	3,420 (126, 20,536)	1,371 (132, 4,733)	3,002 (126, 20,536)
	<200 patients	13% (5/39)	7% (1/14)	11% (6/53)
	200-999 patients	38% (15/39)	36% (5/14)	38% (20/53)
	>1,000 patients	49% (19/39)	29% (4/14)	43% (23/53)
Type of economic evaluation	CEA	68% (27/40)	0% (0/14)	50% (27/54)
	CEA and CUA	18% (7/40)	43% (6/14)	24% (13/54)
	CUA	8% (3/40)	29% (4/14)	13% (7/54)
	CCA	8% (3/40)	0% (0/14)	6% (3/54)
	CUA and CBA	0% (0/40)	7% (1/14)	2% (1/54)
	Depends on results	0% (0/40)	7% (1/14)	2% (1/54)
	Not stated	0% (0/40)	14% (2/14)	4% (2/54)
Cluster-randomised		13% (5/40)	43% (6/14)	20% (11/54)
Type of design	Full factorial	88% (35/40)	86% (12/14)	87% (47/54)
	Partial factorial	10% (4/40)	7% (1/14)	9% (5/54)
	Incomplete factorial	3% (1/40)	7% (1/14)	4% (2/54)
Factorial matrix size	2 factors & 2 levels (e.g. 2x2, 2x2+1, or double 2x2)	78% (31/40)	71% (10/14)	76% (41/54)
	>2 factors each with 2 levels (e.g. 2x2x2)	10% (4/40)	0% (0/14)	7% (4/54)
	2 factors with >2 levels on ≥1 factor (e.g. 3x2)	10% (4/40)	21% (3/14)	13% (7/54)
	>2 factors with >2 levels on ≥1 factor (e.g. 3x2x2)	3% (1/40)	7% (1/14)	4% (2/54)
Primary clinical endpoint	Assumed no interaction	53% (21/40)	29% (4/14)	46% (25/54)
	Included interaction term	23% (9/40)	7% (1/14)	19% (10/54)
	Unclear	25% (10/40)	64% (9/14)	35% (19/54)
Base case economic evaluation	Assumed no interaction	40% (16/40)	7% (1/14)	31% (17/54)
	Included interaction term	53% (21/40)	0% (0/14)	39% (21/54)
	Unclear	8% (3/40)	93% (13/14)	30% (16/54)
Presentation of uncertainty around the ICER	None	48% (19/40)	-	48% (19/40)
	Present using CIs or scatter graphs	15% (6/40)	-	15% (6/40)
	CEACs: pair-wise comparison(s)	23% (9/40)	-	23% (9/40)
	CEACs: multiple comparison(s)	8% (3/40)	-	8% (3/40)
	CEACs: pair-wise and multiple comparison(s)	8% (3/40)	-	8% (3/40)

* Completed studies are categorised by the year when the first paper describing the results of the economic evaluation was published, while protocols are categorised by the date the protocol was published.
† Excludes one cluster-randomised study for which the number of patients was not reported.
‡ Numbers add up to >100% since 16 studies/protocols evaluated two or more different types of intervention.

Around two-thirds of studies and protocols (33/54) were conducted in the US or UK. The remainder were conducted in Europe, North America or Australasia, with the exception of two studies in Africa and one in Chile. Only 10 studies/protocols were predominantly manufacturer-funded; UK public funding bodies (e.g. NIHR, Medical Research Council and Wellcome Trust) funded 18 studies/protocols and US public funders sponsored 14.

Cardiovascular disease interventions were evaluated in 40% (18/40) of completed studies (but no protocols), mirroring the large number of clinical analyses of factorial trials in this disease (1). However, other studies covered diverse conditions including musculoskeletal pain, promoting healthy lifestyles and cancer screening, and treating or identifying addiction to smoking, drugs or alcohol.

Whereas most economic evaluations study new drugs, 81% (44/54) of the factorial trials evaluated non-drug technologies. These included physiotherapy, surgery, nutritional supplements/diets, health promotion materials, talking therapies, screening and logistical interventions (e.g. procedures or incentives) targeting the health service.

The studies were also large, with >3,400 participants randomised in the average completed study. Only 13% (5/39) of completed factorial RCTs had <200 participants, vs. 43% (50/115) of all trial-based economic evaluations identified in a 2003 review (15) ($p=0.06$)[45]; the larger sample size may be a feature of factorial trials, since there was no relationship between sample size and date of publication within my review ($p=0.865$).[46]

Ten completed studies (25%) used CUA (either alone or with other forms of economic evaluation), which is similar to the proportion for all economic evaluations (19.8% (126)).

[45] Chi-squared test conducted in Excel with Yates' correction.
[46] OLS regression conducted in Stata version 12 estimating relationship between sample size and year results were published.

However, 85% (12/14) of the protocols proposed using CUA. All studies/protocols used a parallel-group, rather than crossover, design. However, 20% (11/54) were cluster-randomised; this may result from evaluation of complex interventions and those targeting healthcare providers.

Five studies/protocols were partial factorial, in that some but not all patients were randomised to >1 factor, although none used the term "partial factorial".[47] One of the protocols that I have classed as partial factorial described itself as an "double 2x2": it randomised all patients to intensive vs. standard glycaemic control targeting and also randomised the patients with moderate dyslipidaemia to fenofibrate vs. placebo, while the remaining patients were randomised to intensive vs. standard blood pressure targeting (232). I am classed the prematurely-terminated study as an incomplete 3x3 design: patients were first randomised to one of three first-line laxatives and those who were not satisfied with the response were randomised to add in another type of laxative, omitting the arms adding the same type of laxative already received (233). Another study used an incomplete 2x2x2 design, omitting the *ac* and *bc* cells from the design (234). The other 47 studies/protocols were full factorials. However, three of these full factorials included additional study arms outside the factorial design: two included an additional control group (235, 236), while one added an additional comparison that was not fully crossed with the others (237).

Three quarters of studies/protocols (41/54) randomised patients to two factors at two levels. Four studies used 2x2x2 designs with/without additional arms. Seven studies/protocols used 2x2, 2x3 or 3x3 designs with two factors and >2 levels. Two studies/protocols used 3x2x2 or 4x3x2 designs.

[47] One study describes itself as having a "2x2 factorial component" (231). The others simply describe themselves as factorial.

Four studies were identified in which the interaction between factors was confounded (238-241). Three of these studies were included in the review as the authors described them as factorial (238-240). In one study (240), GPs randomised to one of the three active treatment groups received a training session not given to the control group. Conversely, two trials gave patients in the control group an additional intervention not given to the other three groups.[48] A fourth trial, in which the second factor compared physiotherapy against reinforcement of the advice given as part of factor 1 (a whiplash book or usual advice), was excluded from the review as the authors did not describe it as factorial (241). A further study that was included in the review allowed information sharing between practitioners within the *ab* group that was not possible within the groups receiving <2 interventions (242).

With respect to the clinical analysis, only two of the studies explicitly stated that interactions were taken into account in power calculations (242, 243). However, most studies did not clearly describe how the factorial design was taken into account in power calculations.

Of the completed studies, 23% (9/40) allowed for interactions between factors when analysing the primary clinical endpoint, 53% (21/40) assumed no interaction, while 25% (10/40) did not clearly state their methods. Twenty studies (50%) used regression methods for the primary endpoint, of which five included interaction terms, seven did not and eight did not clearly describe their methods. Four studies used inside-the-table analysis and 14 used at-the-margins. Three studies (8%) found that all factors evaluated had a statistically significant effect on the primary clinical endpoint, while 63% (25/39) observed significant differences between levels for some factors, but no significant difference for others. Five studies (13%) found no significant differences in the primary clinical endpoint, while results were not clearly stated for seven studies. Only three studies (8%) observed statistically significant interactions for the primary endpoint, although nine others (23%) observed large

[48] This additional intervention comprised two social visits within the visual impairment (VIP) trial (238) and three group sessions giving advice osteoarthritis, exercise and diet in the Arthritis, Diet, and Activity Promotion Trial (ADAPT) (239).

or qualitative interactions that did not reach statistical significance or for which significance was not reported. Interaction results were not clearly reported for 15 studies.

By contrast, 53% (21/40) of completed studies allowed for interactions in their base case economic evaluation: more than twice the number allowing for interactions in the primary endpoint. Studies were also more likely to report sufficient information to identify whether interactions were taken into account for cost-effectiveness than primary endpoints. However, in most cases it was necessary to infer the methods used from the tables reported, which was frequently not possible for clinical results reported only at-the-margin. However, only five studies analysed economic results using standard regression analyses, while two used event-based CEA, 17 inside-the-table and 14 at-the-margins; this may reflect the difficulties associated with regression-based CEA discussed in Section 3.6.2.

Fifteen completed studies (38%) presented the probability of treatment being cost-effective within the text or as CEACs. Of these, nine presented pair-wise comparisons giving the probability that one treatment is cost-effective compared with a single comparator, three presented figures showing how the probability of each treatment evaluated in the trial having highest NMB varies with ceiling ratio and a further three presented CEACs for both pair-wise and multiple comparisons. Of the 25 studies that did not present the probability of treatment being cost-effective, six quantified the uncertainty around the ICER in other ways (e.g. scatter graphs or CI). Welton et al also presented the EVPI for two trials (116), as well as the expected value of *partial* perfect information (EVPPI) (115) and the EVSI (118) for one trial.

Sixteen studies (40%) reported results inside-the-table in sufficient detail that interactions for both costs and health benefits could be directly evaluated (Table 4.2). Since four of the studies were larger than 2x2 or reported results by subgroup, this gave 24 interactions for each of three outcomes (cost, QALYs and NMB). Large interactions arose frequently: 33%

(24/72) of interactions had an absolute magnitude larger than one or more simple effect (interaction:effect ratios >1 or <-1), with 25% (18/72) being qualitative (ratios <-1).

However, contrary to what was hypothesised in Section 3.4, large and qualitative interactions occurred at least as commonly for health benefits as for costs and NMB: indeed, 50% (7/14) of interactions for health benefits measured in units other than QALYs were larger than simple effects. To explore whether this finding was due to the scale of analysis, health benefits were also analysed on a logarithmic scale. Interaction:effect ratios were smaller on a logarithmic scale than on a natural scale in 50% (12/24) of cases, regardless of whether benefits were measured in QALYs or other units.

Overall, a third of interactions were super-additive (23/72), 49% (35/72) were sub-additive or qualitative, while 17% (12/72) were mixed. However, although 25% (6/24) of studies had qualitative interactions for NMB, the interaction changed the treatment adoption decision in only one case (244). This illustrates the principle described in Section 3.7.3, whereby only qualitative interactions (which by definition change the ranking of treatments) can change which treatment has highest NMB.

Six studies (reporting nine interactions) reported SDs around both costs and health benefits in each group. Within these studies, 56% (5/9) of interactions for cost were statistically significant ($p<0.05$), although there were no statistically significant interactions for health benefits or NMB.

Table 4.2: Magnitude of interactions for the 16 studies reporting mean costs and mean health benefits for each cell within the factorial design

Study	Description		Treatment with highest expected NMB		Interaction ratio† (p-value‡)		
			Ignoring interactions	Allowing for interactions	Cost	Health benefit	NMB§
Studies reporting SDs around costs and benefits for each cell							
AFIST II (245)	2x2 evaluating amiodarone vs. placebo and pacing vs. no pacing. Cost per case of atrial fibrillation avoided		Amiodarone + pacing	Amiodarone + pacing	-1.589 (0.314)*	-7.000 (0.329)*	-1.667 (0.276)*
ATEAM (246)	2x4 evaluating normal care vs. massage (M) vs. 6 lessons in Alexander technique (6A) vs. 24 lessons in the Alexander technique (24A) and exercise and behavioural counselling (E) vs. no exercise/counselling. Cost per QALY	M x E:	24 Alexander lessons + exercise	24 Alexander lessons + exercise	-0.909 (0.252)	-0.500 (0.265)*	-0.442 (0.409)*
		6A x E:			-0.476 (0.328)	-0.333 (0.579)	-0.280 (0.745)
		24A x E:			-0.489 (0.539)	0.000 (1.00)	0.110 (0.897)
Boyle (235)	2x2 evaluating group vs. individual therapy by therapist vs. assistant. Cost per 1-point change in CELF-3 total language score		Therapist, group	Therapist, group	8.385 (0.002)	-2.372 (0.704)	-2.676 (0.698)
Hollis 2007 (247)	2x3 evaluating brief vs. moderate vs. intensive telephone counselling and offer of free nicotine replacement therapy (NRT) patches vs. no offer. Cost per patient quitting smoking	NRT x Moderate:	Intensive, with NRT	Intensive, with NRT	0.225 (0.044)	0.429 (0.726)	0.554 (0.780)
		NRT x Intensive:			0.154 (0.041)	0.577 (0.561)	1.000 (0.616)
NIMH-MTA (242)	2x2 evaluating intensive medication management vs. not and multi-component behavioural treatment vs. not. Cost per 1 point change in functioning score		Medical management	Medical management	231.750 (0.005)	-1.600 (0.599)	-1.082 (0.238)*
UKBEAM (244)	2x2 evaluating spinal manipulation vs. no manipulation and exercise programme, vs. best care. Cost per QALY		**Best care + manipulation**	**Best care + manipulation + exercise**	-1.500 (0.008)	-1.471 (0.348)	-1.450 (0.595)
Studies reporting means but not SDs for costs and benefits within each cell							
BELLS (248)	2x2 evaluating prednisolone vs. placebo and aciclovir vs. placebo. Cost per QALY		Prednisolone	Prednisolone	-0.059*	-0.553*	-0.017*
Salize (240)	2x2 evaluating training and incentive for general practitioners vs. none and training and medication for medication vs. none. Cost per patient abstaining from smoking		Treatment as usual	Treatment as usual	1.690	2.125	1.681
ADAPT (239)	2x2 evaluating exercise vs. no exercise programme and dietary counselling vs. no dietary counselling. Cost per 100% change in WOMAC		Control	Control	0.200	-11.333*	0.045

Study	Description		Treatment with highest expected NMB		Interaction ratio† (p-value‡)		
			Ignoring interactions	Allowing for interactions	Cost	Health benefit	NMB§
ASCOT UK results (249)	2x2 partial factorial evaluating amlodipine (+perindopril as required) vs. atenolol (+bendroflumethiazide) and atorvastatin vs. placebo. Cost per QALY		Amlodipine + atorvastatin	Amlodipine + atorvastatin	-0.082	1.000	13.000
Cantor (250)	2x2x2 evaluating Exit interview (exit) vs. no exit interview, home interview (home) vs. no interview, small-group sessions with a social worker (group) vs. no sessions. Cost per life-year gained	Exit x Home:	Exit + group	Exit + group	-1.218	-2.511	-2.666
		Exit x Group:			-0.750	-0.679	-0.678
		Home x Group:			-2.125	-0.044	0.017
		3-way:			0.938	0.244	0.236
Brandon (251)	2x2 evaluating 8 vs. 1 mailing of 8 vs. 1 booklets to prevent relapse of smoking. Cost per 24-month smoking abstinence		8 mailings 8 booklets	8 mailings 8 booklets	-0.156	-0.667	-0.816
Barnett (252)	2x2 evaluating voucher for free methadone treatment vs. no voucher and case management vs. none. Cost per 1 unit reduction in heroin use at 3 months		Voucher	Voucher	-0.565	-0.032*	-0.003*
Saillour-Glenisson (253)	2x2 evaluating pocket card aide memoir [PAM] vs. no card and informative test request form vs. no form [form]. Inside-the-table results presented separated into 3 types of centre. Cost per test conforming with regulations	Specialist centres:	PAM+form	PAM+form	0.028	0.000	-0.013
		Small centres:	PAM	Control	0.042	-23.000	-44.269
		Large centres:	Form	Form	-0.026	-1.222	-1.188*
Bankhead (116)	2x2 evaluating screening reminder letter from GP vs. no letter and flag in notes prompting discussion of screening, vs. no flag. Cost per additional woman screened		Letter + flag	Letter + flag	-0.231	-0.514	-0.517
Richards (116)	2x2 evaluating screening reminder letter from GP vs. no letter and flag in notes prompting discussion of screening, vs. no flag. Cost per additional woman screened		Letter + flag	Letter + flag	-0.471	-0.789	-0.794
Types of interaction							
Qualitative					16.7% (4/24)	33.3% (8/24)	29.2% (7/24)
Sub-additive					41.7% (10/24)	25.0% (6/24)	25.0% (6/24)
Super-additive					37.5% (9/24)	20.8% (5/24)	33.3% (8/24)
Mixed non-qualitative*					4.2% (1/24)	12.5% (3/24)	12.5% (3/24)
Additive					0.0% (0/24)	8.3% (2/24)	0.0% (0/24)

Abbreviations: AFIST, Atrial Fibrillation Suppression Trial; ASCOT, Anglo-Scandinavian Cardiac Outcomes Trial; ADAPT, Arthritis, Diet, and Activity Promotion Trial; ATEAM, Alexander Technique, Exercise and Massage; BELLS, Bell's palsy: Early acicLovir and/or prednisolone in Scotland; CELF-3, Clinical Evaluation of Language Fundamentals; NIMH-MTA, National Institute of Mental Health Multimodal Treatment study of children with ADHD; UKBEAM, United Kingdom Back pain Exercise And Manipulation; WOMAC, Western Ontario and McMaster [Osteoarthritis Index].

* Mixed interaction: one factor increases the outcome of interest, while the other decreases it. The interaction therefore cannot be classified as either sub-additive or super-additive (Section 2.5.2).

† For two-way interactions, the interaction:effect ratio (IER_{AB}) was calculated as the ratio of the interaction (I_{AB}) divided by the simple effect of A (δ_A): i.e. as $IER_{AB} = I_{AB} / \delta_A = (\mu_0 - \mu_a - \mu_b + \mu_{ab}) / (\mu_a - \mu_0)$, where the factor defined as A is chosen such that $|\mu_a - \mu_0| < |\mu_b - \mu_0|$ in the case of non-mixed interactions and such that δ_A has the opposite sign to I_{AB} in the case of mixed interactions. For three-way interactions, the interaction:effect ratio was calculated as: $I_{ABC}[\mu_{abc} - (\mu_0 + \delta_A + \delta_B + \delta_C + I_{AB} + I_{AC} + I_{BC})] / \delta_A$, where $\delta_A < \delta_B$ and $\delta_A < \delta_C$ for non-mixed interactions and, for mixed interactions, where δ_A is the smallest of the simple effects that has the opposite sign to the interaction term.

‡ The p-value around the interaction term was calculated in Microsoft Excel using a two-tailed t-test of the ratio of mean interaction/SE around the mean (i.e. $|I_{AB} / SE_{\text{int}\,AB}|$); the t-test was conducted with $n_0 + n_a + n_b + n_{ab} - 4$ degrees of freedom. The SE for the interaction was calculated as $4 \cdot \left(\sum_{k=0,a,b,ab} (n_k - 1) \cdot s_k^2 \right)^{1/2} \div \left[2 \cdot \left(\sum_{k=0,a,b,ab} (n_k) \right)^{1/2} + \sum_{k=0,a,b,ab} (n_k) \right]$, where n_k equals the number of patients in arm k and s_k equals the SD in arm k.

§ A £20,000/QALY (or $20,000/QALY or €20,000/QALY) ceiling ratio was used for QALYs (145) and life-years gained, while a ceiling ratio of £5,000 (or $5,000 or €5,000) per unit of benefit was arbitrarily used for all other health benefits.

4.4 Discussion and conclusions

This review demonstrates that large interactions occurred commonly for clinical endpoints as well as for costs and NMB, which contradicts the hypothesis within Section 3.4. There are several possible explanations for this finding. Firstly, large, non-significant interactions could arise more commonly in clinical endpoints than previously recognised; qualitative interactions are particularly likely to arise by chance for studies observing very small treatment effects.[49] Secondly, interactions could have been substantially smaller for the primary endpoint, which was frequently not the measure used in economic evaluation. Thirdly, some of the non-qualitative interactions could be artefacts of the scale of analysis. Finally, the studies reporting costs and benefits inside-the-table may not be typical: in particular, studies may have reported results inside-the-table *because* interactions were large.

Nonetheless, the finding that 25% of all interactions were qualitative (i.e. change the ranking of treatments and render at-the-margins estimates misleading (20, 33)) is concerning, even if only one changed the treatment adoption decision at the arbitrary ceiling ratio used in this analysis. Chapter 5 evaluates the implications of this finding for the criteria that are used to determine which interactions are taken into account in the analysis. However, more generally, the analysis demonstrates the importance of reporting costs and benefits inside-the-table, to allow readers to assess the importance of interactions for themselves. This also has implications for Bayesian analyses setting informative priors on interactions: in particular, it may not be appropriate to assume that we can be 95% confident that the interaction is not qualitative, as has been done previously (115).

[49] Small clinical differences were common in this sample and only 8% of studies found both factors to have statistically significant effects on the primary endpoint. As observed previously (1), statistically significant interactions were rare for clinical endpoints (0% [0/9] here, vs. 6% [2/31] in the previous study).

Of the completed studies, 53% allowed for interactions in their base case economic evaluation, whereas only 23% considered interactions for the primary clinical endpoint.[50] This could be due to interactions being smaller for the primary clinical analysis than the endpoint used in economic evaluation, or interactions being smaller when analysed on the logarithmic scale, which may be appropriate for many clinical endpoints but not economic evaluation (Section 3.6.1). Alternatively, the increased use of inside-the-table analysis within economic evaluation could reflect economic thinking: particularly the view that inference is irrelevant (157), or that treatment-combinations should be evaluated as mutually-exclusive alternatives. Regression analysis and CEACs for multiple comparisons currently appear to be under-used within economic evaluations of factorial trials.

Combining the results of this review with a review of all RCT-based economic evaluations based published in 2003 (15) suggests that around 3% of trial-based economic evaluations use factorial trials.[51] The number of economic evaluations alongside factorial trials is increasing over time, although the number of trial-based economic evaluations is likely to have also increased. Study identification was conducted in 2010 in order to evaluate the way in which economic evaluations of factorial trials were typically conducted before the programme of work described in this thesis. Further studies meeting inclusion criteria have been published since the review was conducted (e.g. (58, 254)) and conduct and analysis of factorial trials could have changed. However, any such trends are unlikely to have changed the conclusions of this review.

[50] Similar findings were observed in a previous review of clinical endpoints from factorial trials, which found that 97% (43/44) of trials evaluated reported at-the-margins results in the text, but only 39% (17/44) reported inside-the-table results (1). A quarter of studies in the previous review (11/44) reported only at-the-margins results, did not say whether interactions were examined and did not provide data for each cell that would have allowed the reader to do inside-the-table analysis. Twenty-six studies (59%) reported testing for interactions, of which 11 used multivariate analyses, three used ANOVA, two used the Breslow-Day test for homogeneity, two used chi-squared tests and eight did not specify the methods used. All but one study reported the interaction as a p-value rather than giving the interaction term and its confidence interval. No studies reported testing for interactions in AEs.

[51] I identified four factorial trials meeting my inclusion criteria that were published in 2003, while Doshi identified 115 trial-based evaluations published in that year, suggesting that 3% (4/115) of trial-based economic evaluations use factorial trials.

There is no medical subject heading (MeSH) term for factorial trials, making systematic identification of such trials difficult. Many trial reports (particularly those focusing on the economic evaluation) do not mention the factorial design in the abstract and six studies did not use the word “factorial” anywhere in the economic evaluation paper(s). My review may have missed some studies for this reason: particularly if the abstract of the clinical paper did not mention the economic evaluation. Many of the studies that did not mention the factorial design reported results for only one factor, so are unlikely to have allowed for the other factor in the analysis; my review may therefore underestimate the proportion of studies that have ignored the potential for interactions. However, this problem is not unique to economic evaluation: in a systematic review of factorial trials, McAlister et al noted that their electronic search strategy failed to identify 24% (7/29) of the factorial trials identified in a hand search (1). Hand searching was not feasible for this review, since economic evaluation papers are spread over a huge range of journals. This problem highlights a general deficiency in published economic evaluations and the need to ensure that all studies based on factorial trials mention the study design in the abstract to facilitate future systematic reviews and give readers a full understanding of the study design.

Chapter 5: Analysis of simulated data

5.1 Introduction

As discussed in Chapter 3, all analyses of factorial trials must balance bias against inefficiency/imprecision to decide which (if any) interactions are taken into account. Ignoring interactions increases statistical power (3, 5, 47) (unless interactions are very large (5, 30, 40, 42)), but biases treatment effect estimates whenever the true interaction is non-zero (3, 4, 8, 20, 30).

In practice, researchers cannot know whether treatments genuinely interact and have only a single sample in which to decide which interactions matter and estimate treatment effects. Analysts must therefore pre-specify a decision rule or criterion that determines which interactions are included. If possible, this rule should distinguish between those interactions likely to bias treatment effect estimates and those which would simply lose statistical power.

Various different rules or criteria could be used to identify which interactions are important (Section 3.7), including: statistical inference; goodness of fit; including all interactions above a certain size; including all qualitative interactions; or excluding all unexpected interactions not hypothesised a priori, or those that do not have plausible biological explanations.

Although it is often argued (157) that statistical inference is irrelevant to decision-making (Section 3.7.1), there are good reasons to avoid conducting inefficient analyses. Firstly, inefficient analyses will overestimate the value of further information, potentially displacing spending on healthcare today with over-investment in research. Secondly, small sample sizes or inefficient analyses may mean that (by chance) the treatment with highest expected NMB in our sample is not the one that would genuinely maximise NMB in the population.

Within this chapter, I use simulated data to compare the performance of different criteria for selecting interactions. Using simulated data means that: (a) whereas for a real trial, we only see one sample, for simulated data, we can generate as many samples as we wish and see how performance varies; (b) we know the true data-generating mechanism and can compare the conclusions of each individual sample against the true answer; (c) we can vary the characteristics of the data-generating mechanism (e.g. interaction size and sample size) and see the impact on the results. However, for simplicity, simulations focused on balanced 2x2 full factorial designs with no covariates or missing data.

5.2 Methods

To reflect the trends observed in real factorial trials, simulations were based on secondary data from nine factorial RCTs, which were either identified in Chapter 4 or which comprised case studies discussed in this thesis for which results were available at the time of analysis. Secondary data from all six studies reviewed in Chapter 4 that reported SDs around means for each group were used in the simulation (235, 242, 244-247) (Appendix H).[52] For KAT, final results from the two groups of patients randomised to two comparisons were used. Preliminary one-year IVAN results (255) were used, since two-year results were not available at the time simulations were conducted.

Means and SDs for each treatment group were extracted for each study and used to estimate interaction terms and treatment effects (Appendix H, Table H1). The correlation between costs and benefits was estimated from primary data for KAT and IVAN. However, since none of the published trials reported the correlation between costs and benefits,

[52] For simplicity, the analysis focused on 2x2 matrices, so included only the first two levels for each factor evaluated in Hollis (247) and the Alexander Technique, Exercise and Massage (ATEAM) trial (246).

hypothetical correlation coefficients were estimated based on the clinical area and any scatter graphs reported.

In addition to the original studies, five variants of each trial were simulated using interaction terms that were 0%, 50% or 200% of the size observed in the original study, and using double the sample size with either the original interaction or zero interaction (Appendix H, Table H2).

Stata version 12 was used to simulate 300 samples of each of the 54 scenarios from the nine trials. Appendix H shows the Stata code used to generate simulated trials and analyse the results. Costs and QALYs were randomly drawn for each patient in each sample. For simplicity, studies were simulated as perfectly balanced, with equal patient numbers in each treatment arm.

Five trials measured health benefits in QALYs, two used functioning/language scores and two used the proportion of patients meeting endpoints. For simplicity, benefits were simulated as normal distributions,[53] allowing for the correlation between costs and benefits. However, a maximum benefit was also specified for each trial (Appendix H, Table H1), to ensure that percentages remain <100% and that no patient accrued >1 QALY per year of analysis.[54] Any patient with sampled health benefits above the ceiling for that study had benefits reduced to the ceiling. Such ceiling effects introduce a small skew into benefits and could introduce or accentuate interactions, since the treatment group with highest benefits

[53] Assuming that health benefits are continuous and normally distributed means that the data generating process does not realistically replicate the two studies with binomial benefit measures. In principle, health benefits for these studies could have been modelled as a binomial distribution correlated with the cost distribution that modelled the effect of treatment and interactions on a logarithmic scale (e.g. as odds ratios). However, this would have greatly complicated data generation and would have meant that interactions were modelled (and would need to be analysed) on a logarithmic scale rather than a natural scale. By modelling these two trials as though they measured benefits on a continuous scale, this analysis essentially treats them as though they had actually measured benefits in QALYs.

[54] For QALYs, the ceiling allowed for discounting and equalled $1/(1+r)^t$ per year of analysis.

will be truncated more than other groups (Appendix D, example P). An arbitrarily high limit of 50 units was therefore used for all scenarios assuming zero interaction.

Costs were simulated as gamma distributions defined by the mean cost for each treatment arm and its SD. Although SDs for health benefits were approximately homoskedastic across study arms, SDs for cost varied substantially between treatment arms in all but three studies.[55] For five studies, SDs were assumed to be proportional to the mean (Appendix H, Table H1). For two further studies[56] the specific pattern of heteroskedasticity was modelled by using different SD:mean ratios for patients with and without treatment B.

The costs and benefits for each sample were analysed using four mixed models with different combinations of interaction terms: no interactions; interaction for costs only; interaction for health benefits only; and interactions for costs and benefits. The mixed models implement seemingly-unrelated regression allowing for correlations between costs and benefits by predicting outcomes (which could be either costs or benefits) with random effects by patient. However, separate constants, treatment effects and (where necessary) interactions were estimated for costs and benefits and unstructured residuals were used. This approach gives identical results to the `sureg` command. The log-likelihood, degrees of freedom and coefficients and their SEs were recorded for each model.

The coefficients estimated in mixed models were used to calculate NMB. Results focus on ceiling ratios of £20,000/QALY (145) for the five studies measuring benefits in QALYs, and £5,000 (or $5,000) per unit of benefit for other studies. For simplicity, all costs were interpreted as though they were in pounds Sterling.

[55] These comprised the two parts of the KAT trial, plus the National Institute of Mental Health Multimodal Treatment study of children with ADHD (NIMH-MTA).
[56] These comprised the Atrial Fibrillation Suppression Trial (AFIST) and the Alexander Technique, Exercise and Massage (ATEAM) trial.

Fifteen criteria for identifying which interactions to include in the analysis were evaluated on each of the 300 samples of each of the 54 scenarios (Table 5.1). Criteria taking into account prior evidence/hypotheses about the size of interactions were not considered, since this information was rarely reported in the published studies.

Table 5.1. List of the criteria for determining which interactions are taken into account that are evaluated in the study.

	Name of criterion	**Explanatory notes on implementation**
1	Always include all interactions	Interactions were included in all samples. In the literature, this strategy is sometimes referred to as “never pool” (3-5).
2	Never include any interactions	No interactions were included in any samples. In the literature, this strategy is sometimes referred to as “always pool” (4, 5).
3	Include interactions where p<0.05	Interaction for cost [or benefits] was included if it was statistically significantly different from zero in the model that included interactions for both costs and benefits.
4	Include interactions where p<0.10	
5	Include interactions where p<0.25	
6	Include interactions decreasing AIC	Results based on the mixed model with lowest AIC/BIC.
7	Include interactions decreasing BIC	
8	Include qualitative interactions in cost or benefits	Includes interactions for cost [benefits] that change rankings of treatments for cost [benefits]: i.e. those that are at least as large as and have the opposite sign from one or both of the simple effects (which will have interaction:effect ratios <-1).
9	Include interactions for cost or benefits if >simple effect	As for 8, but also including any super-additive interactions for cost or benefits that are at least as large as the smaller of the two simple effects. As result, all interactions with an absolute magnitude at least as large as the smaller of the two simple effects (i.e. all those with interaction:effect ratios <-1 or >1) are included.
10	Include interactions for cost or benefits if p<0.05 or >simple effect	As for 9, but also including smaller interactions that are statistically significantly different from zero.
11	Include qualitative interactions for cost, benefits or NMB	Interactions were calculated at a series of eight ceiling ratios between £5,000 and £40,000 per unit benefit using the coefficients for all four mixed models. Interactions for costs [benefits] were included if the interaction for cost [benefits] was qualitative, or if the coefficients from the mixed model that included an interaction for cost [benefits] but not benefits [cost] produced a qualitative interaction for NMB at any of the ceiling ratios.
12	Include interactions for cost, benefit or NMB if >simple effect	Includes all qualitative interactions in cost, benefits or NMB and any super-additive interactions that are larger than the smaller of the two simple effects. Calculated as for 11, but also including large super-additive interactions.
13	Include |interactions| ≥0.25 or ≥£250	Only interactions above the designated size threshold were taken into account. For example, criterion 13 includes interactions in benefits that are ≥0.25 (or ≤-0.25) units in size and interactions in cost that are ≥£250 (or ≤-£250) in size.
14	Include |interactions| ≥0.5 or ≥£500	
15	Include |interactions| ≥1 or ≥£1,000	

The performance of each of the 15 criteria on seven measures (Table 5.2) was evaluated against the "true" results for each dataset, which (for the purposes of this simulation study) were assumed to equal the mean values for treatment effects and interactions shown in Appendix H, Tables H1-H2.

Table 5.2. The measures used to assess performance of the criteria for deciding which interactions are considered.

	Measure	**Details of calculation**	**Rationale**
1	Coverage for simple effects	The proportion of simulations in which the 95% CI around the effect of A or B from the mixed model favoured by any given criterion included the true simple effect value that was used to generate the simulated trials. The simple effect (i.e. $\mu_a - \mu_o$) was used to estimate coverage in all cases, since an analysis that excludes an interaction implicitly assumes that the at-the-margin estimate $0.5(\mu_a + \mu_{ab} - \mu_0 - \mu_b)$ equals the simple effect $\mu_a - \mu_o$.	Under a purely frequentist framework, the priority is to maximise the statistical power to detect the effect of A and B while keeping the risk of a type I error at alpha. Coverage and power assess the performance of the criterion for hypothesis testing. However, as discussed in Section 3.7.1, this is less relevant for economic evaluation. Nonetheless, coverage will (to some extent) identify analyses that are prone to bias as well as those that are inefficient.
2	Statistical power for main effects	The proportion of simulations in which the coefficient of A (or B) in the analysis chosen by any given criterion is significantly different from zero (p<0.05).	
3	Sensitivity for including non-zero interactions	The proportion of samples in which interactions in cost [benefit] were taken into account in the analysis when the true interaction was not zero	Sensitivity and specificity evaluate the extent to which criteria identify non-zero interactions, but do not reflect the consequences of ignoring them.
4	Specificity for excluding interactions equal to zero	The proportion of samples in which interactions in cost [benefit] were excluded from the analysis when the true interaction equalled zero	
5	Amount of bias in incremental effects	The average bias within estimates of the incremental costs [benefits] of A or B was estimated by multiplying the proportion of samples in which the analysis ignored interactions that were not truly zero by 50% of the absolute true interaction term within those scenarios. For example, if the true interaction were £200 and criterion X allowed for the interaction in 40% of samples: $bias = £40 = \lvert £200 \rvert \cdot 50\% \cdot 40\%$.	The treatment effects estimated within analyses excluding interactions are biased estimates of the simple effect ($\mu_a - \mu_o$), where the bias is half the size of the interaction term (3). However, large interactions may have no impact on our treatment adoption decision, whereas some small interactions may change which treatment has highest NMB.
6	Probability of adopting treatment with highest NMB	The treatment arm with highest expected NMB was identified at the ceiling ratio of interest for (a) the "true" parameters used to generate the data and (b) based on the mixed model	This focuses on the purpose of economic evaluation: namely to inform a treatment adoption decision regarding which treatment has highest expected NMB and to thereby maximise health

	Measure	Details of calculation	Rationale
		coefficients estimated on each sample. The proportion of samples in which the treatment predicted to have highest NMB (b) was the same as the "true" best treatment (a) was calculated for each scenario.	gain from the budget. It assumes inference is irrelevant to decision-making, but nonetheless acknowledges that inefficient analysis and small sample sizes may cause us to adopt the wrong treatment by chance. However, it does not take account of the consequences of making the wrong decision.
7	Opportunity cost associated with adopting a suboptimal treatment	For each sample, the opportunity cost was defined as the NMB for the "true" best treatment (a) minus the NMB for the treatment predicted to have highest NMB in that analysis of that sample (b). In both cases, NMB for each treatment was calculated using the "true" parameters used for data generation. Opportunity cost was therefore zero for all samples in which the "true" best treatment was adopted and positive in all other cases. Opportunity cost was then averaged across samples and scenarios.	This measure takes account of the opportunity cost of adopting the wrong treatment, as well as the probability of adopting the wrong treatment. It is similar to the opportunity cost of ignoring interactions (developed in Section 3.8) but is based on a contrast between the genuine best treatment and the treatment predicted to be best, rather than a comparison between two imperfect analyses on finite samples. As such, the opportunity cost estimated here takes account of situations where allowing for spurious interactions causes us to adopt the wrong treatment by chance, as well as situations where ignoring interactions biases the analysis.

I used the opportunity cost as the primary measure of which criterion works best, since it focuses on the central question of economic evaluation: namely maximising health gains from a finite budget. However, the other six measures are used as secondary measures to see whether they give the same conclusions about which method is best. In principle, the different methods may give similar answers: for example, low sensitivity and/or specificity at picking up interactions may lead to imprecise and/or biased estimates of treatment effect that may cause us to adopt the wrong treatment (leading to an opportunity cost) or reduce our statistical power to detect significant differences between treatments.

5.3 Results

The 15 criteria differed in the proportion and type of interactions that were correctly identified (Table 5.3). Other than the "always include interactions" criterion (criterion 1), including interactions where $p<0.25$ (criterion 5) and including interactions that are significant or

greater than simple effects (criterion 10) resulted in the largest number of cost interactions being included. By contrast, criteria 5 and 9-12 included the largest number of benefit interactions. In general, specificity and sensitivity were inversely proportional; measures based on information criteria or statistical significance at alpha=0.05 tended to have high specificity and low sensitivity.

Across all 54 scenarios from the nine trials, “always include” minimised the opportunity cost from adopting treatments that do not in fact maximise true NMB (Table 5.4). “Never include” performed worst, while criteria 3-7 and 15 (based on statistical significance, information criteria or large absolute thresholds) also performed poorly.

However, the criterion with lowest opportunity cost differed between trials and scenarios. As expected, “never include” was, *on average*, the best criterion for the scenarios that did not have qualitative interactions (Table 5.4), although no criteria had high opportunity costs when interactions were zero. Across the 22 scenarios with qualitative interactions, “always include” performed best, although criteria 11-13 also performed well. However, there were marked variations between individual scenarios (Appendix H, Table H3).

Across all scenarios, criterion 9 (including interactions ≥simple effects) had the highest probability of adopting the treatment that has highest true NMB (Table 5.4). “Never include” performed worst overall on this measure, but performed best in scenarios without qualitative interactions for NMB. “Always include” performed best when there were qualitative interactions. However, probabilities and the best criterion differed substantially between scenarios (not shown).

Table 5.3. Comparison of performance of the different criteria with regards specificity and sensitivity to detect interactions. The values shown in bold represent the most favourable of all criteria for this measure.

Criterion	Proportion of samples in which interactions are included: % (n)*		Sensitivity: proportion of any non-zero interactions taken into account: % (n)†		Specificity: proportion of interactions equal to 0 that are excluded: % (n)‡	
	Cost	Benefit	Cost	Benefit	Cost	Benefit
1: Always include all interactions	**100.00% (16,200)**	**100.00% (16,200)**	**100.00% (10,800)**	**100.00% (10,800)**	0.00% (0)	0.00% (0)
2: Never include any interactions	0.00% (0)	0.00% (0)	0.00% (0)	0.00% (0)	**100.00% (5,400)**	**100.00% (5,400)**
3: Include interactions where p<0.05	45.65% (7,396)	22.31% (3,614)	63.32% (6,839)	28.36% (3,063)	89.69% (4,843)	89.80% (4,849)
4: Include interactions where p<0.10	54.51% (8,830)	34.10% (5,524)	71.44% (7,716)	40.86% (4,413)	79.37% (4,286)	79.43% (4,289)
5: Include interactions where p<0.25	73.81% (11,957)	61.81% (10,014)	85.46% (9,230)	67.31% (7,270)	49.50% (2,673)	49.19% (2,656)
6:Include interactions decreasing AIC	51.00% (8,262)	29.36% (4,757)	68.24% (7,370)	35.82% (3,869)	83.48% (4,508)	83.56% (4,512)
7: Include interactions decreasing BIC	31.94% (5,174)	8.29% (1,343)	47.11% (5,088)	11.61% (1,254)	98.41% (5,314)	98.35% (5,311)
8: Include qualitative interactions in cost or benefits	41.78% (6,769)	53.73% (8,704)	49.14% (5,307)	59.29% (6,403)	72.93% (3,938)	57.39% (3,099)
9: Include interactions for cost or benefits if >simple effect	55.06% (8,919)	63.00% (10,206)	64.82% (7,001)	68.04% (7,348)	64.48% (3,482)	47.07% (2,542)
10: Include interactions for cost or benefits if p<0.05 or >simple effect	67.62% (10,954)	64.25% (10,409)	82.11% (8,868)	69.49% (7,505)	61.37% (3,314)	46.22% (2,496)
11: Include qualitative interactions for cost, benefits or NMB	52.42% (8,492)	67.75% (10,975)	59.13% (6,386)	71.93% (7,768)	61.00% (3,294)	40.61% (2,193)
12: Include interactions for cost, benefit or NMB if >simple effect	63.51% (10,288)	74.78% (12,115)	71.84% (7,759)	78.57% (8,486)	53.17% (2,871)	32.80% (1,771)
13: Include interactions ≥0.25 or ≥£250	53.90% (8,731)	35.32% (5,722)	61.89% (6,684)	37.83% (4,086)	62.09% (3,353)	69.70% (3,764)
14: Include interactions ≥0.5 or ≥£500	42.35% (6,861)	25.75% (4,172)	49.72% (5,370)	28.35% (3,062)	72.39% (3,909)	79.44% (4,290)
15: Include interactions ≥1 or ≥£1,000	32.82% (5,317)	16.88% (2,735)	39.93% (4,312)	19.73% (2,131)	81.39% (4,395)	88.81% (4,796)

* Percentages are out of 16,200 (300 samples of 9 trials, each with six scenarios).

† Percentages are out of 10,800 (300 samples of 9 trials, each with four scenarios with non-zero interactions).

‡ Percentages are out of 5,400 (300 samples of 9 trials, each with two scenarios with zero interactions).

Table 5.4. Comparison of performance of the different criteria with regards the probability and the opportunity cost associated with adopting a treatment that does not have highest true NMB. The values shown in bold represent the most favourable of all criteria for this measure.

Criterion	Opportunity cost of adopting a suboptimal treatment					Probability of adopting treatment with highest NMB				
	All datasets	9 original studies only	Zero interaction	Non-zero non-qualitative interaction	Qualitative interaction	All datasets	9 original studies only	Zero interaction	Non-zero non-qualitative interaction	Qualitative interaction
Number of datasets included	**54**	**9**	**18**	**14**	**22**	**54**	**9**	**18**	**14**	**22**
1: Always include all interactions	**£816**	**£1,107**	£887	£128	**£1,195**	78.85%	74.07%	78.61%	84.24%	**75.62%**
2: Never include any interactions	£1,514	£2,226	**£627**	**£59**	£3,165	73.52%	64.11%	**84.59%**	**92.83%**	52.17%
3: Include interactions where p<0.05	£1,010	£1,644	£695	£79	£1,861	77.94%	68.63%	82.85%	88.57%	67.15%
4: Include interactions where p<0.10	£936	£1,472	£764	£98	£1,609	78.44%	70.63%	81.20%	87.19%	70.62%
5: Include interactions where p<0.25	£864	£1,255	£871	£125	£1,328	78.56%	72.63%	79.00%	84.83%	74.21%
6:Include interactions decreasing AIC	£961	£1,547	£738	£84	£1,701	78.28%	69.56%	81.81%	87.81%	69.33%
7: Include interactions decreasing BIC	£1,193	£1,941	£650	£65	£2,354	76.06%	65.19%	84.13%	90.21%	60.45%
8: Include qualitative interactions in cost or benefits	£842	£1,178	£855	£121	£1,290	79.96%	76.07%	80.93%	90.19%	72.67%
9: Include interactions for cost or benefits if >simple effect	£829	£1,155	£867	£119	£1,250	**80.94%**	**77.15%**	80.91%	90.50%	74.89%
10: Include interactions for cost or benefits if p<0.05 or >simple effect	£830	£1,158	£870	£124	£1,246	79.70%	74.00%	80.59%	86.02%	74.95%
11: Include qualitative interactions for cost, benefits or NMB	£826	£1,133	£876	£129	£1,228	78.94%	74.33%	78.87%	85.17%	75.05%
12: Include interactions for cost, benefit or NMB if >simple effect	£826	£1,129	£882	£128	£1,224	79.09%	74.48%	78.80%	85.21%	75.44%
13: Include interactions ≥0.25 or ≥£250	£820	£1,138	£871	£109	£1,230	79.12%	73.74%	81.15%	89.79%	70.68%
14: Include interactions ≥0.5 or ≥£500	£848	£1,211	£860	£105	£1,312	79.71%	73.63%	81.43%	90.12%	71.68%
15: Include interactions ≥1 or ≥£1,000	£931	£1,365	£774	£108	£1,582	79.31%	72.15%	82.61%	89.95%	69.83%

Doubling sample size reduced the opportunity cost and the probability of adopting the wrong treatment for all criteria. However, *relative to other criteria*, the performance of criteria based on statistical significance or information criteria (which explicitly take account of sample size) did not appear to be any better in larger studies. Furthermore, criterion 11 (including qualitative interactions for cost, benefits or NMB) performed best in scenarios with double the original sample size, whereas “always include” performed best with a smaller sample size.

Averaged across all trials, coverage was <95% for all criteria (Table 5.5). However, some trials had a >95% coverage; the explanation for this is unclear, although it could result from the gamma-distributed costs and the ceiling effect for benefits. Coverage was markedly reduced for criteria that excluded some/all interactions, since it was defined as the proportion of samples in which the 95% CI included the true *simple effect* ($\mu_a - \mu_o$). For analyses excluding interactions, treatment effects are estimated as the effect of A averaged *across patients with and without B*; the 95% CI around this *main effect* may exclude the true simple effect due to either omitted-interaction bias or a failure of distributional assumptions. As result, coverage is affected by both bias and inefficiency and was directly proportional to the proportion of samples in which interactions were included.

On average, the nine original trials had 66% (range: 9-100%) power to detect the main, at-the-margins treatment effect for cost, compared with 37% (range: 8-82%) power for benefits. Although power was, on average, lower for "always include" than for "never include", some criteria (e.g. 8 or 15) had higher power than "never include" (Table 5.5) because pooling treatment arms despite large interactions reduces power (5, 30, 40, 42).

Table 5.5. Comparison of performance of the different criteria with regards coverage, power and bias. The values shown in bold represent the most favourable of all criteria for this measure.

Criterion	Coverage for effects of A and B: % (n)*		Power for effects of A and B: % (n)†		Bias (0.5*ignored interactions): mean (range across scenarios)‡	
	Cost	Benefit	Cost	Benefit	Cost	Benefit
1: Always include all interactions	**94.04% (30,468)**	**94.46% (30,606)**	62.72% (20,322)	35.06% (11,358)	**£0 (£0, £0)**	**0.00 (0.00, 0.00)**
2: Never include any interactions	59.85% (19,392)	83.91% (27,187)	67.62% (21,909)	43.66% (14,145)	£767 (£0, £10,870)	0.18 (0.00, 2.04)
3: Include interactions where p<0.05	88.99% (28,834)	87.33% (28,296)	66.41% (21,516)	44.40% (14,385)	£289 (£0, £3,587)	0.11 (0.00, 1.11)
4: Include interactions where p<0.10	90.17% (29,216)	88.58% (28,699)	66.18% (21,443)	43.45% (14,077)	£215 (£0, £2,808)	0.08 (0.00, 0.78)
5: Include interactions where p<0.25	92.13% (29,850)	91.22% (29,556)	65.07% (21,084)	40.58% (13,147)	£104 (£0, £1,522)	0.04 (0.00, 0.38)
6:Include interactions decreasing AIC	89.64% (29,042)	88.02% (28,519)	66.42% (21,520)	44.03% (14,267)	£253 (£0, £3,189)	0.09 (0.00, 0.90)
7: Include interactions decreasing BIC	84.58% (27,405)	85.61% (27,737)	66.23% (21,458)	44.69% (14,480)	£442 (£0, £6,377)	0.14 (0.00, 1.54)
8: Include qualitative interactions in cost or benefits	73.96% (23,962)	89.40% (28,966)	**70.29% (22,775)**	44.19% (14,316)	£240 (£0, £2,935)	0.04 (0.00, 0.32)
9: Include interactions for cost or benefits if >simple effect	82.14% (26,612)	91.28% (29,574)	64.87% (21,017)	40.73% (13,198)	£158 (£0, £1,431)	0.03 (0.00, 0.30)
10: Include interactions for cost or benefits if p<0.05 or >simple effect	91.37% (29,605)	91.37% (29,604)	64.76% (20,983)	40.07% (12,983)	£88 (£0, £1,196)	0.03 (0.00, 0.30)
11: Include qualitative interactions for cost, benefits or NMB	77.65% (25,159)	90.91% (29,454)	68.61% (22,231)	42.16% (13,659)	£193 (£0, £2,935)	0.03 (0.00, 0.21)
12: Include interactions for cost, benefit or NMB if >simple effect	84.08% (27,241)	92.34% (29,918)	64.46% (20,884)	39.36% (12,752)	£122 (£0, £1,025)	0.02 (0.00, 0.20)
13: Include interactions ≥0.25 or ≥£250	81.07% (26,268)	88.85% (28,787)	64.41% (20,868)	43.25% (14,014)	£23 (£0, £79)	0.02 (0.00, 0.08)
14: Include interactions ≥0.5 or ≥£500	75.67% (24,517)	87.88% (28,472)	64.61% (20,933)	44.61% (14,454)	£44 (£0, £180)	0.04 (0.00, 0.15)
15: Include interactions ≥1 or ≥£1,000	72.74% (23,569)	86.78% (28,116)	65.96% (21,371)	**45.01% (14,584)**	£75 (£0, £303)	0.06 (0.00, 0.29)

* Coverage = proportion of simulations in which the 95% CI around the effect of A or B from the mixed model favoured by any given criterion included the true simple effect value that was used to generate the simulated trials. Coverage is shown over the effects of A and B combined and is out of 32,400 (two treatment effects for each of 300 samples of 54 scenarios).

† Power = proportion of simulations in which the coefficient of A (or B) in the analysis chosen by any given criterion is significantly different from zero (p<0.05). Power is shown over the effects of A and B combined and is out of 32,400 (two treatment effects for each of 300 samples of 54 scenarios).

‡ Average bias within the incremental costs [benefits] = the proportion of samples in which the analysis ignored interactions that were not truly zero, multiplied by 50% of the absolute true interaction term within those scenarios. For example, if the true interaction were £200 and criterion X allowed for the interaction in 40% of samples, $bias = £40 = |£200| \cdot 50\% \cdot 40\%$.

Bias from omitted interactions was substantial for many trials and criteria, but was completely avoided by including all interactions (Table 5.5). As expected, all other criteria produced some bias, although bias was nearly twice as large for "never include" than the second worst criterion (BIC). With the exception of “always include”, the three criteria including interactions above a certain absolute size had lowest absolute mean bias.

5.4 Discussion

The analysis confirms the prediction proposed in Section 3.7.2 that for economic analyses informing treatment adoption decisions, it is better to include interactions that arose by chance than to risk ignoring genuine interactions that could change the conclusions. Across all 54 scenarios, a strategy of including all interactions regardless of size or significance minimised the opportunity cost associated with adopting the wrong treatment. Excluding all interactions, or using information criteria or statistical significance generally performed poorly on the measures most relevant to economic evaluation. However, the best criterion differed substantially depending on the way in which criteria were evaluated. For example, the probability of adopting the treatment with the highest NMB was slightly higher for criterion 9 (including interactions larger than simple effects) than for “always include”. There were also substantial variations in the relative performance of different criteria between trials and scenarios. In particular, criteria that excluded most interactions performed well in scenarios where interactions equalled zero or did not change the ranking of treatments.

I found no evidence that the most appropriate criterion varied with sample size and the best-performing criteria take no account of sample size. This finding was unexpected. If we had an infinitely large sample, there would be no benefits from increasing statistical efficiency and we would only make the wrong treatment adoption decision if systematic bias caused a treatment that is not genuinely the best to have highest NMB in our sample. In this case, it

would always be best to include all interaction terms and the sample mean will always give the “right” answer. By contrast, in a very small sample, there is a high risk that we could draw the wrong conclusions by chance and it may be at least as important to avoid statistical efficiency as systematic bias. This unexpected finding does not appear to result from my simulating unusually large trials: studies randomised between 31 and 1,586 patients per factor-combination and five of the nine trials randomised <50 patients per factor-combination. It would therefore appear that avoiding bias is more important than avoiding inefficiency even when sample size is limited.

This study evaluated a convenience sample of nine factorial trials and a small range of variants on each study. Since the most appropriate criterion differs between studies, different results could have been obtained with a different set of trials or scenarios. The analysis focused on 2x2 full factorial trials and interactions between two factors. Although the same principles are likely to apply to larger factorial designs, ***higher-order interactions*** may be harder to detect. Furthermore, all trials were simulated and analysed as though they measured health benefits on continuous scales and all costs and health benefits were analysed on a natural scale. In principle, the criteria could perform differently if outcomes were analysed on a logarithmic scale.

The data generating mechanism assumed that all trials had equal numbers in each arm, that costs are gamma-distributed with predictable patterns of heteroskedasticity and that health benefits are continuous, Gaussian (subject to ceiling effects) and homoskedastic. The mixed models and the criteria based on statistical significance may perform less well in real trials where these idealised data characteristics do not apply. In addition to being sensitive to the study characteristics varied in this analysis, the optimal choice of criteria may also be sensitive to the features common to all simulated datasets. Unlike the simulated data, real datasets also commonly have missing data and/or or censoring, require extrapolation, have imbalance in baseline characteristics or require exploration of heterogeneity, necessitating

more complicated analytical methods that are explored within the three applied case studies (Chapters 6-8).

In order to approximate the QALY distributions seen in real trials, health benefits for all scenarios other than those assuming no interaction were truncated at a maximum level relevant to each trial. Such ceiling effects reduce the means and SDs for the simulated samples to values less than the secondary data on which they are based and reduce means/SDs more for the treatment arms with highest mean benefits. Such ceiling effects will not change which treatment has highest health benefits or NMB in each sample and will therefore not bias estimates of the probability or opportunity cost of adopting the wrong treatment. However, ceiling effects will mean that the average size of the interaction term across all 300 samples will be systematically different from the values shown in Appendix H. As result, mixed models estimated that the interactions for QALYs were, on average 9.8% lower than the data inputs would suggest; as expected, the discrepancy was greatest for the 20 scenarios where the ceiling effect was <2 SDs from the largest group mean. Ceiling effects will also introduce a systematic error into coverage estimates. Nonetheless, this is unlikely to affect the conclusions since coverage is shown only for information and ceiling effects were not used for the scenarios assuming no interaction.

This analysis suggests that for the average trial, it is generally appropriate to take account of all interactions between factors when conducting economic evaluations informing treatment adoption decisions. However, results varied between trials. Furthermore, the opportunity cost of adopting the wrong treatment may not be the only consideration: risk-averse decision-makers may take account of the probability of adopting the wrong treatment, while those making joint decisions about adoption and future research may be concerned that taking account of all interactions will increase value of information estimates. In such settings, it may therefore be useful to pre-specify criteria that exclude small interactions that cannot

change the conclusions of the analysis in situations where treatments are *expected* to have additive effects: particularly if the sample size is small.

Any prior evidence or beliefs about the size of interactions could be used to select the appropriate criteria or as informative priors in a Bayesian analysis. In particular, including all interactions above a certain size appeared to perform well at minimising bias and opportunity costs; this criterion may work even better if the threshold is based on the treatment effects expected or the amount of bias that is acceptable in a particular setting. However, as for all statistical analyses, criteria must be pre-specified before data analysis: particularly since results are highly sensitive to the criteria used.

Chapter 6: IVAN – A full 2x2 factorial trial

6.1 Introduction

This chapter applies the methods and principles for economic evaluation of factorial trials discussed in Chapter 3 to "real-world" patient-level data from the IVAN trial (56, 57). IVAN evaluates two treatment regimens of two treatments, which raises additional issues over and above those faced by those factorial trials that compare two drugs against their respective placebos. In particular, as I discuss in Section 6.3, comparisons between treatment regimens introduce large super-additive interactions for drug costs. Furthermore, within IVAN, interactions and treatment effects were expected to be larger for economic endpoints than for clinical outcomes, since drug costs are expected to differ substantially between treatments and neither drug nor treatment regimen was expected to affect clinical outcomes in this non-inferiority study. Unlike the simulated data analysed in Chapter 5, the IVAN data included loss to follow-up, missing data and imbalance in baseline utilities, which necessitated a more complex data analysis.

6.2 Case study

IVAN comprises a factorial RCT comparing two treatment regimens of Avastin (bevacizumab) and Lucentis (ranibizumab) in age-related macular degeneration: continuous treatment (monthly injections for two years) vs. discontinuous treatment (giving three monthly injections then monitoring outcomes and giving further courses of three injections to any patients showing disease progression or worsening vision) (56, 57) (Table 6.1). The trial recruited 610 patients and was powered to test whether Avastin is non-inferior to Lucentis and whether discontinuous treatment is non-inferior to continuous therapy for visual acuity at two years (assuming no interaction between drug and treatment regimen).

Table 6.1. Design of the IVAN trial

		Factor 1: Drug	
		Lucentis	**Avastin**
Factor 2: Treatment regimen	**Continuous treatment**	Continuous Lucentis	Continuous Avastin
	Discontinuous treatment	Discontinuous Lucentis	Discontinuous Avastin

A within-trial economic evaluation taking a two-year time horizon was conducted alongside the assessment of clinical outcomes to evaluate costs and cost-effectiveness from an NHS perspective (Appendix I). Missing utility data were imputed using multiple imputation (256) to generate 100 imputed datasets using a model that included dummy variables for each factor and an interaction term. Resource use data and unit costs were combined to estimate quarterly costs of: Avastin/Lucentis; drug administration and ocular monitoring; and hospitalisations and ambulatory consultations for serious adverse events (SAEs) or AEs previously linked with Avastin/Lucentis treatment, and changes in medications licensed for treatment-related AEs/SAEs.

6.3 Analytical methods for dealing with factorial design

Since IVAN is factorial, it is important to consider the potential for interactions between factors: i.e. to examine whether the difference in costs, QALYs or NMB between Avastin and Lucentis is affected by the treatment regimen used. In general, interactions rarely occur between factors that have little/no effect on outcomes (30, 35). Since IVAN is designed to demonstrate non-inferiority and no difference in efficacy is expected, we would a priori therefore not expect any interaction between drug and treatment regimen for QALYs or for non-drug costs unless the number of re-treatments differed between Avastin and Lucentis and/or differences in efficacy/safety were observed.

However, we would expect a large interaction for drug costs, as reducing treatment frequency will have a proportionately greater effect for Lucentis than for the less costly Avastin, as treatment regimen and drug may have multiplicative (rather than additive) effects.

For example, if Avastin cost £49/dose, Lucentis cost £742.17/dose (257) and the average patient receiving discontinuous treatment received 10 doses over the two-year trial period while continuous treatment involves 24 injections, the interaction would equal £9,704 (17812-7422-1176+490). This indicates that the cost of continuous Lucentis is £9,704 higher than what we would expect if the effect of drug and treatment regimen were additive.

Ignoring the interaction by presenting results for continuous vs. discontinuous treatment at-the-margins would therefore be misleading and introduce bias (3, 4, 8, 20, 30). For example, an at-the-margins estimate of the incremental cost of giving continuous rather than discontinuous treatment would be based on an “average” drug cost of £396/dose, systematically overestimating incremental costs for Avastin and underestimating those for Lucentis. Conversely, allowing for interactions when treatment effects are actually additive has been shown to increase SEs and reduce statistical power (3, 5, 47). Although biased point estimates will distort decisions based on economic evaluation more than inflated SEs (Chapter 3), inefficient statistical analysis would nonetheless bias estimates of decision uncertainty, overestimating value of information and biasing CEACs towards 50% when two treatment combinations are compared and towards 25% when four treatment combinations are under consideration.

The anticipated interaction for drug costs means that conclusions about which drug has lower costs or higher NMB cannot be made independently of conclusions about treatment regimen. We must therefore treat the four combinations of drug and treatment regimen as mutually-exclusive strategies and look inside-the-table when drawing conclusions and deciding which strategy(ies) to implement.

However, although we need to *interpret* results inside-the-table, we do not necessarily need to allow for interactions in those components of NMB for which the effect of drug and treatment regimen is additive when *estimating* the costs and QALYs for each treatment arm.

By breaking down NMB into four *components* (QALYs, drug costs, administration/monitoring costs and medication/medical service use costs) we can therefore avoid bias by allowing for interactions when analysing those components where interactions are important, but maximise statistical power by assuming additive effects for those components where interactions are unimportant.

It is therefore critical to make appropriate, pre-specified decisions about *which* interactions are genuine and which are spurious. Collaborators were keen to exclude unimportant interactions to increase statistical efficiency and avoid having conclusions driven by small, chance interactions. Like many trials, the IVAN clinical analysis specified that only interactions significant at the 0.05 level would be taken into account. However, this is inappropriate for economic evaluation, where interpretation is based on the magnitude of ICERs and expected NMB, rather than statistical significance (157), and where the outcomes of interest (costs and QALYs) are often under-powered (139) and more likely to show large interactions (Section 3.4). As a result, large or qualitative interactions that could change conclusions may not be statistically significant.

The base case analysis therefore included all interactions that were *either* statistically significant *or* which had an absolute magnitude larger than either main effect. Since inference is arguably irrelevant to decisions about adoption today (157), I included qualitative interactions and large super-additive interactions since these could change the conclusions of the analysis regardless of statistical significance. No interaction terms or indicators of treatment regimen were included for costs or QALYs in the first quarter (Q1) as all patients meet the re-treatment criteria at baseline and therefore receive three monthly injections. However, for each component of NMB, outcomes in Q2-8 were analysed in the same way, as there is no logical reason why interactions would exist for some but not all quarters.

The correct analysis of trial-based economic evaluations should allow for correlations between costs and QALYs and their typically skewed distributions (Section 3.6). As I discussed in Section 3.6.2.4, correlations between costs and QALYs can be dealt with using various methods, including mixed models, seemingly-unrelated regression, MCMC and bootstrapping, which give very similar results and correctly allow for the correlations between costs and QALYs. Bootstrapping is particularly convenient when we need to subdivide costs and QALYs into components that must be analysed with different assumptions about interactions and also facilitates KMSA and combining multiple imputed datasets.

I therefore conducted the following steps on each of 130 bootstrap replicates[57] drawn independently from each of the 100 imputed datasets, giving a total of 13,000 bootstrap replicates. Point estimates were obtained by conducting each step using the original sample of each imputed dataset (without bootstrapping):

a) OLS regression was conducted separately on QALYs, drug costs, administration/monitoring costs and medication/medical service use costs for each quarter, excluding patients who died before the start of the quarter or dropped out before the end of the quarter. OLS was used despite the skewness, severe heteroskedasticity and non-linear effects seen for some variables, since it is computationally much faster and facilitates inclusion/exclusion of interactions. However, SEs were based on SDs across bootstrap replicates rather than regression output. In a sensitivity analysis, drug costs were analysed using GLM with a logarithmic link function to evaluate whether this eliminated the interaction.

   For quarters 2-8, explanatory variables included one dummy indicating treatment regimen and another indicating drug allocation; such models were estimated with and without

[57] 130 bootstrap replicates were drawn per imputed dataset since this provided >10,000 independent bootstraps across all datasets and simulated in a convenient length of time.

interaction terms equal to drug multiplied by treatment regimen to evaluate the effect of changing the methodology and evaluate the statistical significance and magnitude of interactions. For Q1, only the dummy for drug was included in the base case analysis, since all patients received monthly injections, although one analysis analysing Q1 data in the same way as the base case analysis and another including all interactions were also conducted. Regressions on quarterly QALYs also controlled for baseline EQ-5D utility, since the marked imbalance in mean baseline utilities between groups could otherwise have introduced bias (150). Regression coefficients and the log-likelihood were recorded for each bootstrap replicate on each regression model.

b) I used KMSA to allow for patients withdrawing from the trial. This technique was adapted to prevent chance differences in mortality unrelated to treatment affecting incremental QALYs, by assuming that deaths unrelated to treatment occurred at the same rate in all four study arms. Adjusted Kaplan-Meier estimates of the probability of being alive at the start of each quarter were estimated for each of the four treatment arms. The probability of dying in each quarter was assumed to equal the probability of deaths definitely/probably/possibly related to study medication (referred to hereafter as treatment-related deaths) plus the probability of deaths unlikely to be/not related to treatment (referred to hereafter as unrelated deaths).

Two Kaplan-Meier estimates were calculated: one allowing for interactions in treatment-related deaths and one ignoring such interactions. In both analyses, the probability of unrelated deaths in each quarter was calculated as the total number of deaths unlikely to be/not related to study medication (across all four treatment arms) divided by the total number of people alive and non-censored at the start of that quarter. The analysis allowing for interactions for deaths estimated the probability of treatment-related deaths in quarter q as the proportion of patients *in that treatment arm* who had treatment-related deaths that quarter; this method was used in the base case since it is less complicated.

The sensitivity analysis excluding interactions for treatment-related deaths used logistic regression without interaction terms and simple mathematical rules to estimate the probability of treatment-related deaths.[58]

c) Predicted quarterly costs and QALYs in each arm were calculated from each regression, multiplied by the probability of being alive at the start of each quarter and summed over the four quarters, discounting values in Year 2 at 3.5% (145). These calculations were replicated for all 32 (2^5) combinations of assumptions about interactions, in addition to two analyses allowing for the effect of interactions and assignment to discontinuous therapy within Q1 and one assuming no interaction for drug costs on a natural (rather than logarithmic) scale.

Analyses were conducted in Stata version 12. Analyses with and without interactions were conducted on the same set of bootstrap replicates to eliminate chance variations between the sets of results presented. Results across the 100 imputed datasets were pooled using Rubin's rule (258) in Microsoft Excel 2007; the resulting pooled means and SEs were used to estimate CIs and p-values. For each model specification and each bootstrap replicate, AIC was calculated from the log-likelihood as 2k-2loglikelihood. AIC was summed across all

[58] Since only seven IVAN participants had treatment-related deaths before either withdrawing from the study or attending visit 24, all bootstrap replicates had at least one quarter in which one or more treatment arms had no deaths, which frequently meant that logistic regression models could not be estimated without perfect prediction. Continuity corrections (e.g. adding 0.5 deaths to each cell) were explored but were not feasible since the resulting mortality rates were extremely sensitive to the magnitude of the arbitrary correction factor. To get around this problem, the pattern of deaths across treatment arms within each quarter of each bootstrap replicate was categorised into one of five types. Firstly, if no patients died in any treatment arm in that quarter, the probability of treatment-related deaths was zero for all treatment arms. Secondly, if three of the four treatment arms had no treatment-related deaths that quarter, the number of treatment-related deaths was halved in the arm with deaths, while the number of deaths in the two adjacent cells was assumed to be a quarter of the number of observed in the cell with deaths. Thirdly, if treatment-related deaths were observed in both continuous groups but neither discontinuous group (or vice versa), or if deaths were observed in both Avastin groups, but neither Lucentis group (or vice versa), the probability of treatment-related deaths in the two cells with deaths was assumed to be the average across those cells, while the number of deaths in the other two cells was assumed to be zero. For all other mortality patterns, the probability of treatment-related deaths was estimated using logistic regression with dummies indicating drug and treatment regimen.

quarters for each model specification and then averaged across bootstrap replicates and across imputed datasets to give the mean total AIC for that specification.[59]

The interactions included in the base case analysis were decided by examining the total QALYs, drug cost, administration/monitoring cost and medication/medical service use cost that were estimated in the analysis including all interactions; interactions that were either statistically significant at the 0.05 level or which were larger than either main effect were included. For simplicity, the base case analysis included interactions for mortality in all quarters.

The conclusions of all 35 combinations of assumptions about interactions were evaluated. However, detailed results are presented for four analyses:

1) Base case;
2) Assuming additive effects for all NMB components;
3) Allowing for multiplicative effects with no interaction for drug cost and additive effects for all other components;
4) Including all interactions for all NMB components, including in Q1.

[59] The approach used to calculate AIC from multiply imputed data and a complex series of bootstrap models has several major shortcomings. Firstly, it does not allow for the impact of mortality on model fit (since AIC is calculated on models estimated specifically for patients who are alive and non-censored in that quarter) and also cannot be used to compare models of mortality (because a full set of treatment dummies and interaction terms will perfectly predict mortality given the small number of deaths observed in IVAN). Secondly, it is not valid to apply Rubin's rule to AIC (256) and it can be argued that the missing data does not have a likelihood function; a recent paper developed methods for estimating the DIC given missing data (259), although I am aware of no comparable methods for AIC. However, this method does ensure that AIC is calculated on the same dataset and using exactly the same models as used in other analyses. Furthermore, alternative methodologies, such as estimating mixed models specifically to compare model specifications (Chapter 7), do not exactly replicate the models used in bootstrapping. Additionally, if AIC were to be estimated solely on those patient-quarters with complete data, it would be based on a different dataset from the main analyses: a dataset that would be more likely to exclude patients' last months of life and quarters in which SAEs occurred. Due to the limitations of alternative methods, I therefore averaged AIC across all bootstrap replicates and all datasets and also estimated the proportion of datasets in which the model with interactions had lower AIC.

EVPI was estimated as the mean maximum NMB for each bootstrap replicate minus the expected NMB for the treatment that would be adopted based on point estimates. For simplicity, EVPI calculations excluded evidence from other trials. EVSI and population EVPI were not calculated here (or in Chapter 7) as the aim was to explore the effect of interactions on EVPI, rather than make research recommendations. CEACs were estimated as the proportion of bootstrap replicates (across all 100 datasets) where each treatment maximised NMB. Base case conclusions assume a £20,000/QALY ceiling ratio (130). The methods described in Section 3.8 were used to calculate the opportunity cost of ignoring interactions under current and perfect information from the mean NMB for the original sample and for each bootstrap replicate.

6.4 Results

6.4.1 Identifying important interactions

The analysis including interactions for all quarters and all components of NMB (Analysis 4) suggested that there were no statistically significant interactions between drug and treatment regimen for any outcome other than drug cost, where large super-additive interactions were observed in Q2-8 ($p<0.001$; Table 6.2). For drug cost, the magnitude of the interaction term was markedly larger than the difference between continuous and discontinuous Avastin and that between continuous and discontinuous treatment when averaged over both drugs. The interactions for drug cost therefore met both of my criteria for inclusion in the base case, being statistically significant and larger than $\geq$1 main effect.

For administration and monitoring costs, interactions were non-significant ($p\geq0.697$) and substantially smaller than the main effects of drug or treatment regimen; drug and treatment regimen were therefore assumed to have additive effects on administration and monitoring costs.

Table 6.2. Mean results for each treatment arm including interactions for all outcomes and all quarters. Results were calculated from Analysis 4 (including all interactions).

	Quarter	Mean (95% CI)				Interaction (SE)
		Continuous Lucentis (n=158)	Discontinuous Lucentis (n=156)	Continuous Avastin (n=150)	Discontinuous Avastin (n=146)	
Drug cost	1	£2,189 (£15)	£2,207 (£9)	£144 (£1)	£145 (£1)	-£18 (-£52, £17)
	2	£2,113 (£23)	£953 (£69)	£137 (£2)	£78 (£5)	£1,101 (£957, £1,245)*
	3	£2,097 (£27)	£1,164 (£66)	£137 (£2)	£80 (£5)	£877 (£736, £1,017)*
	4	£2,110 (£26)	£890 (£69)	£136 (£2)	£77 (£5)	£1,160 (£1,016, £1,304)*
	5	£1,994 (£31)	£1,084 (£68)	£128 (£2)	£75 (£5)	£857 (£710, £1,003)*
	6	£1,927 (£38)	£997 (£65)	£128 (£2)	£64 (£5)	£866 (£721, £1,011)*
	7	£1,936 (£34)	£1,018 (£66)	£126 (£2)	£68 (£5)	£859 (£716, £1,002)*
	8	£1,912 (£41)	£926 (£69)	£128 (£2)	£65 (£5)	£924 (£768, £1,081)*
	TOTAL	£16,277 (£141)	£9,239 (£329)	£1,064 (£8)	£652 (£24)	£6,626 (£5,930, £7,321)*
Administration & monitoring cost	1	£296 (£6)	£298 (£6)	£295 (£7)	£296 (£7)	£0 (-£25, £24)
	2	£251 (£6)	£246 (£12)	£246 (£7)	£239 (£12)	-£3 (-£40, £35)
	3	£250 (£6)	£224 (£10)	£246 (£7)	£227 (£11)	£6 (-£29, £41)
	4	£251 (£6)	£220 (£11)	£245 (£7)	£221 (£11)	£7 (-£29, £44)
	5	£238 (£7)	£221 (£11)	£230 (£7)	£215 (£11)	£2 (-£33, £37)
	6	£229 (£7)	£212 (£10)	£230 (£7)	£209 (£12)	-£4 (-£40, £32)
	7	£232 (£7)	£208 (£11)	£228 (£7)	£210 (£11)	£5 (-£30, £40)
	8	£226 (£7)	£209 (£11)	£230 (£7)	£209 (£12)	-£4 (-£41, £32)
	TOTAL	£1,970 (£45)	£1,838 (£72)	£1,950 (£47)	£1,826 (£77)	£9 (-£232, £250)
Medications & medical service cost	1	£43 (£13)	£27 (£16)	£62 (£26)	£19 (£5)	-£26 (-£91, £39)
	2	£31 (£9)	£72 (£23)	£106 (£84)	£36 (£9)	-£111 (-£283, £61)
	3	£18 (£4)	£102 (£71)	£21 (£4)	£42 (£18)	-£63 (-£208, £81)
	4	£34 (£8)	£26 (£6)	£52 (£32)	£49 (£24)	£4 (-£76, £84)
	5	£75 (£34)	£23 (£4)	£45 (£21)	£22 (£5)	£29 (-£51, £109)
	6	£58 (£25)	£81 (£45)	£32 (£8)	£230 (£186)	£175 (-£204, £554)
	7	£47 (£12)	£55 (£18)	£223 (£133)	£74 (£49)	-£157 (-£441, £127)
	8	£36 (£8)	£38 (£8)	£65 (£29)	£32 (£13)	-£34 (-£100, £32)
	TOTAL	£342 (£62)	£424 (£92)	£606 (£178)	£505 (£195)	-£183 (-£739, £373)

	Quarter	Mean (95% CI)				Interaction (SE)
		Continuous Lucentis (n=158)	Discontinuous Lucentis (n=156)	Continuous Avastin (n=150)	Discontinuous Avastin (n=146)	
QALYs	1	0.206 (0.002)	0.207 (0.002)	0.206 (0.002)	0.204 (0.003)	-0.003 (-0.010, 0.004)
	2	0.207 (0.003)	0.209 (0.004)	0.209 (0.003)	0.205 (0.004)	-0.006 (-0.019, 0.006)
	3	0.205 (0.004)	0.206 (0.004)	0.208 (0.003)	0.204 (0.004)	-0.005 (-0.018, 0.007)
	4	0.205 (0.004)	0.204 (0.004)	0.205 (0.003)	0.203 (0.004)	-0.001 (-0.015, 0.013)
	5	0.198 (0.004)	0.194 (0.004)	0.199 (0.003)	0.194 (0.004)	-0.001 (-0.015, 0.012)
	6	0.197 (0.003)	0.190 (0.004)	0.197 (0.003)	0.194 (0.003)	0.005 (-0.008, 0.017)
	7	0.195 (0.003)	0.187 (0.004)	0.192 (0.004)	0.192 (0.004)	0.008 (-0.006, 0.021)
	8	0.195 (0.004)	0.185 (0.005)	0.189 (0.004)	0.188 (0.004)	0.009 (-0.007, 0.024)
	TOTAL	1.608 (0.022)	1.583 (0.027)	1.605 (0.021)	1.583 (0.024)	0.003 (-0.078, 0.085)

* Statistically significant interaction ($p<0.05$).

Interactions for medication/medical service costs associated with expected AEs/SAEs were non-significant (p≥0.206), although the overall interaction across all eight quarters was larger than the difference between drugs (Table 6.2). These interactions were qualitative in that that the sign (not just the magnitude) of the effect of treatment regimen changed depending on which drug was used. For example, the continuous Lucentis group accrued substantially lower medication/medical service costs than the discontinuous Lucentis group, whereas the continuous Avastin group accrued higher medication/medical service costs than the discontinuous Avastin group.

For QALYs, interactions between drug and treatment regimen were non-significant (p≥0.274), although the overall interaction across all eight quarters was larger than the difference between drugs. Although all QALY differences were extremely small and may not be clinically meaningful, the interaction was qualitative, since the continuous Lucentis group accrued 0.003 more QALYs than the continuous Avastin group, while the discontinuous Lucentis group accrued 0.0006 fewer QALYs than the discontinuous Avastin group (Table 6.2).

Allowing for interactions in all quarters (Analysis 4), rather than excluding all interactions (Analysis 2), reduced AIC for drug cost, but increased AIC for administration and monitoring costs, medication and medical service use costs and QALYs (Appendix I, Table I1).

The base case analysis (Analysis 1) therefore allowed for interactions in drug cost, medication/medical service use and QALYs in Q2-8, but assumed additive effects for administration and monitoring costs.

6.4.2 Base case results

The base case results (Analysis 1; Table 6.3) demonstrated that (as expected) continuous Lucentis is the most costly treatment and discontinuous Avastin is the least costly. The continuous Lucentis arm also accrued highest mean QALYs. Although patients receiving continuous Avastin accrued slightly (but not significantly) more QALYs than those on discontinuous Avastin (p=0.452), discontinuous Avastin had highest NMB at a £20,000/QALY ceiling ratio. Continuous Avastin cost £30,220/QALY gained vs. discontinuous Avastin. Continuous Lucentis cost £3.5 million/QALY vs. continuous Avastin, while discontinuous Lucentis was strongly dominated by continuous and discontinuous Avastin.

6.4.3 At-the-margins (Analyses 2 and 3)

Analysis 2 evaluated the impact of ignoring all interactions and assuming that drug and treatment regimen had additive effects on mortality, QALYs and all costs. This gave quite different results from the base case (Appendix I, Table I2). In particular, the interaction for drug cost was so large that OLS with no interaction term predicts that drug costs for discontinuous Avastin will be *negative*, as well as giving unrealistic estimates of the cost of other treatment regimens. As result of the implausibly low cost estimated for discontinuous Avastin, this analysis estimated the ICER for continuous vs. discontinuous Avastin to be £190,594/QALY: >6 times higher than the base case estimate. This analysis was not explored further due to these implausible results.

Table 6.3. Base case results (Analysis 1): Allowing for interactions in QALYs, drug costs, medication/medical service use costs and deaths related to treatment, but assuming no interactions for administration/monitoring costs.

	Mean (95% CI) QALYs†	Mean (95% CI) drug cost†	Mean (95% CI) administration & monitoring cost	Mean (95% CI) medications/ medical service cost†	Mean (95% CI) total cost†	Mean (95% CI) NMB†‡
Discontinuous Avastin	1.584 (1.538, 1.630)	£651 (£605, £698)	£1,825 (£1,708, £1,941)	£526 (£144, £908)	£3,002 (£2,601, £3,403)	£28,683 (£27,707, £29,658)
Continuous Avastin	1.604 (1.563, 1.645)	£1,065 (£1,048, £1,081)	£1,952 (£1,860, £2,043)	£585 (£250, £919)	£3,601 (£3,259, £3,943)	£28,480 (£27,548, £29,412)
Discontinuous Lucentis	1.582 (1.530, 1.634)	£9,229 (£8,584, £9,875)	£1,838 (£1,724, £1,952)	£432 (£253, £611)	£11,500 (£10,798, £12,202)	£20,142 (£18,963, £21,321)
Continuous Lucentis	1.608 (1.565, 1.651)	£16,286 (£16,011, £16,562)	£1,970 (£1,883, £2,057)	£334 (£215, £452)	£18,590 (£18,258, £18,922)	£13,576 (£12,769, £14,383)
Difference: Lucentis vs. Avastin§	Continuous: 0.004 (-0.046, 0.054) Discontinuous: -0.002 (-0.064, 0.060)	Continuous: £15,222 (£14,948, £15,495)* Discontinuous: £8,578 (£7,932, £9,225)*	£16 (-£109, £141)	Continuous: -£251 (-£604, £102) Discontinuous: -£94 (-£514, £326)	Continuous: £14,989 (£14,522, £15,456)* Discontinuous: £8,498 (£7,700, £9,295)*	Continuous: -£14,904 (-£15,995, -£13,813)* Discontinuous: -£8,541 (-£9,939, -£7,144)*
Difference: continuous vs. discontinuous§	Lucentis: 0.026 (-0.032, 0.085) Avastin: 0.020 (-0.032, 0.071)	Lucentis: £7,057 (£6,364, £7,750)* Avastin: £413 (£365, £462)*	£130 (£20, £239)*	Lucentis: -£98 (-£310, £113) Avastin: £59 (-£438, £556)	Lucentis: £7,090 (£6,337, £7,844)* Avastin: £599 (£91, £1,107)*	Lucentis: -£6,566 (-£7,861, -£5,271)* Avastin: -£203 (-£1,372, £967)
Interaction	0.006 (-0.071, 0.084)	£6,643 (£5,949, £7,338)*	£5 (-£31, £42)	-£157 (-£696, £381)	£6,491 (£5,604, £7,379)*	-£6,363 (-£8,088, -£4,638)*

* Significantly different from zero ($p<0.05$)
† Analysis allowed for interactions
‡ NMB equals QALYs multiplied by ceiling ratio minus costs; the table shows NMB at a £20,000/QALY ceiling ratio.
§ For administration/monitoring costs (where drug and treatment regimen were assumed to have additive effects), treatment effects were calculated as the average of the continuous Lucentis and discontinuous Lucentis groups minus the average of the continuous Avastin and discontinuous Avastin groups (e.g. £16= 0.5*(£1,970+£1,838)–0.5*(£1,952+1,825)), or as the average of the continuous Lucentis and continuous Avastin groups minus the average of the discontinuous Lucentis and discontinuous Avastin groups. For the other outcomes (for which interactions were included), treatment effects were calculated as the difference between outcomes in the relevant to cells (e.g. for continuous Lucentis vs. continuous Avastin, 0.004= 1.608-1.604). Numbers may not sum exactly due to rounding.

On a logarithmic scale (Analysis 3), the interaction between drug and treatment regimen was non-significant and negligible for drug costs in Q3 and Q5-8 ($p \geq 0.426$), increasing drug costs by between -4% and 7%, but was markedly larger and statistically significant in Q2 and Q4 ($p \leq 0.022$), increasing costs by 26-34%.[60] The interactions in Q2 and Q4 reflect the observation that patients randomised to discontinuous Avastin required 22-31% more injections in Q2 and Q4 than those randomised to discontinuous Lucentis ($p \geq 0.163$). This analysis therefore largely supports the hypothesis that drug and treatment regimen have multiplicative effects on drug cost, but highlights statistically significant interactions that are taken into account in Analysis 1 but not Analysis 3. Ignoring the interactions that are not eliminated by analysing drug costs on a logarithmic scale would have introduced bias into the analysis. Using GLM with log-link and no interaction decreased total AIC from 66,688 to 66,675 compared with Analysis 1, although this AIC reduction of only 1.9 per model may not be meaningful.

Analysis 3 assumed that drug and treatment regimen had additive effects on QALYs, mortality, administration/monitoring costs and medication/medical service use costs, but multiplicative effects on drug cost. This analysis gave plausible estimates of the drug cost within each treatment arm (Table 6.4) that were broadly similar to the base case analysis. Ignoring interactions reduced SEs and narrowed 95% CI in almost all cases, which reflects the increased statistical efficiency demonstrated previously (3, 5, 47). However, there were some exceptions to this trend, which are likely to have arisen due to heteroskedasticity: for example, SEs around drug cost and total cost were larger than the base case for continuous Avastin, but smaller for other arms, while SEs around medication/medical service use costs were larger than the base case for Lucentis-treated patients.

[60] Notably, the statistically significant interactions in Q2 and Q4 were only observed when SEs were estimated using bootstrapping or robust methods; SEs estimated using the observed information matrix were around 100 fold higher than bootstrap estimates for all quarters. This may be due to the large number of patients in the discontinuous arms who have zero drug costs, which introduces a large amount of heteroskedasticity.

Table 6.4. Results of Analysis 3: Ignoring all interactions, but analysing drug costs on a multiplicative scale

	Total QALYs (SE)	**Mean (95% CI) drug cost†**	**Mean (95% CI) administration & monitoring cost**	**Mean (95% CI) medications/ medical service cost**	**Total cost (SE)†**	**Total NMB (SE)†‡**
Discontinuous Avastin	1.584 (1.543, 1.624)	£613 (£576, £651)	£1,824 (£1,708, £1,940)	£566 (£262, £871)	£3,004 (£2,678, £3,331)	£28,668 (£27,803, £29,533)
Continuous Avastin	1.604 (1.566, 1.643)	£1,082 (£1,054, £1,110)	£1,951 (£1,859, £2,043)	£544 (£267, £821)	£3,577 (£3,287, £3,867)	£28,505 (£27,660, £29,350)
Discontinuous Lucentis	1.584 (1.539, 1.629)	£9,240 (£8,599, £9,880)	£1,840 (£1,727, £1,953)	£394 (£210, £578)	£11,473 (£10,778, £12,169)	£20,202 (£19,128, £21,277)
Continuous Lucentis	1.607 (1.567, 1.646)	£16,273 (£16,016, £16,529)	£1,969 (£1,883, £2,055)	£372 (£214, £530)	£18,613 (£18,284, £18,943)	£13,520 (£12,767, £14,273)
Difference: Lucentis vs. Avastin	0.001 (-0.038, 0.041)	Continuous: £15,191 (£14,939, £15,443)* Discontinuous: £8,626 (£8,020, £9,232)*	£17 (-£108, £141)	-£172 (-£452, £108)	Continuous: £15,036 (£14,632, £15,441)* Discontinuous: £8,469 (£7,778, £9,160)*	Continuous: -£14,985 (-£15,847, -£14,123)* Discontinuous: -£8,465 (-£9,489, -£7,442)*
Difference: continuous vs. discontinuous	0.022 (-0.016, 0.060)	Lucentis: £7,033 (£6,362, £7,705)* Avastin: £468 (£417, £520)*	£128 (£19, £237)*	-£22 (-£287, £243)	Lucentis: £7,140 (£6,398, £7,882)* Avastin: £573 (£281, £865)*	Lucentis: -£6,683 (-£7,708, -£5,658)* Avastin: -£163 (-£990, £664)
Interaction	0.002 (-0.005, 0.010)	£6,565 (£5,943, £7,187)*	£2 (-£6, £11)	£0 (-£4, £4)	£6,567 (£5,945, £7,190)*	-£6,520 (-£7,149, -£5,891)*

* Significantly different from zero ($p<0.05$)

† Analysis assumes that drug and treatment regimen have multiplicative effects on drug cost.

‡ NMB equals QALYs multiplied by ceiling ratio minus costs; the table shows NMB at a £20,000/QALY ceiling ratio.

6.4.4 Inside-the-table (Analysis 4)

Analysis 4 included interactions for all NMB components in all quarters (including Q1) and gave very similar point estimates to the base case (Table 6.5). The continuous Lucentis group accrued highest QALYs and continuous Avastin had highest NMB, costing £29,683/QALY vs. discontinuous Avastin. This demonstrates that ignoring all interactions in Q1 and interactions for administration/monitoring costs in Q2-8 introduced relatively little bias into the analysis. However, including all interactions increased almost all SEs, since all regression analyses include an additional variable that has negligible effect.

The increase in uncertainty associated with including additional interactions also affected CEACs and EVPI (Table 6.6, Figure 6.1). EVPI is always higher than the base case when all interactions are included and is always lower than the base case when all interactions are excluded (Figure 6.1). In all analyses, the probability of either Lucentis regimen maximising NMB remained <1% at all ceiling ratios below £190,000.

Table 6.6. Impact of assumptions about interactions on uncertainty measures

		£20,000/QALY ceiling ratio			**£30,000/QALY ceiling ratio**		
		1: Base case	**3: Exclude all inter-actions***	**4: Include all inter-actions**	**1: Base case**	**3: Exclude all inter-actions***	**4: Include all inter-actions**
EVPI per patient		£152	£98	£163	£337	£222	£354
Probability cost-effective	Continuous Lucentis	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Discontinuous Lucentis	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Continuous Avastin	37.1%	35.7%	37.5%	49.8%	53.7%	50.4%
	Discontinuous Avastin	62.9%	64.3%	62.5%	50.2%	46.3%	49.6%
Opportunity cost of ignoring interactions: current information		£0.00	£0.00	-	£6.80	£0.00	-
Opportunity cost of ignoring interactions: perfect information		£2.11	£73.91	-	£2.27	£116.73	-
Probability that ignoring interactions leads to adoption of a suboptimal treatment		4%	24%	-	3%	26%	-

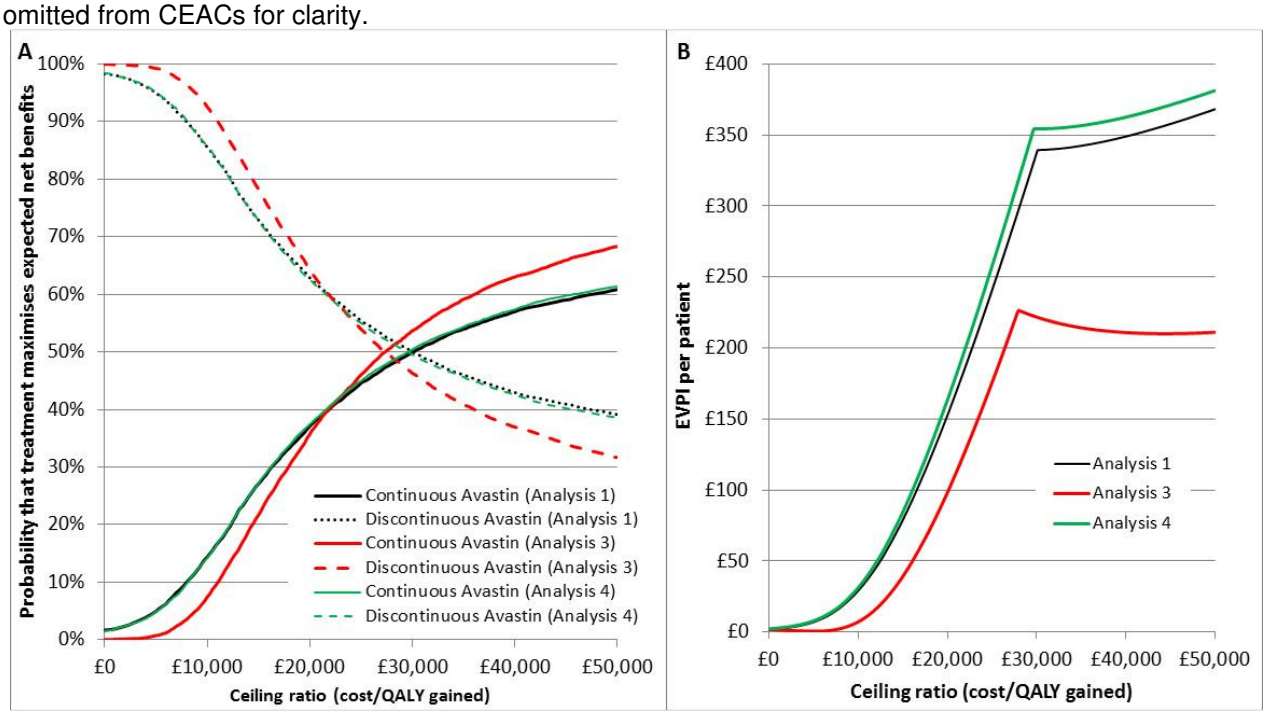
* Drug costs are assumed to have no interaction when analysed on a multiplicative scale.

Table 6.5. Results of Analysis 4: Including interactions for all components of NMB in all quarters (inside-the-table analysis)

	Mean (95% CI) QALYs†	**Mean (95% CI) drug cost†**	**Mean (95% CI) administration & monitoring cost**	**Mean (95% CI) medications/ medical service cost†**	**Mean (95% CI) total cost†**	**Mean (95% CI) NMB†‡**
Discontinuous Avastin	1.583 (1.536, 1.630)	£652 (£605, £698)	£1,826 (£1,676, £1,977)	£505 (£124, £886)	£2,983 (£2,572, £3,394)	£28,685 (£27,687, £29,684)
Continuous Avastin	1.605 (1.563, 1.646)	£1,064 (£1,047, £1,081)	£1,950 (£1,857, £2,042)	£606 (£258, £954)	£3,620 (£3,265, £3,975)	£28,478 (£27,528, £29,427)
Discontinuous Lucentis	1.583 (1.530, 1.636)	£9,239 (£8,593, £9,885)	£1,838 (£1,698, £1,979)	£424 (£244, £604)	£11,501 (£10,790, £12,212)	£20,156 (£18,958, £21,354)
Continuous Lucentis	1.608 (1.564, 1.651)	£16,277 (£16,001, £16,553)	£1,970 (£1,881, £2,059)	£342 (£221, £462)	£18,589 (£18,250, £18,928)	£13,562 (£12,739, £14,386)
Difference: Lucentis vs. Avastin	Continuous: 0.003 (-0.048, 0.054) Discontinuous: -0.001 (-0.064, 0.063)	Continuous: £15,213 (£14,939, £15,487)* Discontinuous: £8,587 (£7,941, £9,234)*	Continuous: £21 (-£101, £142) Discontinuous: £12 (-£201, £224)	Continuous: -£264 (-£631, £102) Discontinuous: -£81 (-£501, £339)	Continuous: £14,969 (£14,488, £15,450)* Discontinuous: £8,518 (£7,696, £9,340)*	Continuous: -£14,915 (-£16,034, -£13,796)* Discontinuous: -£8,530 (-£9,963, -£7,096)*
Difference: continuous vs. discontinuous	Lucentis: 0.025 (-0.037, 0.086) Avastin: 0.021 (-0.033, 0.076)	Lucentis: £7,038 (£6,344, £7,732)* Avastin: £412 (£364, £461)*	Lucentis: £132 (-£32, £297) Avastin: £123 (-£53, £299)	Lucentis: -£82 (-£299, £135) Avastin: £101 (-£414, £615)	Lucentis: £7,088 (£6,311, £7,865)* Avastin: £636 (£95, £1,178)*	Lucentis: -£6,593 (-£7,943, -£5,243)* Avastin: -£208 (-£1,442, £1,027)
Interaction	0.003 (-0.078, 0.085)	£6,626 (£5,930, £7,321)*	£9 (-£232, £250)	-£183 (-£739, £373)	£6,452 (£5,506, £7,398)*	-£6,385 (-£8,204, -£4,567)*

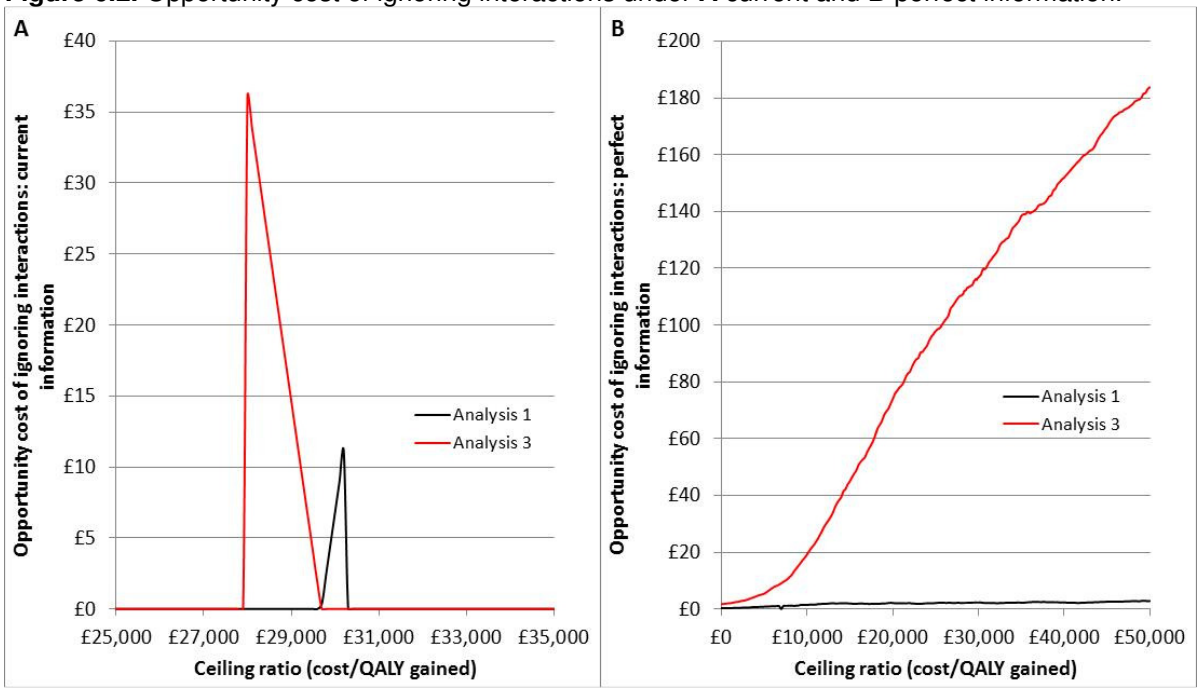
* Significantly different from zero ($p<0.05$)
† Analysis allowed for interactions
‡ NMB equals QALYs multiplied by ceiling ratio minus costs; the table shows NMB at a £20,000/QALY ceiling ratio.

Figure 6.1. CEACs (**A**) and EVPI (**B**) for analyses 1 (base case), 3 (excluding all interactions, but assuming multiplicative effects on drug cost), and 4 (including all interactions). The probability that either Lucentis regimen maximised NMB was zero at all ceiling ratios shown, so these treatments are omitted from CEACs for clarity.



The opportunity cost of ignoring interactions under current information was zero for both Analyses 1 and 3 at ceiling ratios <£27,958 and >£30,220 (when all three analyses draw the same conclusions), but rose to a maximum of £37/patient with Analysis 3 and £12/patient with Analysis 1 (Figure 6.2A). For Analysis 1, the opportunity cost of ignoring interactions under perfect information reached a plateau at around £3/patient, whereas that for Analysis 3 was always markedly higher and continue to rise with ceiling ratio (Figure 6.2B).

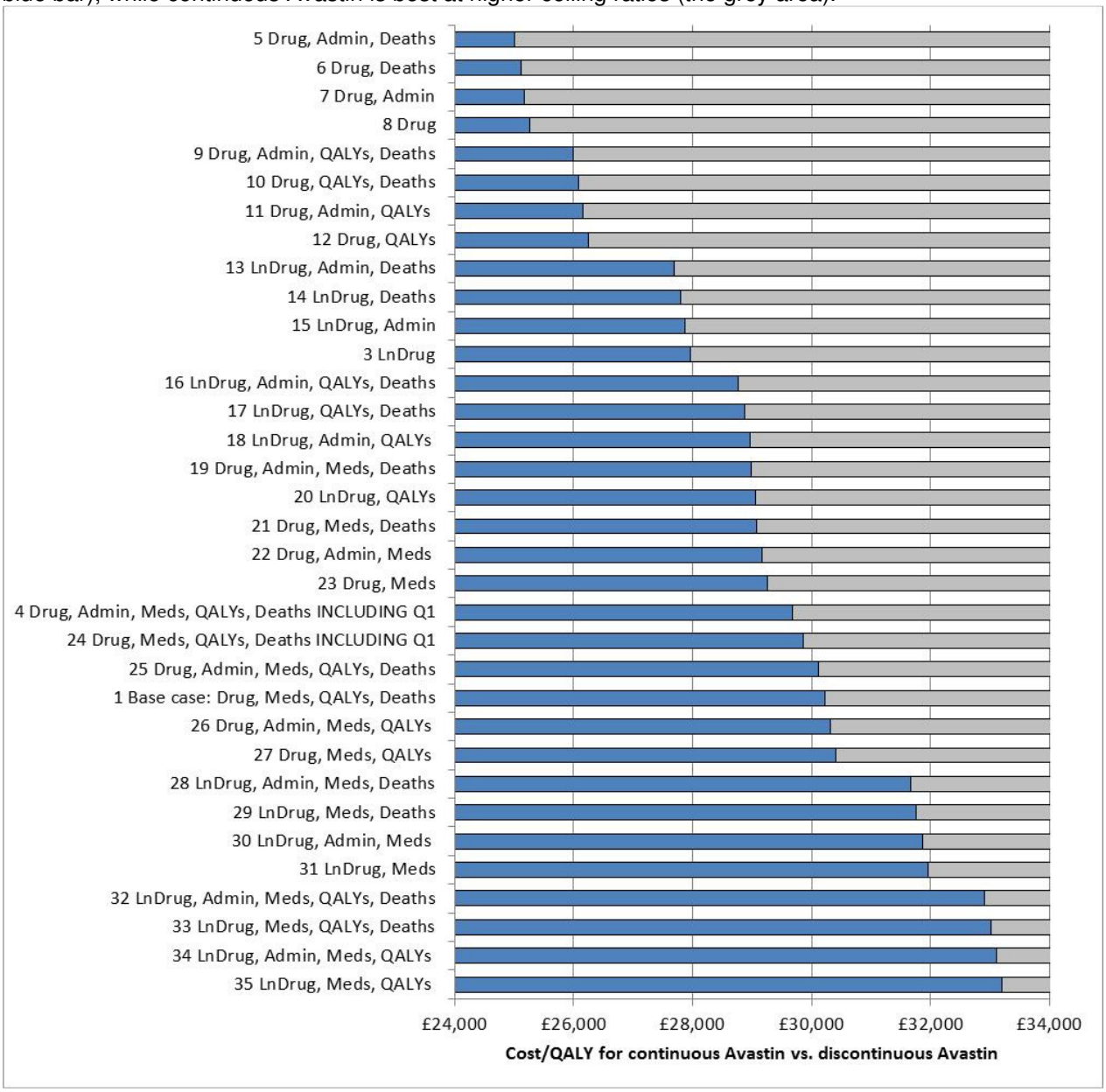
Figure 6.2. Opportunity cost of ignoring interactions under **A** current and **B** perfect information.



6.4.5 Alternative analyses

I also evaluated what the conclusions would be at different ceiling ratios for all 32 analyses with different combinations of interactions in the five NMB components, as well as two additional methods for analysing Q1 (Figure 6.3). None of these analyses changed the conclusion that Lucentis costs >£3 million/QALY gained vs. Avastin. Furthermore, all analyses that allowed for either interactions or multiplicative effects for drug cost found that continuous Avastin cost between £25,000 and £34,000 per QALY gained compared with discontinuous Avastin. However, conclusions were highly sensitive to the assumptions about interactions within this range.

Figure 6.3. Impact of including different combinations of interactions on the conclusions. The y-axis can be interpreted in one of two ways. Firstly, the point at which the blue bar ends and the grey bar begins represents the ICER for continuous Avastin vs. discontinuous Avastin. Secondly, the area covered by the blue bar represents the range of ceiling ratios at which discontinuous Avastin is the most cost-effective strategy considered in the trial, while the area shown by the grey bar represents the range of ceiling ratios at which continuous Avastin is most cost-effective. For example, Analysis 5 allows for interactions in drug costs ("Drug"), administration/monitoring costs ("Admin") and mortality ("Deaths"), the ICER for continuous Avastin vs. discontinuous Avastin is £25,010, indicating that discontinuous Avastin is the most cost-effective strategy at ceiling ratios below £25,010 (shown by the blue bar), while continuous Avastin is best at higher ceiling ratios (the grey area).



Abbreviations: Admin, interactions for administration/monitoring costs were included; Deaths, interactions for mortality were included; Drug, interactions for drug cost were included; LnDrug, drug and treatment regimen were assumed to have no interaction on drug costs when analysed on a logarithmic scale; Meds, interactions for medication/medical service use costs were included; QALYs, interactions for QALYs were included.

Interactions for deaths and administration/monitoring costs had negligible impact on the ICER for continuous vs. discontinuous Avastin, while interactions for medication/medical service use costs had the largest impact (excluding analyses assuming additive effects on drug cost).

However, including interactions for one variable sometimes altered the impact of interactions in another variable: this effect could be termed “interactions between interactions”. For example, the impact of allowing for interactions in QALYs was larger when interactions in medication/medical service use costs were also included, and vice versa.

As a result of the differences between these analyses, some of the findings are sensitive to the criteria used to select which interactions were considered “important” or “genuine”. Including all interactions beyond Q1 (Analysis 25) would have reduced the ICER for continuous Avastin vs. discontinuous Avastin by £4 to £30,116/QALY and would not have changed SEs. If I had included only those interactions that were statistically significant (at either alpha=0.05 or alpha=0.10), I would have included only interactions for drug cost (Analysis 8); this would have reduced SEs and reduced the ICER for continuous Avastin to £25,263/QALY. Basing our choice of analysis on AIC would have led us to assume multiplicative effects for drug costs and exclude all other interactions (Analysis 3); this would have reduced the ICER for continuous Avastin to £27,958/QALY. Analysing Q1 outcomes in the same way as other quarters (Analysis 24) increased SEs around NMB and reduced the ICER for continuous Avastin to £29,857/QALY.

6.5 Discussion

Within IVAN, assumptions about interactions had a marked effect on decision uncertainty and conclusions. In particular, analysing and interpreting the results at-the-margin and

assuming no interactions on a natural scale gave negative cost estimates and suggested that there was no decision uncertainty.

Analyses including all interactions have been shown to generate unbiased estimates of incremental outcomes (3, 4); as a result, the conclusions of Analysis 4, which allows for all interactions, may be taken as a “gold standard” against which we can compare mean results for any given analysis. However, results also showed that allowing for unimportant interactions increases estimates of SEs, the probability of adopting the wrong treatment and EVPI. Although it is unclear which analysis gives the “correct” measure of decision uncertainty, I feel that the base case results are most realistic as they minimise bias and it would be inappropriate to commission further research when a more efficient statistical analysis could have sufficed.

Tailoring assumptions about interactions to each component of NMB and interpreting results inside-the-table helped to reduce bias in both point estimates and uncertainty measures. The base case analysis achieved point estimates comparable with analyses including all interactions, but generally produced smaller SEs, lower EVPI and a lower probability of making the wrong decision. At a £20,000/QALY ceiling ratio, ignoring interactions for administration/monitoring costs had zero opportunity cost given current information (since we would adopt discontinuous Avastin regardless), although an opportunity cost of £7/patient would be incurred at a £30,000/QALY ceiling ratio.

Drawing conclusions inside-the-table also enables us to decide between the four cells of the factorial design as mutually-exclusive alternatives, rather than making separate decisions on drug and treatment regimen as though these decisions were independent. However, one shortcoming of my approach is that it explicitly adjusts group means for assumptions about interactions. As result, the main results (Table 6.3) artificially remove interactions that are

excluded from the analysis; I therefore recommend that unadjusted group means (e.g. Table 6.2) are also presented wherever possible.

However, my approach increases the complexity of statistical analysis and interpretation, as well as requiring researchers to pre-specify and apply appropriate criteria to determine which interactions are included in the analysis and which can be ignored. The simulation work described in Chapter 5 suggested that including all interactions minimised the opportunity cost of adopting a suboptimal treatment across all simulated trials, although excluding all interactions or including only interactions that were qualitative or larger than simple effects performed better for the one-year IVAN results (Appendix H, Table H3) since ignoring for interactions increased the probability that discontinuous Avastin was cost-effective.

A slightly different decision rule was used for the one-year analysis: namely including interactions that were statistically significant and those that changed conclusions about total costs, QALYs or NMB at ceiling ratios between £10,000 and £40,000/QALY gained (255). However, Figure 6.3 demonstrates that interactions of almost any magnitude can change the conclusions at some specific ceiling ratios if the point estimate lies within this range; applying this rule strictly would therefore result in inclusion of all interactions.

Nonetheless, some of the qualitative interactions observed could have arisen by chance. In particular, the large interaction for medication/medical service use costs appears to be due to chance as three of the five patients with costs >£10,000 were in the continuous Avastin group and omitting these five patients greatly reduced the interaction. Similarly, the interaction for QALYs is qualitative simply because the difference in QALYs between Avastin and Lucentis is <0.003.

Bayesian informative priors could have down-weighted the qualitative interaction for medication/medical service use costs. However, it is unclear how such priors should be

elicited. The methods used by Welton et al (which assume that we can be 95% confident that the interaction is smaller than main effects (116), or sub-additive but not qualitative (115)) might be suitable for QALYs and non-drug costs, but not for drug costs. Including only those interactions specified a priori is a variant of this approach, which effectively assigns 0% weight on any unexpected interactions. In this case, assuming multiplicative effects for drug cost and additive effects for all other NMB components (Table 6.4) would have given similar conclusions at a £20,000/QALY ceiling ratio, but would have ignored the unexpected qualitative interactions in QALYs and medication/medical service use costs.

Since my results highlight the potential impact of interactions on conclusions, I recommend that the criteria used to identify interactions are pre-specified in advance of data analysis to avoid data mining or bias. It is also important to ensure that costs and QALYs are analysed in the same way across all time periods unless there is a good reason to expect interactions to differ over time (e.g. if the trial is non-factorial for some periods).

The same analytical approach could be applied in other situations where interactions are expected for some types of costs or health benefits but not others. The analysis uses a single case study where drug costs are the main cost driver, which has a non-inferiority design and evaluates two drugs and two treatment regimens, rather than comparing two drugs against their respective placebos. Partly as result of this atypical trial design, the interactions in IVAN may be more predictable and more clear-cut than other studies. Furthermore, the bias introduced by ignoring interactions could be substantially smaller in trials with smaller interactions and/or larger treatment effects. However, ignoring interactions can introduce bias whenever treatment effects are not truly additive (3, 4, 8, 20, 30) and including all interactions appeared to perform better for most trials in simulation work (Chapter 5). A strategy of including all interactions may therefore perform better for analyses that will only inform decisions about treatment adoption today, not further research.

IVAN also raises the question of whether different types of economic evaluation can be used for different factors, analogous to the relatively common practice of using different primary endpoints for different factors (e.g. (85) or AFORRD). Within the primary IVAN publications (260) (Appendix I), drugs were compared using CMA, while treatment regimens were compared using CUA (both stratified by the other factor), in addition to the four-way comparison presented here.[61] However, presenting two evaluation types complicated presentation of results, without reducing analytical complexity. In principle different denominators could also be used for different factors within CEA (e.g. cost/pain-score reduction for Factor A and cost/depression-score reduction for Factor B), although it is unclear how such an analysis would deal with interactions. I would recommend that future studies present only CUA (or CEA with a common denominator), to facilitate interpretation of ICER and evaluation of interactions.

The study also highlights the more general benefits of analysing different components of NMB separately. Whereas guidelines often recommend regression directly on NMB (261), this study highlights the importance of considering the most appropriate model for QALYs and for different types of cost and the benefits of conducting multivariate regression.

[61] Although CMA will almost always bias uncertainty estimates as well as potentially giving misleading conclusions, such bias is minimal in situations where the difference in drug costs is so large that no plausible difference in efficacy will change the conclusion that the cheaper drug is cost-effective (126). In principle (as occurs for IVAN), large differences in drug costs may be anticipated for one factor (in this case the comparison between drugs), but not the other (in this case the comparison between treatment regimens). This means that CMA is inappropriate for the comparison between treatment regimens, but not for the comparison between drugs. Following discussions with the trial collaborators, it was therefore decided that different types of analysis would be presented for different factors within IVAN. As result, the main papers reporting results of IVAN (accepted by *Health Technology Assessment* (260) and *BMJ Open* [Appendix I]), report results of CMA comparing drugs and CUA comparing treatment regimens (both stratified by the other factor) , as well as a four-way comparison between all four study arms.

Chapter 7: KAT – A partial factorial trial

7.1 Introduction

Partial factorial trials[62] evaluate multiple treatments simultaneously on the same patient group by randomising a subset of patients to >1 factor in a factorial manner, while other patients are randomised to just one factor or to a different combination of factors (Figure 7.1). Although this design is rare (even among factorial trials, Chapter 4), high-profile examples include the Women's Health Initiative (62) and the UKPDS (263).

Figure 7.1. Schematics of full and partial factorial designs

Full factorial trial (n=400)

	Placebo of A	**Drug A**
Placebo of B	105 pts (***00, 0***)	97 pts (***a0, a***)
Drug B	98 pts (***0b, b***)	100 pts (***ab***)

Partial factorial trial (n=400)

	Placebo of C	**Drug C**	**Not randomised to comparison C**
Placebo of D	62 pts (***00***)	61 pts (***c0***)	65 pts (***-0***) of whom 25 had C
Drug D	58 pts (***0d***)	55 pts (***cd***)	67 pts (***-d***) of whom 35 had C
Not randomised to comparison D	15 pts (***0-***), of whom 7 had D	17 pts (***c-***) of whom 6 had D	

By comparing multiple treatment factors on overlapping populations, partial factorial trials can address multiple questions in the same study, reducing the fixed costs of each research question and the overall sample size compared with conducting several non-overlapping trials. Since some patients are randomised to different combinations of factors in a factorial manner, partial factorial trials can also investigate interactions: i.e. whether the effect of C differs depending on whether D is also given. Partial factorial designs are often used when economic, geographic, or clinical constraints restrict the comparisons to which patients can be randomised (61), since this design facilitates flexible recruitment strategies, such as

[62] I define "partial factorial trials" as studies in which some but not all patients are randomised in a factorial manner to combinations of two or more treatments (61, 62, 262). However, the same term is also used to describe "incomplete factorial trials" in which all experimental units are randomised to a subset of the possible combinations of factors (30, 68). My literature reviews suggest that most studies meeting my definition of partial factorial trials describe their design as simply "factorial".

letting patients[63] (62) or clinicians (59) choose which comparisons to be randomised into, or recruiting only patients from certain countries (264) or those with specific comorbidities, clinical characteristics or laboratory findings to a second (or even third) comparison (62, 265).

Despite extensive research and established guidelines on factorial trials in general, only one review article has discussed partial factorial trials (61). Partial factorial trials raise several additional issues over and above those introduced by full factorial designs and the appropriate analytical methods are less clear and under-researched. I propose two methods that could be used to evaluate interactions and conduct inside-the-table analysis:

- Firstly, we could focus on the subset of patients randomised to >1 comparison and analyse cells *00*, *c0*, *0d* and *cd* (Figure 7.1) “inside-the-table” like a full factorial trial. This analysis ensures unbiased estimation of interactions, but excludes large numbers of patients (41% in the case of Figure 7.1). As result, interactions can only be evaluated on a small sample, power to detect main effects is reduced and the value of information may be overestimated. Furthermore, the patients randomised to >1 comparison may not be representative of the whole trial population, potentially reducing generalisability.
- Secondly, we could analyse the entire trial population “as-treated” and subdivide all patients into ≥4 groups based on the combination of treatments that they actually received. For example, in Figure 7.1 we might pool patients in cell *cd* together with the six patients in cell *c-* who were randomised to Drug C and received D without being randomly assigned in this comparison. Inside-the-table analysis can therefore be done by comparing outcomes for the four combinations of received treatment. However, this approach analyses patients according to the treatment that they actually received (rather than their randomised allocation) and is therefore prone to the selection bias associated

[63] In this sense, partial factorial trials are like several overlapping "patient preference trials" (also known as the "comprehensive cohort design"), in which patients who have a strong preference for one treatment over the other receive their preferred treatment, while those who have no strong views are randomised (58).

with per protocol analysis (266). Since the patients in cells *0-*, *c-*, *-0* and *-d* are not randomly assigned to the second factor, any observed effect of this second factor or any observed interactions between the two factors could be caused by confounding rather than causal effects. For example, if we find that patients in cell *0-* who have D accrue fewer QALYs than those in cell *0-* who do not have D, this could be due to patient characteristics that both affected outcomes and the probability of receiving treatment D, rather than a causative effect of D. As such, subgroup analysis of partial factorial trials carries many of the same hazards and biases as analyses of observational studies, or subgroup analysis of two-arm trials (267, 268). By contrast, full factorial randomisation with intention-to-treat (ITT) analysis and concealment of allocation avoids selection bias and ensures that the only systematic difference between randomised groups is the allocated treatment.

Although it is prone to bias when there are interactions (3, 4, 8, 20, 30), at-the-margins analysis (e.g. evaluating Drug C by comparing outcomes in cells *c0*, *cd* and *c-* with those in cells *00*, *0d* and *0-*) may give a good indication of average effects for partial factorial trials. Several reviews and textbooks on full factorial trials argue that at-the-margins estimates of the main effect of factor A may be informative even in the presence of interactions if the ratio of B to not-B in the trial reflects that in the setting of interest (e.g. routine clinical practice) (10, 30, 36). While this is unlikely for many full factorial trials, partial factorial trials enable the distribution of patients between B and not-B to be governed by routine clinical practice for those patients not randomised to this comparison. As result, at-the-margins analysis will give a good estimate of average incremental effects across the whole population for a partial factorial trial, even if there is an interaction, although there are still benefits from evaluating interactions and exploring whether outcomes differ with B.

My systematic review identified five partial factorial trials with economic evaluations (Chapter 4, Table 7.1). All were analysed at-the-margins as overlapping trials addressing independent

questions (232, 263, 269-271). Two studies considered interactions, of which one proposed analysing the subsets of patients randomised to each combination of factors separately (232) and the other did not clearly describe their methods (269). The three remaining studies reported the different factors in different papers (263, 270, 271).

Table 7.1. Partial factorial trials with economic evaluations identified in the systematic review described in Chapter 4. Appendix G gives more details on each study.

	Publication	**How authors described trial**	**Methods for economic evaluation**	**% patients in ≥1 comparison**
ASCOT 1: (231, 249, 271-275)	Factors presented in separate papers	Some but not all of the clinical and economic papers describe it as "factorial"	No mention of interactions	53%
ACCORD (232)	Protocol only	Described as "factorial"	Propose conducting at-the-margins and inside-the-table analysis. Hypothesise that there will be non-additive effects	100%*
FRISC II (270, 276-280)	Factors presented in separate papers	Clinical papers describe as "factorial"	No mention of interactions	Unclear (likely to be around 60%)
SAPPHIRE (269, 281)	Factors presented separately in the same paper	Described as factorial	Interactions evaluated but excluded as non-significant	78%
UKPDS (211, 263, 282-288)	Factors presented in separate papers	Clinical paper describes as factorial	No exploration of interactions	Unclear (likely to be around 50%)

Abbreviations: ACCORD, Action to Control Cardiovascular Risk in Diabetes; ASCOT, Anglo-Scandinavian Cardiac Outcomes Trial; FRISC II, FRagmin and Fast Revascularisation during InStability in Coronary artery disease; SAPPHIRE, shoulder acute pain in primary health care.
* All patients were randomised in the glycaemia comparison; 54% of participants were also randomised in the lipid comparison, while the remainder were randomised in the blood pressure comparison.

This study aims to explore the implications of partial factorial design on the methods and results of economic evaluation and compare how the results and conclusions of an applied partial factorial trial differ between at-the-margins analysis and inside-the-table analysis on patients randomised to >1 comparison. Results for inside-the-table subgroup analysis are shown in Appendix J.

7.2 Case study

KAT is a pragmatic partial factorial RCT evaluating three[64] aspects of knee prosthesis design (58-60):

A) Using a mobile bearing in the joint between the tibia and femur, vs. a fixed bearing;
B) Using a metal-backed tibial component, vs. one made of solid polyethylene;
C) Resurfacing the patella (replacing part of the knee-cap with plastic), vs. no resurfacing.

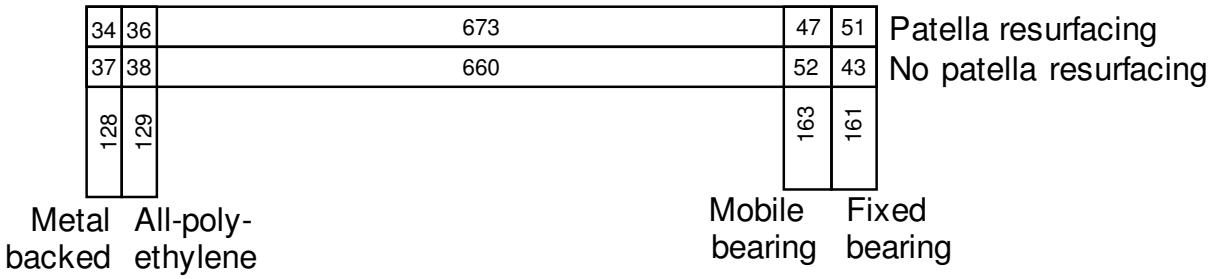
The partial factorial design enabled patients to be randomised to Comparison C as well as either Comparison A or Comparison B (Figure 7.2). Surgeons recruited patients only to those comparisons for which they were happy for the treatment decision to be made by random allocation. Patients about to undergo primary knee replacement under the care of a collaborating surgeon were recruited and randomised to those comparisons for which the surgeon was in equipoise; patients whose surgeon considered a particular type of surgery to be clearly indicated were therefore not randomised in that comparison. Surgeons were free to use their clinical judgement on all aspects of patient care to which patients were not randomised: for example, surgeons treating patients randomised only in Comparison A were free to choose whether or not to resurface the patella.[65]

This design and recruitment strategy was chosen to maximise recruitment of surgeons and patients and to acknowledge the marked variations between surgeons in the comparisons for which they are willing to accept randomisation. A full factorial design is unlikely to have been feasible as only one centre recruited patients to all three comparisons. Given that there was no previous evidence on interactions and no clinical reason why one would be expected, the

[64] A fourth aspect was also evaluated (58), comparing total knee replacement with unicompartmental replacement (replacing only one side of the knee joint), but is not discussed in this chapter since it did not overlap with other comparisons and stopped early.
[65] However, some limitations were imposed by the fact that all mobile bearings are metal-backed. Patients randomised to mobile bearings could therefore not receive all-polyethylene components, while those randomised to all-polyethylene could not receive mobile bearings.

primary clinical and economic analyses were conducted at-the-margins and the sample size was calculated on that basis, although exploration of interactions was pre-specified in the protocol (262).

Figure 7.2. Design of KAT (n=2252, excluding patients who died/withdrew before surgery)



Numbers and rectangle areas represent the number of patients randomised to each arm.

CUA was conducted on results at a median of 10 years' post-operation follow-up; methods have been reported previously (58). Costs were discounted at 3.5% per annum and are presented in 2011/12 pounds.

Multiple imputation (256) was used to impute missing data on utilities and resource use. Partial factorial designs raise particular challenges for multiple imputation. Firstly, imputation could be conducted once on the entire trial population, or separately for each comparison. I used the former approach to maximise the amount of data available for imputation and thereby get more precise imputations and facilitate conditional imputation models. This also facilitates inside-the-table analyses and ensures consistency between comparisons. The alternative strategy of conducting imputation separately for each of the three comparisons would have greatly reduced the sample size for imputation models (and therefore the precision of imputed values) and produced two different sets of imputed values for patients randomised to two comparisons. Secondly, the best way to define treatment variables in the imputation model is unclear. To avoid bias, the variables, functional forms and interaction terms in any imputation model should match those used in the analysis model (256). For full factorial trials, variables for all potential analysis models can be captured by including one

variable for each factor and one for each interaction term. However, for partial factorial trials, having one dummy per factor is not feasible: patients can, for example, be randomised to mobile bearing, randomised to fixed bearing or not randomised in Comparison A (in which case they may receive either mobile or fixed). I therefore included six treatment dummies in the imputation model (RandtoPatella, RandtoNoPatella, RandtoMobile, RandtoFixed, RandtoMetal, RandtoPoly), each equal to 1 if the patient was randomised to that treatment and 0 if they were randomised to the alternative or not randomised in that comparison. This coding is equivalent to an at-the-margins model:

$$if\ (Randto\text{Patella} = 1, or Randto\text{NoPatella} = 1), then: y = \beta_0 + \beta_{\text{Patella}} Randto\text{Patella}$$
$$= \beta_{\text{PatellaNoCons}} Randto\text{Patella} + \beta_{\text{NoPatellaNoCons}} Randto\text{NoPatella} \qquad (7.1)$$

where y is costs or QALYs. This imputation model does not allow for interactions between factors and may therefore cause inside-the-table analysis to underestimate the magnitude of interactions. In principle, interaction terms (RandtoPatella*RandtoMobile and RandtoPatella*RandtoMetal) could have been added into the model, although they may have hindered convergence of the imputation model, since <50 patients would have a 1 in each interaction term.

All preparatory work needed to get total annual costs and QALYs (including data cleaning and multiplication of resource quantities with unit costs) was also done on the entire dataset simultaneously. Once annual costs and QALYs had been calculated for each patient, I generated separate datasets for each comparison, and (for the base case analysis) ran bootstrapping separately on each comparison in a do file that looped over each of the three comparisons. This approach to data management facilitated the methodological analyses described in this paper, saved analysis time and ensured that all data cleaning and processing steps were conducted identically on all comparisons.

I used inverse probability weighting (IPW) (155) to allow for administrative censoring and OLS to adjust QALYs for imbalance in baseline utility (150). Bootstrapping was used to quantify uncertainty and was done 100 times for each of the 100 imputed datasets; point estimates were based on means for raw data (with no bootstrapping). Results for the 100 imputed datasets were combined using Rubin's rule (256) and used to calculate CEACs. EVPI was calculated as the mean maximum NMB accrued in each bootstrap replicate minus the mean NMB for the treatment with highest expected NMB.

7.3 Analytical methods

Bootstrapping and analysis of results using IPW were conducted using at-the-margins analysis and inside-the-table analysis on patients randomised to >1 comparison to evaluate the impact on the results. A third analysis using the subgroup approach is described in Appendix J. Unless otherwise specified, all analyses were conducted in Stata version 12, all p-values are two-sided and NMB is reported at a £20,000/QALY ceiling ratio (130).

7.3.1 Analysis 1 (base case) methods: at-the-margins analysis

The base case analysis comprised at-the-margins analysis to ensure consistency with the primary clinical analysis and to take account of all participants as-randomised. Furthermore, at-the-margins estimates are particularly likely to reflect population average effects for KAT, since patients in Comparisons A or B were only randomised in Comparison C if surgeons were in equipoise, while patella resurfacing in those patients randomised to only comparison A or B will have reflected routine clinical practice. For this analysis, bootstrapping was conducted separately for each of the three comparisons, excluding all patients not randomised to that comparison. Costs and QALYs in each year were calculated for each group and combined using IPW.

7.3.2 Analysis 2 methods: Inside-the-table analysis

Analysis 2 considered the subset of patients randomised to >1 comparison on an ITT basis as a small full-factorial trial. In this analysis, bootstrapping was conducted twice: once on the 193 patients randomised to Comparisons A and C, and once on the 145 patients randomised to Comparisons B and C. OLS was used to predict the costs and QALYs accrued in each year of the trial conditional on a dummy indicating randomised allocation in Comparison A or B, a dummy for whether patients were randomised to patella resurfacing, an interaction term equalling RandtoPatella*RandtoX, and (for QALYs only) baseline utility. Regression analyses were replicated on each bootstrap replicate for each imputed dataset to generate 10,000 predictions of mean annual costs and QALYs; total costs and QALYs were calculated using IPW.

The impact of interactions was estimated using several measures:

- Statistical significance: The interaction between treatment allocations was calculated for total costs, QALYs and NMB in each bootstrap replicate (interaction=$\mu_{ab}-\mu_a-\mu_b+\mu_0$) and results across imputed datasets were pooled using Rubin's rule (256) to estimate the mean, SE and two-sided p-value (which was evaluated at alpha=0.05).
- AIC: No established methods for estimating a single AIC value for datasets with administrative censoring, multiple imputation and excess zeros have been developed to date and developing such methods is beyond the scope of this thesis. I therefore estimated AIC for models with and without interactions separately for each of the 100 imputed datasets.[66] Since AIC values cannot be combined across imputed datasets

[66] I used mixed models to predict the QALYs accrued by each patient in each of the 10 study years and estimated a single AIC value for each imputed dataset across the whole study period. The mixed model included random constants for each patient, baseline utility, a dummy for randomised patella allocation and a dummy for randomised allocation in Comparison A or B; the model was estimated with and without an interaction term. Within this analysis, QALYs in years after administrative censoring were treated as missing. Since mixed models of annual costs did not converge due to excess zeros in Years 2-10, GLM with log-link and gamma family was used to predict total cost over the period from randomisation to 10 years (or death/study exit) with the same explanatory variables (excluding baseline utility). Costs in years after the patient was administratively censored were assumed to be equal to 0 within this analysis.

(256), I present the proportion of datasets in which AIC was lower with interactions as well as mean AIC.

- Opportunity cost of ignoring interactions under current and perfect information and the probability that at-the-margins analysis gives misleading conclusions: These measures used the methods described in Section 3.8 to compare the results of Analysis 2 with at-the-margins analyses conducted on the subset of patients randomised to >1 comparison, since Analyses 1 and 2 include different patient groups.
- Estimated bias from ignoring interactions: The bias in main effect size estimates caused by ignoring interaction terms was estimated as half the size of the interaction term (3).

7.4 Results

7.4.1 Analysis 1 (base case) results: at-the-margins

Analysis 1 evaluated each of the three comparisons separately on all patients randomised to that comparison, treating each comparison as an independent decision. Comparing outcomes for all patients randomised to patella resurfacing against all those randomised to no resurfacing suggested that patella resurfacing dominated no resurfacing, generating an additional 0.19 QALYs and saving an average of £104 per patient treated with a 96% chance of being cost-effective[67] (Table 7.2; Figure 7.3). Analysis 1 also suggested that we can be 91% confident that metal-backed components are cost-effective compared with non-metal backed, with an ICER of just £35 per QALY gained. On average, mobile bearings were cost-effective compared with fixed bearings, although QALY gains were very small and there was substantial uncertainty around this finding. If we were to make separate decisions treating the different aspects of knee component design as independent options, we would therefore recommend patella resurfacing, metal-backing and (more hesitantly) mobile bearing.

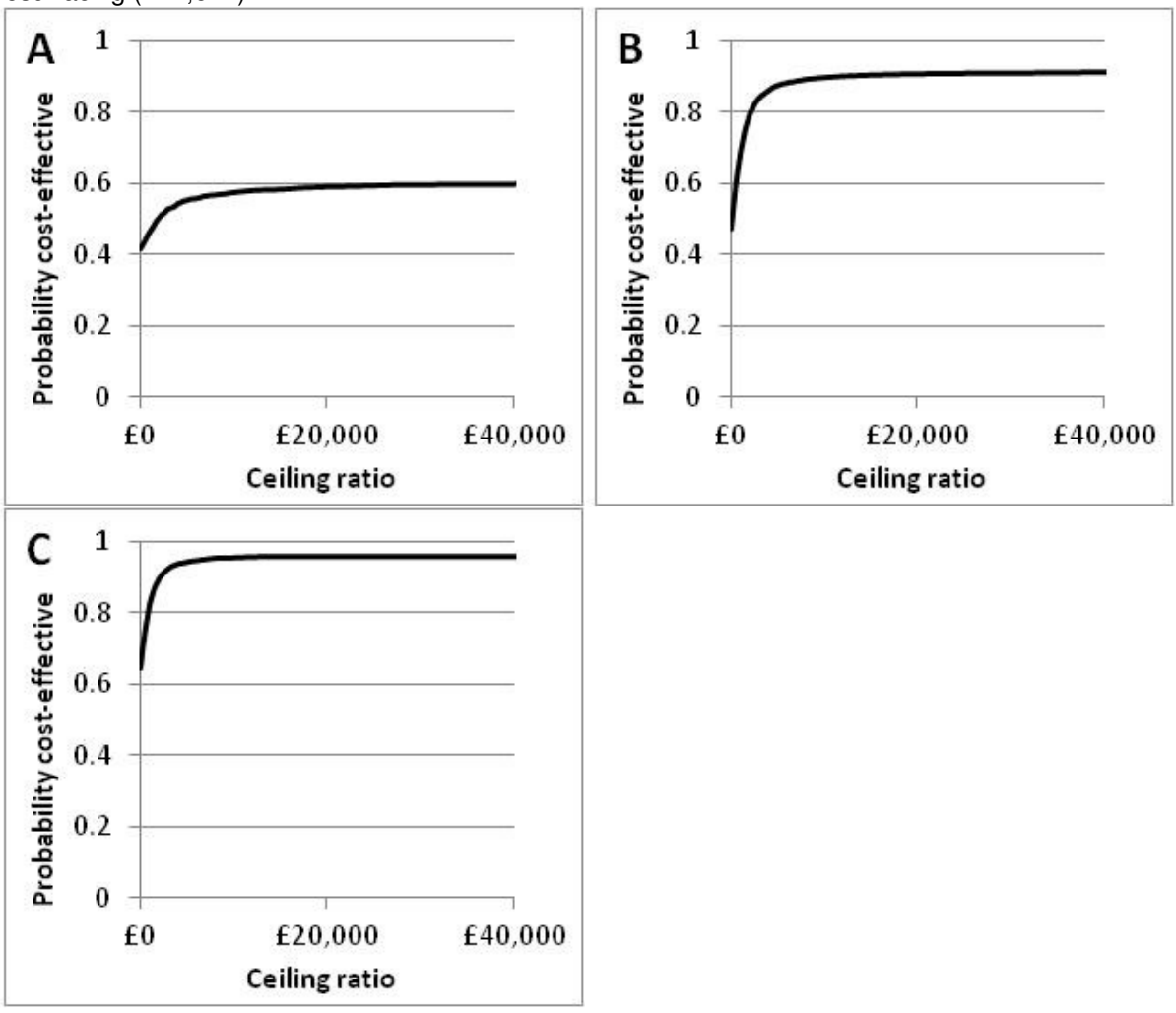
[67] Notably, this analysis provides a real-life example of the situation identified hypothetically previously (126, 136), in which NMB differs significantly (p=0.96 on a one-tailed test), despite non-significant differences in both costs and effects.

Table 7.2. Base case results for all three comparisons. Numbers in brackets are SEs.

		Mobile bearing (n=262) vs. fixed bearing (n=255)	**Metal backed (n=199) vs. non-metal backed (n=203)**	**Patella resurfacing (n=841) vs. no resurfacing (n=830)**
Treatment group	Cost	£8,998 (£310)	£8,235 (£272)	£8,785 (£161)
	QALYs	5.007 (0.143)	5.219 (0.151)	5.297 (0.076)
	NMB*	£91,145 (£2,968)	£96,145 (£3,112)	£97,158 (£1,551)
Control group	Cost	£8,913 (£405)	£8,225 (£344)	£8,889 (£211)
	QALYs	4.956 (0.141)	4.926 (0.152)	5.110 (0.080)
	NMB*	£90,209 (£2,938)	£90,290 (£3,144)	£93,308 (£1,662)
Difference	Cost	£85 (£508)	£10 (£440)	-£104 (£269)
	QALYs	0.051 (0.196)	0.293 (0.210)	0.187 (0.108)
	NMB*	£936 (£4,087)	£5,854 (£4,343)	£3,849 (£2,235)
ICER (per QALY gained)		£1,666	£35	Dominant
Probability cost-effective*		0.59	0.91	0.96
Probability cost-saving		0.42	0.47	0.64

* At a £20,000/QALY ceiling ratio.

Figure 7.3. CEACs for independent decisions based on base case analysis: including all patients randomised in each comparison. **A**: mobile bearing (n=517). **B**: metal backing (n=402). **C**: Patella resurfacing (n=1,671).



7.4.2 Analysis 2 results: Inside-the-table analysis

Analysis 2 aimed to inform joint decisions between combinations of treatment strategies by analysing the subset of patients randomised to >1 comparison as a full factorial trial. The analysis of mobile bearings and patella resurfacing included 193 patients: 37% of those randomised in Comparison A. The subset of patients in these two comparisons tended to have higher costs than the average patient in Comparison A (Table 7.3) and all SEs were at least twice as large as those in Analysis 1 due to the substantially smaller sample size.

Table 7.3. Primary analysis of interactions: including only those patients randomised to >1 comparison. Numbers in brackets are SEs.

		Mobile bearing	Fixed bearing	Metal backing	All-polyethylene
Patella resurfaced	No. pts	47	51	34	36
	Cost	£9,068 (£466)	£9,169 (£1,165)	£8,036 (£411)	£7,833 (£567)
	QALYs	5.559 (0.264)	4.959 (0.289)	5.518 (0.337)	5.046 (0.330)
	NMB*	£102,110 (£5,372)	£90,015 (£6,256)	£102,327 (£6,870)	£93,087 (£6,837)
No patella resurfacing	No. pts	52	43	37	38
	Cost	£11,100 (£1,147)	£8,481 (£464)	£7,782 (£384)	£8,085 (£409)
	QALYs	4.732 (0.311)	5.029 (0.294)	4.976 (0.311)	5.569 (0.248)
	NMB*	£83,533 (£6,755)	£92,104 (£6,014)	£91,745 (£6,348)	£103,293 (£5,031)

* At a £20,000/QALY ceiling ratio.

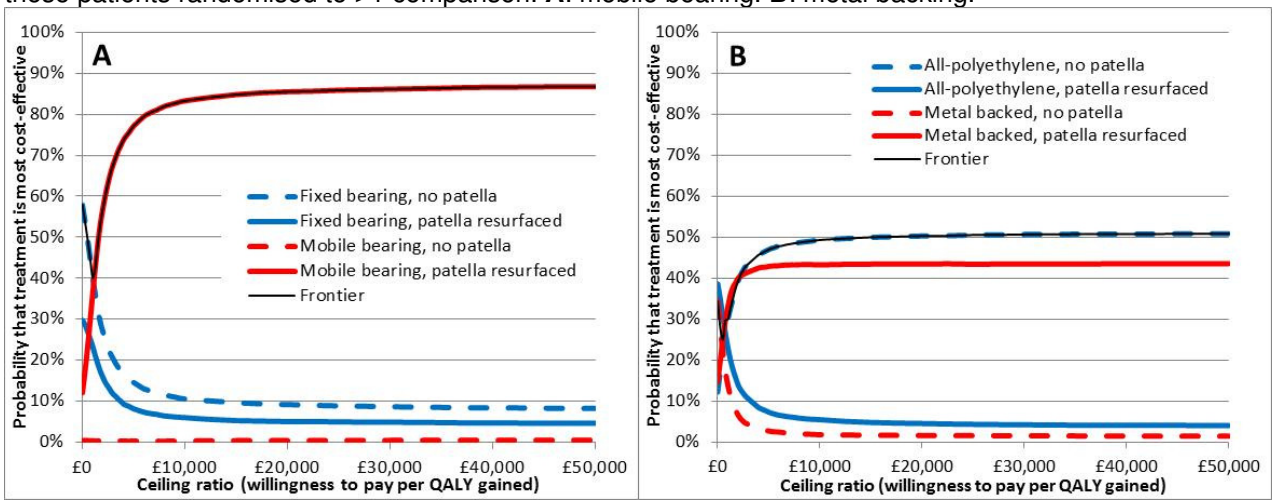
Large, non-significant interactions were observed for costs (-£2,720 [SE: £1,751]; p=0.12), QALYs (0.90 [SE: 0.51]; p=0.08) and NMB (£20,667 [SE: £10,820]; p=0.06). This suggests that main effect estimates from at-the-margins analysis include substantial bias: around £10,334 (£20,667/2) on a NMB scale. Allowing for interactions in QALYs increased AIC in 100% of imputed datasets (from -239 to -237, on average), whereas allowing for interactions in costs decreased AIC in 91% of cases (from 3,938 to 3,932, on average). This suggests that allowing for interactions in costs improves model fit but that interactions in QALYs worsen fit.

All three interactions were qualitative, with the incremental costs, QALYs and NMB for mobile bearings changing sign depending on whether patients were allocated to patella resurfacing or no resurfacing (Table 7.3). In particular, patients randomised to mobile bearing with no patella resurfacing accrued substantially higher costs and substantially fewer QALYs than

the other three groups. Mobile bearings therefore dominated fixed bearings in patients who were also randomised to patella resurfacing, but were dominated in patients randomised to no resurfacing. However, making a joint decision about mobile bearings and patella resurfacing based on this analysis and adopting the treatment with the highest expected NMB would still suggest that mobile bearings with patella resurfacing should be recommended – the same conclusion as was drawn from independent assessments in Analysis 1.

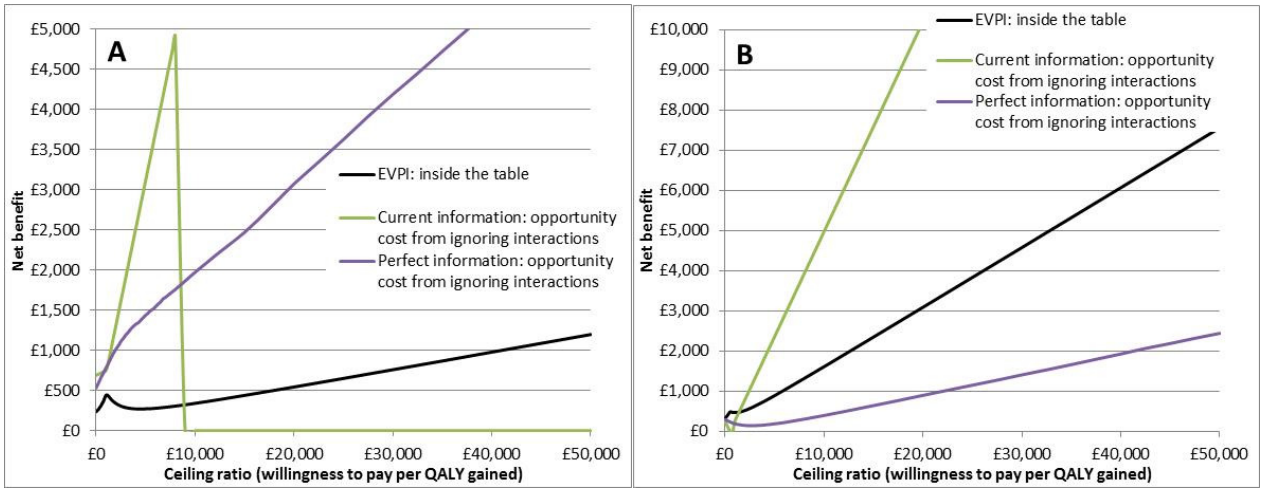
However, despite the smaller sample size, the amount of confidence we can have in this finding is much higher than in Analysis 1 because the expected NMB for mobile bearings with patella resurfacing is much higher than that of the alternatives. Mobile bearings and patella resurfacing have an 86% chance of being cost-effective in this analysis (Figure 7.3A), whereas in Analysis 1, the probability of mobile bearings being cost-effective was only 59%. Similarly, the EVPI associated with the joint decision about patella resurfacing and mobile bearing in Analysis 2 (£545/patient) was also substantially lower than the Analysis 1 estimate of the EVPI to inform the decision about mobile bearings (£726/patient), despite the substantially smaller sample size.

Figure 7.3. CEACs for multiple comparisons based on primary analysis of interactions: including only those patients randomised to >1 comparison. **A**: mobile bearing. **B**: metal backing.



Since inside-the-table analysis and at-the-margins analysis both suggested that mobile bearing with patella resurfacing was the best treatment at ceiling ratios >£8,000/QALY gained, there is no opportunity cost from ignoring interactions under current information at a £20,000/QALY ceiling ratio (Figure 7.4A). However, the two analyses gave different conclusions in 37% of bootstrap replicates, giving an opportunity cost of £3,071 under perfect information.

Figure 7.4. EVPI and expected opportunity cost of ignoring interactions: including only those patients randomised to >1 comparison. **A**: mobile bearing. **B**: metal backing.



Analysing the 145 patients randomised to the metal backing and patella comparisons in the same way also highlighted a significant interaction for QALYs (1.06 [SE: 0.54]; p=0.05) and non-significant qualitative interactions for costs (£506 [SE: £907]; p=0.58) and NMB (£20,788 [SE: £11,094]; p=0.06). This suggests that the bias within at-the-margins estimates of INB could be £10,394 (£20,788/2). In all imputations, allowing for interactions in costs and QALYs increased AIC (from -2,903 to -2,905 for costs and from -302 to -300 for QALYs, on average).

Both costs and QALYs were substantially higher in the group randomised to no metal backing and no patella resurfacing and the group randomised to metal-backing and patella resurfacing than in the other two groups (Table 7.3). Making a joint decision on this basis

would suggest that the treatment maximising NMB is all-polyethylene with no patella resurfacing, with the opposite combination (metal backing with patella resurfacing) coming a close second (Table 7.3). However, the subset of patients randomised in Comparisons B *and* C tended to accrue lower costs and higher QALYs than those randomised to just Comparison B (Tables 7.2-8.3), suggesting that this patient population may not be typical.

There is also substantial uncertainty around this finding (Figure 7.3B), with a 43% chance that metal backing with patella resurfacing has highest NMB and a 50% chance that all-polyethylene with no patella resurfacing is best. Similarly, the EVPI to inform the joint decision about metal backing and patella resurfacing was £3,094/patient (Figure 7.4B): substantially larger than the EVPI to inform the metal backing decision in Analysis 1 (£184/patient).

At-the-margins analysis on the subset of patients randomised to comparisons B and C suggested that all-polyethylene and patella resurfacing were cost-effective at a £20,000/QALY ceiling ratio. On this basis, the opportunity cost of ignoring interactions and making separate decisions in this population is £10,205 (£103,293-£93,087) under current information, although only £965 (£103,293-£102,327) would be gained by adopting the treatment with highest expected NMB based on inside-the-table analysis (all-polyethylene with no resurfacing) rather than the treatment that would be recommended based on Analysis 1 (metal backing with resurfacing). At-the-margins analysis and inside-the-table analysis gave different conclusions in 48% of bootstrap replicates and the opportunity cost of ignoring interactions under perfect information was £4,176.

7.5 Discussion

This study suggests that there is evidence that patella resurfacing may affect the incremental costs, QALYs and cost-effectiveness of mobile bearings and metal backing. Interactions were generally qualitative (changing the direction of incremental effects) and approached statistical significance. Although interactions were not expected a priori, it is plausible that patella resurfacing could affect the movement of the knee joint or the amount of wear in the joint between the femur and patella and therefore affect patients' HRQoL and/or their risk of readmission. However, the exact form of the interactions varies between analyses and is difficult to explain clinically: for example, it is unclear why all-polyethylene with patella resurfacing should have lower NMB than either metal backing with patella resurfacing, or all-polyethylene without resurfacing. Furthermore, only one of the 12 interaction terms estimated was statistically significant at the 0.05 level, suggesting that the observed interactions could be due to chance. Nonetheless, such interactions warrant further investigation.

7.5.1 Which analysis?

Each analysis provided different estimates of incremental costs, QALYs and NMB. Conclusions for Comparison B also varied between analyses: Analysis 2 suggested that all-polyethylene with no patella resurfacing was best, whereas Analysis 1 found metal backing with patella resurfacing to be best. In general, the choice of analysis depends on the decision-maker's research question and the impact of interactions. For knee replacement, surgeons have to simultaneously decide whether or not to resurface the patella and whether to use a mobile bearing, a fixed metal-backed bearing or an all-polyethylene tibia. As such, evaluating the incremental cost-effectiveness of different prosthesis combinations would provide the best evidence to inform this joint decision. However, as discussed in Section 3.5 and Chapter 9, at-the-margins analysis and inside-the-table would give the same conclusions

if there were no interactions, but at-the-margins will give biased and potentially misleading results if interactions are present.

Analysing the subset of patients randomised to >1 comparison inside-the-table (Analysis 2) provides an unbiased estimate of interactions, since all patients are analysed in the groups to which they were randomised, ensuring that there is no systematic difference between groups other than the treatment they received. However, by restricting analysis to those in >1 comparison, this analysis excluded most of the trial population, greatly increasing SEs and substantially reducing the power to detect significant interactions or differences between treatment groups. The small sample size also means that chance fluctuations between patients may have large effects. Nonetheless, EVPI and the probability of making the wrong decision were not necessarily any larger in this analysis as differences in expected NMB were sometimes larger. Furthermore, the subset of patients included in >1 comparison may not be typical of the overall trial population. In this case, patients in Comparisons B and C tended to accrue lower costs and higher QALYs, with the opposite trends among patients in Comparisons A and C. This may mean that the interactions observed in this patient subgroup may not generalise to the wider population. However, the impact of these disadvantages depends on how many patients are randomised to >1 comparison. Since only 15% (338/2,252) of KAT participants were randomised to >1 comparison, inside-the-table analysis is likely to perform worse in KAT than in the other partial factorial trials with economic evaluations, where ≥50% of participants were in >1 comparison (232, 263, 269-271) and where inside-the-table analysis may therefore have greater power and greater generalisability.

By contrast, almost all patients recruited to Comparisons A or B could be considered using the subgroup approach. A sensitivity analysis using this approach (Appendix J) produced smaller SEs and smaller interactions than Analysis 2, since the larger sample size increased statistical power and reduced chance fluctuations in outcomes between cells. However,

evaluating interactions in patients randomised to only one comparison required analysing the patella comparison as-treated, not as-randomised, which introduces substantial selection bias. As result, any observed interactions may be due to patient characteristics (e.g. age, physical activity or severity of bone damage) that affect the chance of surgeons choosing to undertake patella resurfacing as well as the costs and QALYs accrued over the time horizon.[68] This analysis therefore does not inform causal inferences about the main effect of patella resurfacing or interactions between factors. The approach used in Analysis 2 is therefore likely to be the only way to get unbiased estimates of interactions.

However, at-the-margins analysis may provide a useful estimate of average costs and benefits for the population of interest even when interactions exist. This is particularly relevant to KAT, since decisions about those aspects of implant design that are not randomly assigned will reflect routine clinical practice and patients were randomised only to those comparisons about which surgeons are in equipoise. As result, we would expect the proportion of patients having patella resurfacing in KAT to be similar to that in routine clinical practice, regardless of whether they are randomised to Comparison C; this hypothesis is partially supported by National Joint Registry (NJR) data suggesting that 39% of patients undergoing total knee replacement in 2003 (the last year of recruitment to KAT and the first year for which NJR data are available) had patella resurfacing, vs. 50% in KAT. Furthermore, at-the-margins analysis maximises statistical power and generalisability by analysing the entire trial population. Nonetheless, if the observed interaction between metal backing and patella resurfacing is genuine and all-polyethylene with no patella resurfacing truly has higher NMB than metal backing with patella resurfacing, ignoring this interaction and basing decisions on at-the-margins analysis would fail to maximise health gains from the budget.

[68] In principle, we could reduce bias by adjusting for observed confounders (e.g. using propensity scores or genetic matching (289, 290)), although the subgroup approach would remain vulnerable to bias from unobserved confounders.

One alternative approach that would use all the data without assuming additive effects would be to use a Bayesian analysis in which the incremental effects estimated using at-the-margins analysis of patients in one comparison are used as priors in the factorial analysis of those patients randomised to >1 comparison. Preliminary analysis on Comparison B using a vetted bootstrap (213) suggested that this approach is feasible, giving smaller SEs than Analysis 2, finding metal backing with patella resurfacing to be the best treatment and increasing the magnitude of the interaction between patella resurfacing and metal backing. Bayesian methods also enable use of sceptical priors to down-weight interactions (42, 115), which may provide a compromise between including or excluding interaction terms.

7.5.2 Inference and interactions

Claxton previously argued that statistical inference is irrelevant to healthcare adoption decisions as it is based on arbitrary significance levels, favours treatments already in routine use and fails to maximise health gains produced from the healthcare budget (157). This study highlights issues around the relevance of statistical inference for both treatment adoption decisions and identification of important interactions. Firstly, although no comparisons reached statistical significance, I take the view that KAT provides sufficient evidence to conclude that both metal backing and patella resurfacing are good value for money since they have positive expected NMB. However, the case for adopting mobile bearings is less clear, since QALY gains are modest and there is substantial uncertainty.

Secondly, it is unclear how useful statistical inference is as a measure of the importance of interactions. In particular, even full factorial trials are generally under-powered to detect interaction effects for primary outcomes (20), while partial factorial trials will generally have even lower power due to the smaller number of patients in whom interactions can be evaluated. AIC provides one measure that trades the statistical efficiency of the

parsimonious model without interactions against the bias overcome by allowing for interactions. However, it includes only a small penalty for including additional variables and appeared sensitive to the scale of analysis: preliminary analyses found interactions to be important on a NMB scale, but not for costs and QALYs. I also estimated AIC using a separate mixed model analysis, since it is unclear whether it is possible to estimate AIC directly from IPW bootstrap results. Estimates of the likely bias and opportunity cost introduced by ignoring interactions may be more informative, since they directly relate the method of analysis to the decision-making context. These measures suggest that there is a 24-48% probability that ignoring interactions would lead to a different treatment adoption decision and that ignoring interactions could bias INB estimates by up to £10,000.

7.5.3 Conclusions

This study demonstrates that partial factorial trials can be used to evaluate interactions. Such trials are more informative, as well as cheaper, than conducting two separate RCTs on non-overlapping populations. However, compared with full factorials, a partial factorial design reduces the size and generalisability of the sample available for evaluating interactions. Nonetheless, a partial factorial design may substantially increase recruitment in situations where not all patients can be randomised to >1 comparison, thereby increasing the size and generalisability of the sample available for at-the-margins analysis with little/no impact on the number randomised to >1 comparison. Nonetheless, given the difficulties evaluating interactions, partial factorial designs are best reserved for situations where interactions are expected to be negligible.

At-the-margins analysis may provide a useful estimate of average treatment effects for pragmatic partial factorial trials in which the proportion of patients receiving each treatment is likely to be similar to that in routine clinical practice. Nonetheless, at-the-margins analysis is

prone to bias whenever interactions exist (3, 20, 30), so the potential for interactions should always be evaluated in sensitivity analysis. Interactions should be evaluated using ITT inside-the-table analysis on patients randomised to >1 comparison; evaluating interactions on the total sample by subgrouping patients by the treatment they receive is inappropriate as it breaks randomisation and is prone to bias. Bayesian approaches may provide a useful compromise to make use of all available information, while making fewer assumptions about interactions.

Chapter 8: AFORRD – A full 2x2x2 factorial trial

8.1 Introduction

This chapter extends the simulation work and the within-trial analyses of IVAN and KAT to a setting where evaluating the incremental cost-effectiveness of treatment required extrapolation of short-term outcomes using a patient simulation model. Like the worked example discussed in Chapter 3, AFORRD evaluates treatments aiming to reduce the risk of clinical events that in turn affect costs and QALYs: in this case cardiovascular disease events in patients with diabetes (23). AFORRD is typical of published factorial trials accompanied by trial-based economic evaluations (Chapter 4) in that it compares drugs and non-pharmacological interventions against their respective placebos (in contrast to IVAN and KAT which compared different types of active treatment). AFORRD also used a 2x2x2 design (enabling estimation of higher-order interactions) and included one cluster-randomised factor.

8.2 Case study

AFORRD (ISRCTN number: 76737502) comprised a double-blind RCT on 800 adults with type 2 diabetes and no known cardiovascular events, who were managed in primary care and were not thought to require immediate lipid-lowering treatment (23). Patients were randomised to receive 20 mg atorvastatin or matching placebo, and to either omega-3 EE90 (Omacor 2 g/day; Pronova BioPharma Norge AS, Lysaker, Norway, containing 46% eicosapentaenoic acid and 38% docosahexaenoic acid; referred to hereafter as omega-3) or matching placebo. GPs were also cluster-randomised to receive a paper-based behavioural

intervention intended to increase compliance with study medication[69] (referred to hereafter as "leaflet") or standard care.

Cardiovascular risk factors (e.g. low-density lipoprotein [LDL] and triglycerides) were measured at baseline and 16 weeks. All patients who had a >20% chance of cardiovascular events at 16 weeks were then given 20 mg atorvastatin tablets in addition to their initial therapy for the remainder of the trial. To assess the cost-effectiveness of atorvastatin vs. no atorvastatin, the current analysis focuses on the costs and clinical endpoints observed over the first 16 weeks of the trial, before high-risk patients were given additional medication and before the leaflet was reinforced by study nurses.

CUA was conducted to assess the cost-effectiveness of atorvastatin, omega-3 and leaflet alone and in combination, from the perspective of the UK NHS based on within-trial costs and clinical outcomes, which were extrapolated to estimate long-term costs and QALYs (Appendix K). I also aimed to estimate the magnitude of interactions between atorvastatin, omega-3 and leaflet on within-trial costs, events and long-term cost-effectiveness and explore the extent to which the results are sensitive to the assumptions made about interactions. The results presented here are preliminary and may change if a better analytical approach is identified.

[69] The behavioural intervention comprised a leaflet and face-to-face reinforcement/discussion (291). The leaflet was posted to the patient along with other questionnaires two weeks after receiving tablets and starting treatment. It comprised one A4 sheet, which encouraged patients to set and write down an action plan detailing when and where they would take the medication. Research nurses at the centres randomised to receive the behavioural intervention were trained to reinforce the information in the leaflet and discuss the patients' action plan during routine face-to-face visits 18 and 32 weeks after the start of treatment. Nurses at the control centres received no additional training and their patients received only the standard questionnaires.

8.3 Analytical methods

Cardiovascular risk factors observed at Week 16 for each trial participant were extrapolated for 40 years using the UKPDS Diabetes Outcomes Model version 1.3 (Diabetes Trial Unit, Oxford) (210, 292), which predicts the frequency and timing of vascular events and their impact on costs and QALYs. To simplify the analysis for this methodological study, the 68 patients with missing Week 16 data were omitted from the analysis and conditional mean imputation was used for all remaining missing data. Appendix K gives further details on methods for costing, imputation, extrapolation and analysis.

AFORRD was powered assuming that the three interventions have additive effects on clinical endpoints and the factorial design was chosen to address multiple questions without increasing the sample size (293). Indeed, the interventions were evaluated using different primary endpoints.[70] Nonetheless, as described in Section 3.4, we would expect treatments with multiplicative effects on the incidence of clinical events to have non-additive effects on long-term costs and QALYs. Furthermore, we would expect any intervention improving compliance to directly introduce interactions by increasing treatment effects by X%, but have no impact on outcomes with placebo.

The magnitude and significance of interactions for 16-week resource use and costs was evaluated using GLM with gamma family and identity link function.[71] GLM analyses using identity link and Poisson family were also conducted on the number of clinical events within the extrapolated period. Change in triglycerides and LDL were analysed using OLS, since

[70] The primary endpoint comprised LDL for atorvastatin, triacylglycerol for omega-3 and medication compliance for leaflet.

[71] The distributional family was selected using a modified Park test and confirmed using AIC and deviance residual plots (162) conducted on total 16-week cost. Gamma family and logarithmic link gave identical predictions and model fit to gamma family and identity link since the model includes all interaction terms; identity link was therefore used for simplicity. The function fitting total cost best (gamma family and identity link) was then used for all resource use and cost variables to estimate the interaction terms shown in Table 6.2, except in cases where this function did not converge. Modified Park test, AIC and deviance residual plots were also used to select the family and link function for clinical events.

they were symmetrically distributed. All such analyses included dummy variables for each of the three treatments, three interaction terms capturing two-way interactions and one three-way interaction term. Regression analyses were conducted in Stata version 12. For simplicity, SEs were adjusted for clustering by GP using the `cluster` option within Stata, in preference to random-effects models.

The UKPDS model was run with 200,000 Monte Carlo loops and 100 sets of bootstrapped parameters. Within-trial costs and QALYs for each patient were added to the costs and QALYs extrapolated over a 40-year period. Rubin's rule (210, 256) was used to estimate the mean costs and QALYs in each study arm and combine parameter uncertainty around extrapolation with sampling uncertainty around the AFORRD sample (Appendix K). Since this method cannot easily be adapted to consider covariates, meta-regression was used as an approximation within a post hoc secondary analysis that controlled for covariates while combining parameter and sampling uncertainty (Appendix K). The cost of atorvastatin and omega-3 was based on full compliance and was assumed to be accrued until death. The leaflet was assumed to comprise a one-off intervention received only at baseline.

Economic evaluation results are presented for three analyses, which were all evaluated using the same set of extrapolated costs and outcomes for each patient:

1) Base case: This analysis took account of all two- and three-way interactions between the three interventions in the 2x2x2 design in order to minimise the opportunity cost of adopting a suboptimal treatment (Chapter 5). Rubin's rule was used to estimate mean outcomes and SEs for the eight cells of the 2x2x2 design; meta-regression included three main effects, and four interaction terms.
2) 2x2 inside-the-table: This analysis evaluated the impact of atorvastatin and omega-3 and interactions between these two interventions. However it ignored the leaflet, since preliminary clinical analyses suggested that it had no significant effect on compliance and that compliance was generally high (291) (suggesting that the effect on within-trial costs

or 16-week cardiovascular risk will be minimal). Here, means and SEs were calculated for four study arms (pooling patients with and without the leaflet): placebo; atorvastatin alone; omega-3 alone; and atorvastatin+omega-3. Meta-regression included two main effects and one interaction term.

3) At-the-margins: This analysis independently evaluated atorvastatin, omega-3 and leaflet, ignoring all interactions. Mean outcomes and SEs were calculated at-the-margins for three overlapping sets of two patient groups: all those randomised to atorvastatin; all those randomised to atorvastatin placebo; all those randomised to omega-3; all those randomised to omega-3 placebo; all those randomised to leaflet; and all those randomised to no leaflet. Meta-regression included only the three main effects.

Group means and SEs for QALYs, costs, numbers of events and NMB (at a £20,000/QALY ceiling ratio (130)) were calculated using Rubin's rule (Appendix K) and used to estimate the relevant between-group differences and interaction terms for each analysis. Variances around between-group differences and interactions allowed for the correlations between group means that are introduced by using the same risk equation parameters (Appendix K).

For simplicity, neither the primary analysis nor meta-regression adjusted for clustering. Although it is generally recommended that methods such as multi-level models are used to allow for between-centre variations in how cluster-randomised interventions are administered (194, 294), such variations are unlikely within the first 16 weeks of AFORRD, since all patients were posted the same leaflet and no reinforcement was received until after 16 weeks.

The opportunity cost of ignoring interactions based on current information (Section 3.8) was calculated for Analyses 2 and 3 by subtracting the NMB for the treatment that would be

adopted within that analysis from the maximum NMB within Analysis 1.[72] CEACs, EVPI and the opportunity cost of ignoring interactions with perfect information were not calculated due to time constraints and difficulties constructing variance-covariance matrices allowing for correlations between analyses.

8.4 Preliminary results

Regression analyses with three two-way interactions, a three-way interaction and no covariates suggested that patients randomised to atorvastatin had significantly greater reductions in LDL ($p<0.001$) and triglycerides ($p=0.005$). Omega-3 monotherapy was associated with a significant increase in LDL ($p=0.043$) and no effect on triglycerides ($p=0.826$), while leaflet had no significant effect on either ($p<0.410$). There was a significant mixed interaction between atorvastatin and omega-3 for change in LDL ($p=0.038$) and triglycerides ($p=0.037$) and between atorvastatin and leaflet for triglycerides ($p=0.042$), although no other interactions were significant ($p\geq0.070$).

There was some evidence of baseline imbalance (Table 8.1). Patients at GP practices randomised to the leaflet were more likely to be white ($p=0.001$) and have higher total cholesterol ($p=0.004$) and non-significantly lower cardiovascular risk ($p=0.066$). While it is relatively straightforward to allow for covariates in within-trial analyses, the methods for doing so within a patient-level extrapolation model combining parameter and sampling uncertainty are unclear. The base case analysis therefore makes no allowance for covariates.

[72] NMB for both treatments was based on the results of Analysis 1 for consistency. The opportunity cost of ignoring interactions was calculated with and without covariates. The opportunity cost without covariates was estimated by comparing the results of Analysis 2 (or 3) without covariates against the results of Analysis 1 without covariates, while the opportunity cost with covariates was estimated by comparing the results of Analysis 2 (or 3) with covariates against the results of Analysis 1 with covariates. In comparisons between Analyses 1 and 2, it was assumed that the leaflet would not be adopted regardless of the results of Analysis 2 (since it is not used in current clinical practice).

All three interventions individually increased total within-trial cost (Table 8.2), largely due to intervention costs, although only the simple effect for omega-3 was statistically significant (p=0.004).[73] Non-significant qualitative interactions for total within-trial cost were observed between atorvastatin and leaflet (p=0.333) and between omega-3 and leaflet (p=0.347), while the interactions between atorvastatin and omega-3 (p=0.604) and the three-way interaction (p=0.690) were super-additive.

Table 8.1. Baseline characteristics for the eight groups

Treatment group*	**000**	**A00**	**0F0**	**AF0**	**00L**	**A0L**	**0FL**	**AFL**	**p†**
White ethnicity	86% (35%)	91% (29%)	88% (33%)	90% (30%)	96% (20%)	96% (19%)	97% (16%)	95% (22%)	0.021
Female	42% (50%)	33% (47%)	38% (49%)	45% (50%)	45% (50%)	42% (50%)	49% (50%)	45% (50%)	0.462
Age at diagnosis of diabetes (years)	59 (12)	56 (11)	60 (12)	59 (13)	56 (11)	57 (12)	55 (12)	56 (11)	0.109
Duration of diabetes (years)	7 (6)	7 (7)	5 (5)	5 (6)	5 (4)	6 (5)	5 (5)	6 (6)	0.224
Weight (kilograms)	85.8 (17.7)	88.3 (19.7)	86.8 (16.2)	86.5 (20.7)	89.4 (20.2)	88.7 (17.7)	90.4 (21.4)	89.2 (20.9)	0.727
Height (metres)	1.69 (0.09)	1.70 (0.10)	1.69 (0.09)	1.67 (0.10)	1.69 (0.09)	1.69 (0.11)	1.70 (0.10)	1.69 (0.09)	0.644
Atrial fibrillation at screening	3% (17%)	2% (14%)	0% (0%)	0% (0%)	0% (0%)	1% (11%)	1% (11%)	1% (11%)	0.731
Current smoker	12% (33%)	21% (41%)	20% (40%)	13% (33%)	16% (37%)	21% (41%)	14% (35%)	12% (32%)	0.400
Total cholesterol at screening (mmol/L)	4.79 (0.96)	5.00 (0.93)	4.93 (0.86)	5.01 (0.95)	5.25 (0.92)	5.08 (0.77)	4.99 (0.75)	5.19 (0.85)	0.020
High-density lipoprotein at screening (mmol/L)	1.16 (0.30)	1.18 (0.26)	1.16 (0.27)	1.16 (0.27)	1.14 (0.32)	1.12 (0.23)	1.16 (0.29)	1.15 (0.29)	0.952
Systolic blood pressure at screening (mmHg)	145 (17)	143 (17)	141 (17)	141 (16)	145 (15)	145 (16)	141 (14)	143 (17)	0.240
Glycated haemoglobin (HbA1c) at screening (%)	6.84 (1.01)	7.02 (1.04)	7.08 (1.17)	6.84 (1.09)	6.86 (1.08)	7.13 (1.05)	6.87 (1.08)	6.90 (1.06)	0.381
Baseline 10-year cardiovascular risk (%)‡	34.6 (18.2)	33.6 (16.9)	34.0 (15.9)	32.3 (16.9)	31.9 (16.5)	34.2 (17.0)	27.2 (15.7)	32.0 (16.6)	0.126

* Treatment is coded as atorvastatin (A or 0), omega-3 (F or 0), leaflet (L or 0).
† Global tests for differences between groups, based on regression analyses with three main effects and four interactions (F test based on OLS for continuous endpoints and chi-squared test based on logistic regression for binary endpoints). Omega-3*leaflet and three-way interactions were omitted from the analysis on atrial fibrillation to enable convergence.
‡ Calculated using the UKPDS risk engine (295, 296).

[73] For consistency with the base case analysis, no adjustment was made for baseline imbalance.

Table 8.2. Costs accrued within the 16-week randomised trial period. Values for each cell or treatment-combination represent crude mean ±SD across the patients randomised to that treatment-combination who had 16-week data. Interactions are shown as mean ±SE; these were calculated using GLM with gamma family and identity link (not allowing for covariates), with the exception of cost of hospitalisations and study medication and the number of inpatient days, which were calculated using GLM with Gaussian family since the model with gamma family did not converge.

		Ambulatory consultations		Concomitant medications		Hospitalisations		Study treatment	Total within-trial cost
		Number of visits	Cost	Number of days	Cost	Number of days	Cost	Cost	
No treatment		4.28 (2.59)	£142 (£92)	563 (343)	£134 (£114)	0.086 (0.709)	£40 (£298)	£0 (£0)	£316 (£358)
Atorvastatin		4.10 (2.60)	£139 (£107)	492 (273)	£116 (£92)	0.038 (0.306)	£64 (£503)	£6 (£0)	£326 (£538)
Omega-3		4.96 (4.61)	£164 (£158)	508 (313)	£120 (£125)	0.140 (1.177)	£70 (£444)	£114 (£0)	£468 (£498)
Atorvastatin+omega-3		4.62 (3.21)	£162 (£118)	552 (404)	£136 (£139)	0.255 (1.874)	£111 (£751)	£120 (£0)	£529 (£786)
Leaflet		3.27 (2.28)	£109 (£79)	500 (299)	£116 (£96)	0.082 (0.595)	£120 (£787)	£5 (£0)	£346 (£821)
Atorvastatin+leaflet		3.98 (3.41)	£117 (£123)	515 (344)	£122 (£118)	0.000 (0.000)	£0 (£0)	£12 (£0)	£245 (£192)
Omega-3+leaflet		4.56 (3.90)	£138 (£114)	440 (316)	£111 (£99)	0.039 (0.344)	£35 (£308)	£119 (£0)	£398 (£338)
Atorvastatin+omega-3+leaflet		4.76 (3.65)	£135 (£110)	517 (303)	£119 (£101)	0.013 (0.113)	£27 (£235)	£126 (£0)	£400 (£292)
Inter-action	Atorvastatin*omega-3	-0.15 (0.72)	£1 (£23)	116 (62)	£33 (£20)	0.162 (0.127)	£16 (£78)	£0 (£0)	£51 (£98)
	Atorvastatin*leaflet	0.90 (0.63)	£10 (£21)	86 (65)	£23 (£24)	-0.034 (0.103)	-£145 (£108)	£0 (£0)	-£111 (£115)
	Omega-3*leaflet	0.61 (0.73)	£7 (£23)	-4 (52)	£9 (£20)	-0.097 (0.156)	-£115 (£112)	£0 (£0)	-£100 (£106)
	Atorvastatin*omega-3*leaflet	-0.36 (0.96)	-£11 (£31)	-54 (85)	-£31 (£28)	-0.107 (0.151)	£96 (£128)	£0 (£0)	£53 (£133)

* Statistically significant interaction ($p<0.05$).

Patients randomised to atorvastatin had significantly fewer diabetic events during the 40-year extrapolated period ($p<0.001$), although omega-3 and leaflet had no significant effect when baseline covariates[74] were taken into account ($p\geq0.820$). For diabetic events, there were no significant interactions between interventions on either a logarithmic scale ($p\geq0.195$), or a natural scale ($p\geq0.263$) and all interactions were smaller than the simple effect of atorvastatin.

8.4.1 Analysis 1 results: 2x2x2 (base case)

The base case analysis (allowing for all interactions) found that omega-3 significantly increased total costs ($p<0.001$) and that atorvastatin significantly increased QALYs relative to no treatment ($p=0.035$; Table 8.3). However, very large qualitative interactions were observed between atorvastatin and leaflet for non-intervention costs, life-years, QALYs and NMB, which (if real) have a large effect on the results. In particular, both omega-3+leaflet and leaflet alone were found to have higher QALYs, life-years and NMB than any other treatment arms, including those receiving atorvastatin (Table 8.3, Figure 8.1). This finding is contradicted by meta-analyses of long-term trials demonstrating that statins reduce cardiovascular events (297, 298) and the clinical findings of AFORRD (which suggest that atorvastatin reduced cardiovascular risk (23), while the leaflet had minimal impact on compliance (291)). In addition to suggesting that a one-page action plan is more effective than atorvastatin, the results also suggest that the leaflet significantly increases QALYs in the absence of active treatment.

[74] The covariates included in the model were: white race; female gender; age at diagnosis; duration of diabetes; weight; height; atrial fibrillation; current smoker; total cholesterol; high-density lipoprotein; systolic blood pressure; and glycated haemoglobin (HbA1c). All parameters were based on the screening visit (which normally took place two weeks before randomisation and start of treatment).

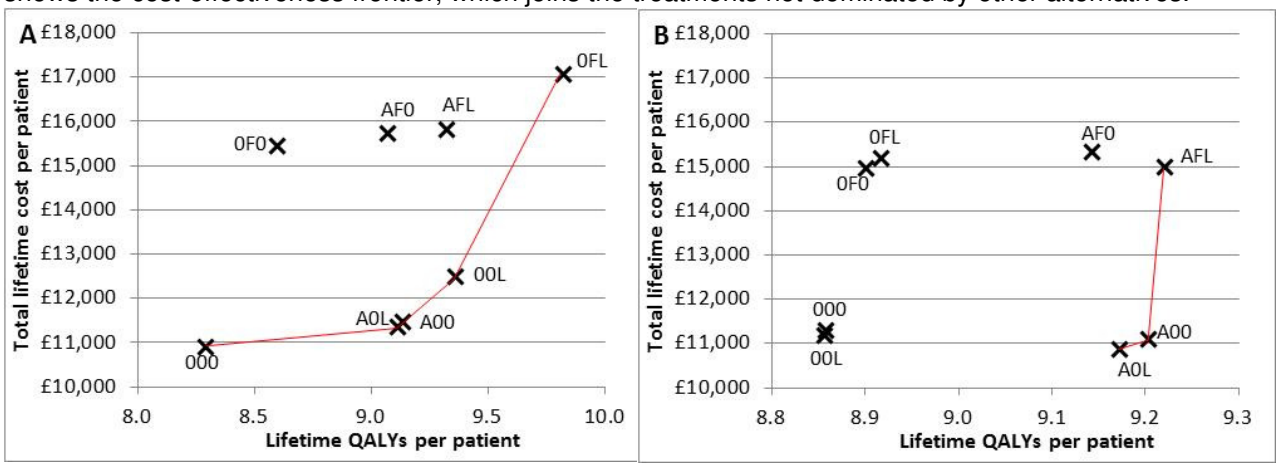
Table 8.3. Analysis 1 (without covariates): Base case total costs and QALYs over 40.3 years, including interactions between atorvastatin, omega-3 and leaflet. All values are discounted at 3.5% per annum and represent mean ±SE.

		Study treat-ment†	Compl-ication costs†	Within -trial costs‡	Total costs	Life-years	QALYs	NMB§
No treatment		£0 (£0)	£10,586 (£2,533)	£316 (£35)	£10,902 (£2,533)	10.89 (0.74)	8.29 (0.56)	£154,932 (£10,914)
Atorvastatin		£243 (£15)	£10,884 (£2,499)	£326 (£52)	£11,453 (£2,502)	11.95 (0.74)	9.14 (0.55)	£171,248 (£10,736)
Omega-3		£4,084 (£256)	£10,895 (£2,150)	£468 (£48)	£15,448 (£2,242)	11.31 (0.69)	8.60 (0.52)	£156,474 (£9,880)
Atorvastatin+omega-3		£4,535 (£280)	£10,677 (£2,343)	£529 (£75)	£15,741 (£2,441)	11.87 (0.71)	9.07 (0.54)	£165,722 (£10,411)
Leaflet		£0 (£0)	£12,151 (£2,527)	£346 (£96)	£12,497 (£2,529)	12.30 (0.71)	9.36 (0.54)	£174,709 (£10,377)
Atorvastatin+leaflet		£243 (£16)	£10,867 (£2,424)	£245 (£22)	£11,355 (£2,428)	11.93 (0.76)	9.11 (0.57)	£170,941 (£11,093)
Omega-3+leaflet		£4,664 (£275)	£12,015 (£2,912)	£398 (£39)	£17,076 (£2,997)	12.87 (0.74)	9.82 (0.56)	£179,395 (£10,858)
Atorvastatin+omega-3+ leaflet		£4,660 (£295)	£10,751 (£2,345)	£400 (£33)	£15,812 (£2,450)	12.19 (0.75)	9.32 (0.57)	£170,644 (£10,958)
Difference: Atorvastatin vs. placebo	No omega-3 or leaflet	£243 (£15)*	£298 (£661)	£10 (£63)	£551 (£658)	1.06 (0.51)*	0.84 (0.40)*	£16,315 (£7,676)*
	With omega-3 no leaflet	£451 (£201)*	-£218 (£665)	£61 (£89)	£293 (£813)	0.56 (0.52)	0.48 (0.41)	£9,248 (£7,577)
	No omega-3 with leaflet	£243 (£16)*	-£1,284 (£670)	-£101 (£99)	-£1,143 (£681)	-0.36 (0.60)	-0.25 (0.47)	-£3,767 (£9,096)
	With omega-3 and leaflet	-£4 (£215)	-£1,263 (£940)	£2 (£51)	-£1,264 (£1,042)	-0.68 (0.56)	-0.50 (0.44)	-£8,750 (£8,176)
Difference: Omega-3 vs. placebo	No atorvastatin or leaflet	£4,084 (£256)*	£309 (£766)	£152 (£59)*	£4,546 (£805)*	0.41 (0.53)	0.30 (0.41)	£1,542 (£7,775)
	With atorvastatin no leaflet	£4,292 (£269)*	-£207 (£653)	£203 (£91)*	£4,288 (£713)*	-0.08 (0.51)	-0.06 (0.40)	-£5,526 (£7,538)
	No atorvastatin with leaflet	£4,664 (£275)*	-£137 (£859)	£52 (£104)	£4,579 (£958)*	0.57 (0.58)	0.46 (0.45)	£4,686 (£8,604)
	With atorvastatin and leaflet	£4,418 (£285)*	-£116 (£652)	£156 (£40)*	£4,457 (£727)*	0.26 (0.59)	0.21 (0.46)	-£297 (£8,720)
Difference: leaflet vs. no leaflet	Placebo+ placebo	£0 (£0)	£1,566 (£662)*	£30 (£102)	£1,595 (£668)*	1.40 (0.55)*	1.07 (0.43)*	£19,777 (£8,332)*
	With atorvastatin +placebo	£0 (£12)	-£16 (£665)	-£82 (£57)	-£98 (£667)	-0.02 (0.58)	-0.02 (0.45)	-£306 (£8,562)
	Placebo+ omega-3	£579 (£206)*	£1,119 (£1,077)	-£70 (£62)	£1,629 (£1,178)	1.56 (0.56)*	1.23 (0.43)*	£22,921 (£8,063)*
	Atorvastatin and omega-3	£125 (£207)	£75 (£621)	-£128 (£82)	£71 (£770)	0.32 (0.53)	0.25 (0.41)	£4,923 (£7,715)
Interaction	Atorvastatin *omega-3	£208 (£200)	-£516 (£1,019)	£51 (£109)	-£258 (£1,121)	-0.50 (0.74)	-0.37 (0.58)	-£7,068 (£10,866)
	Atorvastatin * leaflet	£0 (£12)	-£1,582 (£1,019)	-£111 (£117)	-£1,693 (£1,024)	-1.43 (0.80)	-1.09 (0.62)	-£20,083 (£11,972)

		Study treat-ment†	Compl-ication costs†	Within -trial costs‡	Total costs	Life-years	QALYs	NMB§
	Omega-3*leaflet	£579 (£206)*	-£446 (£1,329)	-£100 (£120)	£33 (£1,417)	0.16 (0.79)	0.16 (0.62)	£3,144 (£11,738)
	Atorvastatin *omega-3*leaflet	-£454 (£294)	£537 (£1,802)	£53 (£156)	£136 (£1,926)	0.18 (1.12)	0.11 (0.87)	£2,085 (£16,586)

* Statistically significant (p<0.05).
† During the 40-year extrapolated period
‡ During the 16-week trial period
§ At a £20,000/QALY ceiling ratio

Figure 8.1. Point estimates from Analysis 1 (base case 2x2x2) plotted on the cost-effectiveness plane: **A** Group means with no adjustment for covariates; **B** Meta-regression with covariates. The red line shows the cost-effectiveness frontier, which joins the treatments not dominated by other alternatives.



A secondary analysis used meta-regression to give an initial indication of whether this unexpected finding was due to between-group imbalance in baseline characteristics. Meta-regression underestimated SEs, but gave broadly similar results to Table 8.3 when no covariates were considered (Table 8.4). However, controlling for baseline risk factors completely changed the conclusions, reducing the magnitude of all simple effects and all but one interaction term.

After controlling for covariates, neither leaflet nor omega-3 significantly affected QALYs (p≥0.234), although omega-3 continued to have the greatest effect on cost (p<0.001) and atorvastatin had greatest effect on QALYs (p<0.001). In particular, controlling for covariates reduced QALYs for the omega-3+leaflet and leaflet alone groups and increased mean QALYs in the no treatment group (Figure 8.1). Within this analysis, atorvastatin alone

(with/without leaflet) dominated no treatment and atorvastatin+omega-3+leaflet cost £231,624/QALY vs. atorvastatin alone. At a £20,000/QALY ceiling ratio, we would therefore adopt atorvastatin alone, since this has highest NMB. By contrast, the analysis without covariates suggested that omega-3+leaflet had highest NMB (Table 8.3).

Table 8.4. Analysis 1: Results of meta-regression analyses with all interactions, with and without covariates. Meta-regression was conducted using the `megareg` command to pool predicted outcomes for each patient, weighted by the between-bootstrap SE for each patient.

		No covariates			Controlling for baseline covariates†		
		Total costs	**QALYs**	**NMB‡**	**Total costs**	**QALYs**	**NMB‡**
Atorvastatin simple effect		£279 (£460)	0.85 (0.35)*	£16,444 (£6,543)*	-£193 (£204)	0.35 (0.06)*	£6,999 (£1,378)*
Omega-3 simple effect		£4,126 (£456)*	0.32 (0.35)	£1,745 (£6,516)	£3,662 (£199)*	0.04 (0.06)	-£3,618 (£1,341)*
Leaflet simple effect		£1,527 (£514)*	1.07 (0.38)*	£19,751 (£7,226)*	-£107 (£233)	0.00 (0.07)	-£158 (£1,481)
Interaction	Atorvastatin* omega-3	£179 (£644)	-0.37 (0.49)	-£7,209 (£9,178)	£559 (£286)	-0.10 (0.09)	-£2,220 (£1,912)
	Atorvastatin*leaflet	-£1,478 (£719)*	-1.09 (0.54)*	-£20,055 (£10,137)*	-£129 (£321)	-0.03 (0.10)	-£262 (£2,109)
	Omega-3*leaflet	-£60 (£722)	0.15 (0.54)	£2,916 (£10,140)	£344 (£326)	0.02 (0.09)	£645 (£2,078)
	Atorvastatin* omega-3*leaflet	£244 (£1,011)	0.12 (0.75)	£2,205 (£14,241)	-£431 (£457)	0.09 (0.13)	£1,628 (£2,958)

* Statistically significant ($p<0.05$).
† The covariates included in the model were: white race; female gender; age at diagnosis; duration of diabetes; weight; height; atrial fibrillation; current smoker; total cholesterol; high-density lipoprotein; systolic blood pressure; and glycated haemoglobin (HbA1c). All parameters were based on the screening visit (which normally took place two weeks before randomisation and start of treatment).
‡ At a £20,000/QALY ceiling ratio

8.4.2 Analysis 2 results: 2x2 inside-the-table

When AFORRD was analysed as a 2x2 trial with no covariates, considering only atorvastatin and omega-3, there were qualitative interactions between atorvastatin and omega-3 for complication costs and within-trial costs, but not total costs, life-years, QALYs or NMB (Table 8.5; Figure 8.2). Interactions had similar magnitude to the atorvastatin*omega-3 interactions in Analysis 1, although SEs were slightly smaller. Based on this analysis, we would adopt atorvastatin monotherapy at a £20,000/QALY ceiling ratio. Atorvastatin+omega-3 generated

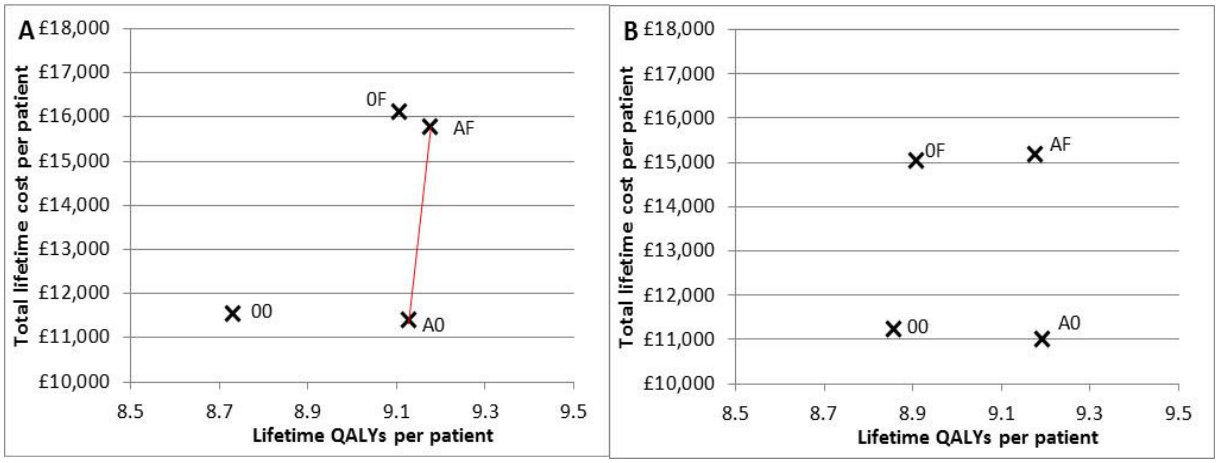
non-significantly more QALYs (p=0.868), but cost £86,813/QALY compared with atorvastatin alone.

Table 8.5. Analysis 2 (without covariates): Total costs and QALYs over 40.3 years, allowing for interactions between atorvastatin and omega-3, but ignoring the compliance leaflet intervention. All values are discounted at 3.5% per annum and represent mean ±SE.

		Study treat-ment†	Compl-ication costs†	Within-trial costs‡	Total costs	Life-years	QALYs	NMB§
Placebo+placebo		£0 (£0)	£11,228 (£2,514)	£328 (£44)	£11,556 (£2,514)	11.47 (0.68)	8.73 (0.51)	£163,043 (£9,928)
Atorvastatin+placebo		£243 (£14)	£10,877 (£2,450)	£292 (£32)	£11,412 (£2,453)	11.94 (0.69)	9.13 (0.52)	£171,119 (£10,039)
Placebo+omega-3		£4,325 (£244)	£11,360 (£2,469)	£439 (£32)	£16,124 (£2,548)	11.95 (0.66)	9.11 (0.49)	£165,993 (£9,551)
Atorvastatin+omega-3		£4,587 (£268)	£10,708 (£2,328)	£476 (£46)	£15,771 (£2,419)	12.00 (0.68)	9.18 (0.51)	£167,764 (£9,948)
Difference: Atorvastatin vs. placebo	No omega-3	£243 (£14)*	-£351 (£520)	-£36 (£55)	-£145 (£520)	0.48 (0.39)	0.40 (0.31)	£8,076 (£5,923)
	With omega-3	£262 (£150)	-£652 (£543)	£37 (£56)	-£354 (£638)	0.05 (0.39)	0.07 (0.30)	£1,771 (£5,672)
Difference: Omega-3 vs. placebo	No atorvastatin	£4,325 (£244)*	£132 (£548)	£111 (£55)*	£4,568 (£601)*	0.49 (0.40)	0.38 (0.31)	£2,950 (£5,864)
	With atorvastatin	£4,344 (£256)*	-£169 (£509)	£184 (£56)*	£4,359 (£549)*	0.06 (0.39)	0.05 (0.30)	-£3,355 (£5,720)
Interaction		£19 (£149)	-£301 (£838)	£73 (£78)	-£209 (£907)	-0.43 (0.56)	-0.33 (0.44)	-£6,305 (£8,281)

* Statistically significant (p<0.05).
† During the 40-year extrapolated period
‡ During the 16-week trial period
§ At a £20,000/QALY ceiling ratio

Figure 8.2. Point estimates from Analysis 2 (2x2) plotted on the cost-effectiveness plane: **A** Group means with no adjustment for covariates; **B** Meta-regression with covariates. The red line shows the cost-effectiveness frontier, which joins the treatments that are not dominated by other alternatives; no frontier is shown on **B** since A0 dominates all alternatives.



Controlling for baseline risk factors using meta-regression increased the statistical significance of the simple effect of atorvastatin on QALYs and reduced the magnitude of the coefficients for omega-3 and the atorvastatin*omega-3 interaction (Table 8.6), but did not change the conclusions at ceiling ratios <£86,000/QALY.

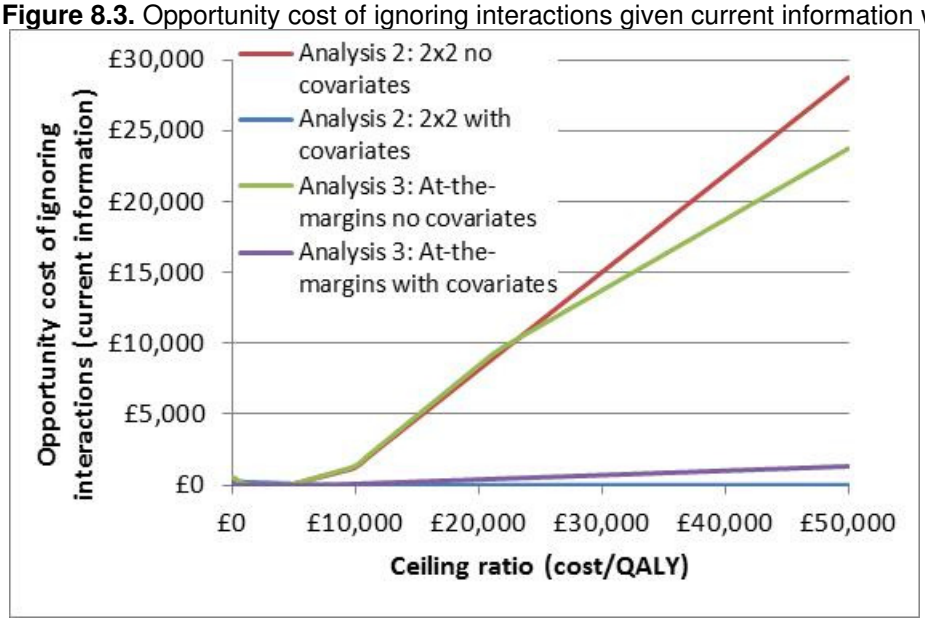
Table 8.6. Analysis 2: Results of meta-regression analyses of AFORRD as a 2x2 trial, with and without covariates.

	No covariates			Controlling for baseline covariates†		
	Total costs	**QALYs**	**NMB‡**	**Total costs**	**QALYs**	**NMB‡**
Atorvastatin simple effect	-£308 (£355)	0.40 (0.27)	£8,177 (£5,015)	-£253 (£158)	0.33 (0.05)*	£6,888 (£1,043)*
Omega-3 simple effect	£4,105 (£355)*	0.38 (0.27)	£2,987 (£5,010)	£3,789 (£159)*	0.05 (0.05)	-£3,350 (£1,023)*
Interaction: Atorvastatin* omega-3	£272 (£498)	-0.33 (0.37)	-£6,329 (£7,042)	£405 (£224)	-0.07 (0.07)	-£1,557 (£1,461)

* Statistically significant ($p<0.05$).
† The covariates included in the model were: white race; female gender; age at diagnosis; duration of diabetes; weight; height; atrial fibrillation; current smoker; total cholesterol; high-density lipoprotein; systolic blood pressure; and glycated haemoglobin (HbA1c). All parameters were based on the screening visit (which normally took place two weeks before randomisation and start of treatment).
‡ At a £20,000/QALY ceiling ratio

Comparing the results of Analyses 1 and 2 without covariates suggested that ignoring interactions has an opportunity cost of £8,147/patient (given current information), since Analysis 2 would lead us to adopt atorvastatin monotherapy, whereas Analysis 1 suggests that omega-3+leaflet has higher NMB. The opportunity cost is <£1,000 at ceiling ratios <£10,000/QALY, but rises linearly thereafter (Figure 8.3). However, as discussed above, controlling for covariates completely changes the conclusions of Analysis 1. Comparing the results of Analyses 1 and 2 controlling for covariates using meta-regression suggests that there is no opportunity cost of ignoring interactions at ceiling ratios between £8,000 and £231,624/QALY (Figure 8.3).

Figure 8.3. Opportunity cost of ignoring interactions given current information within Analyses 2 and 3



8.4.3 Analysis 3 results: At-the-margins

Ignoring all interactions and conducting at-the-margins analysis with no covariates suggests that atorvastatin dominates no atorvastatin, omega-3 costs £21,169/QALY vs. no omega-3, while leaflet costs £1,252/QALY vs. standard care (Table 8.7). If we were to make independent decisions about the three interventions at a £20,000/QALY ceiling ratio, we would therefore adopt atorvastatin+leaflet. Comparing the results of this analysis against Analysis 1 with no covariates suggests that the opportunity cost of ignoring interactions and conducting at-the-margins analysis is £8,453/patient (Figure 8.3). SEs around almost all of the main effects estimated in Analysis 3 were also markedly smaller than those around the simple effects estimated in Analysis 1 (Tables 8.3 and 8.7); however, this was not uniformly true due to the heteroskedasticity for intervention costs.

However, controlling for covariates changed the results and conclusions of this analysis, reducing the p-value for atorvastatin and reducing the main effects of omega-3 and leaflet by >10-fold (Table 8.8). Meta-regression with covariates suggested that atorvastatin+leaflet had highest NMB at a £20,000/QALY ceiling ratio and strongly dominated all alternatives other

than atorvastatin+omega-3+leaflet, which cost £219,740/QALY compared with atorvastatin+leaflet. Both at-the-margins analysis with covariates and Analysis 1 with covariates would lead to adoption of atorvastatin+leaflet at ceiling ratios <£8,000/QALY, although the opportunity cost of ignoring interactions for the comparison between these analyses rose slowly with increasing ceiling ratio, equalling £386/patient at £20,000/QALY.

Table 8.7. Analysis 3 (without covariates): Total costs and QALYs over 40.3 years, excluding interactions. All values are discounted at 3.5% per annum and represent mean ±SE.

	Study treat-ment†	Compl-ication costs†	Within-trial costs‡	Total costs	Life-years	QALYs	NMB§
Atorvastatin placebo	£2,192 (£201)	£11,295 (£2,481)	£384 (£27)	£13,872 (£2,520)	11.71 (0.64)	8.92 (0.48)	£164,538 (£9,297)
Atorvastatin	£2,444 (£216)	£10,791 (£2,380)	£385 (£28)	£13,620 (£2,427)	11.97 (0.66)	9.15 (0.49)	£169,419 (£9,584)
Difference: Atorvastatin vs. placebo	£252 (£178)	-£504 (£429)	£1 (£40)	-£251 (£491)	0.26 (0.28)	0.23 (0.22)	£4,881 (£4,142)
Omega-3 placebo	£123 (£12)	£11,050 (£2,472)	£310 (£27)	£11,483 (£2,474)	11.71 (0.66)	8.93 (0.49)	£167,137 (£9,546)
Omega-3	£4,458 (£246)	£11,029 (£2,388)	£458 (£28)	£15,945 (£2,467)	11.98 (0.64)	9.14 (0.48)	£166,890 (£9,345)
Difference: Omega-3 vs. placebo	£4,335 (£237)*	-£21 (£426)	£148 (£39)*	£4,462 (£452)*	0.27 (0.28)	0.21 (0.22)	-£246 (£4,138)
Standard care	£2,247 (£199)	£10,760 (£2,358)	£411 (£28)	£13,419 (£2,399)	11.51 (0.65)	8.78 (0.48)	£162,131 (£9,416)
Leaflet	£2,423 (£223)	£11,433 (£2,528)	£347 (£27)	£14,203 (£2,574)	12.32 (0.65)	9.40 (0.49)	£173,883 (£9,464)
Difference: leaflet vs. standard care	£176 (£182)	£673 (£452)	-£64 (£39)	£785 (£514)	0.81 (0.28)*	0.63 (0.22)*	£11,752 (£4,135)*

* Statistically significant (p<0.05).
† During the 40-year extrapolated period
‡ During the 16-week trial period
§ At a £20,000/QALY ceiling ratio

Table 8.8. Analysis 3: Results of meta-regression analyses at-the-margins, with and without covariates.

	No covariates			Controlling for baseline covariates†		
	Total costs	QALYs	NMB‡	Total costs	QALYs	NMB‡
Atorvastatin main effect	-£180 (£249)	0.23 (0.19)	£4,913 (£3,507)	-£47 (£112)	0.30 (0.03)*	£6,095 (£728)*
Omega-3 main effect	£4,249 (£249)*	0.21 (0.19)	-£183 (£3,507)	£3,993 (£112)*	0.02 (0.03)	-£4,103 (£734)*
Leaflet main effect	£803 (£253)*	0.62 (0.19)*	£11,641 (£3,558)*	-£116 (£116)	0.02 (0.03)	£449 (£753)

* Statistically significant (p<0.05).
† The covariates included in the model were: white race; female gender; age at diagnosis; duration of diabetes; weight; height; atrial fibrillation; current smoker; total cholesterol; high-density lipoprotein; systolic blood pressure; and glycated haemoglobin (HbA1c). All parameters were based on the screening visit (which normally took place two weeks before randomisation and start of treatment).
‡ At a £20,000/QALY ceiling ratio

8.5 Discussion

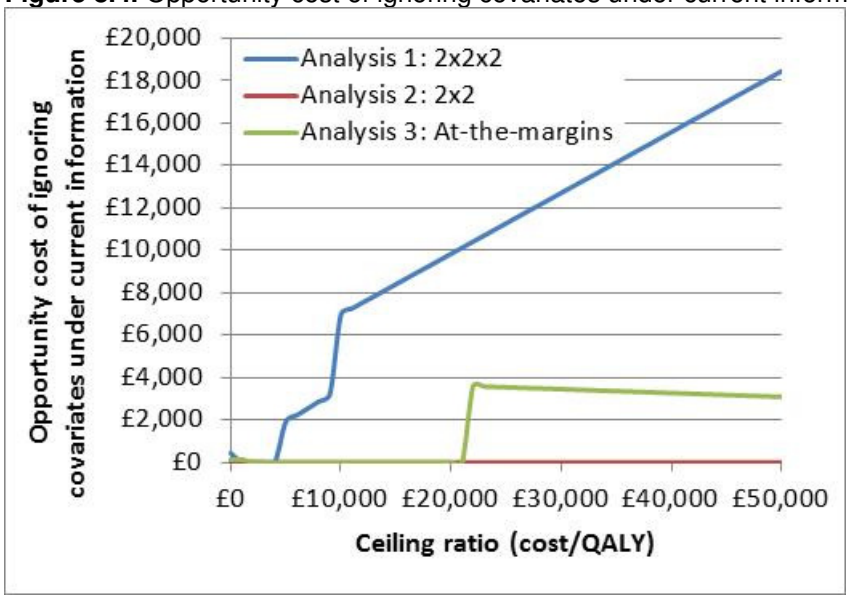
The results of AFORRD were sensitive to the interactions included, but were still more sensitive to baseline characteristics: particularly when the trial was analysed as a 2x2x2 design. Compared with other studies, unexpected results were easier to identify in AFORRD (since it evaluated one treatment with established effectiveness and two with minimal effect) and the influential prognostic variables (those in the UKPDS) were readily available and easy to identify. This raises the question of how frequently baseline imbalance influences the conclusions of other trial-based economic evaluations (factorial or otherwise): particularly since the imbalance was not particularly pronounced for individual covariates (Table 8.1). However, studies extrapolating trial results using a prognostic model are likely to be more sensitive to imbalance in baseline prognostic variables, since these are directly incorporated in estimates of lifetime costs and QALYs. Just as imbalance in baseline utility introduces bias (not just chance variations) within unadjusted QALY calculations (150), imbalance in pre-treatment prognostic factors is likely to introduce bias within AFORRD.

Although controlling for covariates changed point estimates for all three analyses, it had most effect on Analysis 1. This may suggest that factorial trials may be more sensitive to baseline imbalance: particularly those like AFORRD that have numerous study arms and relatively few patients per arm, where there is a high chance that some arms may, by chance, have a higher proportion of healthy patients. However, at-the-margins analysis treating AFORRD as overlapping two-arm trials was also sensitive to covariates. Cluster-randomised trials may be particularly sensitive if baseline variables for patients at the same centre are correlated.

Meta-regression with covariates may give a better indication of cost-effectiveness than unadjusted group means. This would suggest that patients with diabetes should receive atorvastatin without omega-3 or leaflet. Assuming that meta-regression with covariates gives the true NMB for each treatment, we can calculate the opportunity cost of ignoring *covariates*

under current information (Appendix K) to be £9,838/patient for Analysis 1 (Figure 8.4). However, the results of meta-regression should also be interpreted with caution, since this gives lower weight to patients with high between-bootstrap SEs, who tend to be those with longer life expectancy. Meta-regression also assumes that outcomes are independent between patients, ignoring correlations introduced by risk equation parameters common to all patients. As result, meta-regression gives less robust estimates of unadjusted group means than the primary analysis and underestimates SEs.

Figure 8.4. Opportunity cost of ignoring covariates under current information



As result, neither of the analyses presented in this chapter are entirely appropriate and alternative methods for adjusting UKPDS model output for covariates are needed to correctly analyse the AFORRD data. If the matrix of outcomes for all bootstraps on all patients were available, this could be done relatively easily using the `mim` or `micombine` Stata commands, although it is unclear whether any such adjustment can be made on the model output currently available. Given these constraints, centring data inputs before running the model may be the easiest solution.[75]

[75] Continuous risk factors could be centred by subtracting the mean baseline values for each study arm and adding the grand mean baseline values. For example, if patient 1 received atorvastatin monotherapy and reduced total cholesterol from 4.5 to 3.8 mmol/L, their baseline cholesterol could be adjusted to 4.4 by subtracting the mean baseline cholesterol for the atorvastatin-only group (e.g. 5.1)

The analysis with no adjustment for covariates highlights the fact that comparisons between analyses with and without interactions can be misleading if the inside-the-table analysis against which results are compared is prone to error. In particular, the opportunity cost of ignoring interactions is likely to be unrealistically high in analyses without covariates since the baseline imbalance appears to have biased estimates of the INB for omega-3+leaflet vs. atorvastatin±leaflet. After adjusting for covariates, interactions were surprisingly small for two treatments affecting the risk of clinical events and a leaflet intended to improve compliance with both drugs: particularly since the UKPDS model is based on a series of proportional hazards models. The small interactions are most likely to be due to the small main effects of omega-3 and leaflet (4, 30, 35).

The conclusions also should be interpreted with caution because outcomes are extrapolated over a period of 120 times longer than the study period. However, the UKPDS model takes account of the uncertainty around the extrapolation and the importance of interactions is unlikely to be affected by this limitation. Within-trial QALYs also cannot be accurately assessed as HRQoL was not measured at 16 weeks.

I excluded patients with missing 16-week data and used conditional mean imputation to impute missing 52-week data, rather than using multiple imputation. Conducting multiple imputation would have substantially complicated the analysis, requiring each patient to be duplicated M times within the UKPDS model, followed by a more complicated series of Rubin's rule calculations to combine uncertainty around imputation, extrapolation and

---

and adding the grand mean (e.g. 5.0), while their 16-week cholesterol could be adjusted to 3.7 by the same method. Although the same approach cannot be used for binary/categorical variables (sex, smoking, ethnicity or atrial fibrillation), in principle the model could be run for the whole cohort with all combinations of the most important binary variables and outcomes could be weighted by the overall proportion of patients with each combination of variables. An alternative would be to use Monte Carlo simulations to approximate the matrix of bootstrap replicates for each patient by drawing values from the 95% CI for each patient that allowed for correlations between patients and between costs, QALYs and life-years and then analyse the results using `mim`.

sampling. However, this would not have materially changed the conclusions with respect to the factorial design.

To my knowledge, this is the first application of Rubin's rule to combine parameter uncertainty and sampling uncertainty within the Excel version of the UKPDS Outcomes Model, although this was conducted using the original Stata model (210). My method makes several approximations to overcome the limited data recorded for each bootstrap replicate within the UKPDS software version 1.3 (Appendix K). Nonetheless, this method could be applied in other trials, although the inability to control for covariates is a major shortcoming. Taking account of both parameter uncertainty and sampling uncertainty is particularly important in this setting as it enabled me to estimate the efficiency gains from ignoring interactions.

Like 20% of factorial trial-based economic evaluations (Chapter 4), AFORRD includes cluster randomisation. In general, multi-level models or two-stage bootstrap procedures are the most robust methods for dealing with clustering (294). These were not used here to simplify the analysis and as the delivery of the cluster-randomised intervention (leaflet) is unlikely to have varied between centres during the 16-week period analysed. It is also unclear how to adjust either the primary analysis or meta-regression for clustering. More generally, parametric models allowing for clustering can easily be adjusted to allow for interactions by simply adding in an interaction term (as conducted by (116)). However, two-stage bootstrapping is likely to be problematic for trials where ≥1 factor (e.g. A and B) is cluster-randomised and ≥1 other (e.g. C) is individually-randomised within clusters. Two-stage bootstrapping involves first sampling clusters (e.g. GPs) with replacement and then sampling patients with replacement from within each cluster (294). Allowing for interactions requires that each cell in the factorial design is bootstrapped independently. In principle this could be conducted within a two-stage bootstrap by individually sampling the patients randomised to *0*, *a*, *b* or *ab* from within each cluster. For each bootstrap replicate, the mean costs and outcomes for patients

in group *abc* can then be calculated by averaging over the patients in clusters randomised to *c* who were also individually randomised to *a* and *b*.

Chapter 9: Decision-making for healthcare resource allocation: Joint vs. separate decisions on interacting interventions

9.1 Objectives

As discussed in Section 3.5, standard guidance for allocating healthcare resources based on cost-effectiveness recommends using different decision rules for independent and mutually-exclusive alternatives (16, 125, 128, 169-177), although the definition of "mutually-exclusive" varies. Given a set of independent alternatives, we maximise the health gains from the budget by directly comparing ICERs in a league table and adopting multiple options from the lowest ICER upwards.[76] Given a set of mutually-exclusive alternatives, dominated alternatives must be excluded from consideration before adopting the treatment that maximises NMB at each ceiling ratio. At-the-margins analysis of factorial trials naturally lends itself to using the decision rules for independent treatments, while inside-the-table naturally aids consideration of treatment combinations incrementally, as "mutually-exclusive" alternatives. However, interactions can also have profound effects on the relative costs and effects of diverse treatments, whether these are evaluated using factorial trials, two-arm RCTs or decision-analytical models.

This chapter aims to:

- Review the definitions used in the literature to distinguish between independent and mutually-exclusive alternatives (Section 9.2).

[76] See Section 3.5. It is generally stated that health benefits are maximised only if the last treatment adopted is divisible and has constant returns to scale. Although these assumptions are not realistic, this is unlikely to matter for two reasons. Firstly, within a complete league table of all healthcare interventions, the cost of each individual intervention will be a very small fraction of the total budget: particularly if the cost-effectiveness of individual treatments is stratified, with the ICER for treatment in different patient groups appearing separately in the league table. Secondly, healthcare systems typically have a soft budget, such that any money left over from one year can be spent in the next (183).

- Identify additional situations in which interventions may interact, outside the context of factorial trials (Section 9.3).
- Assess the impact of using the "wrong" decision rule: i.e. making separate decisions on interventions that actually have non-additive effects and treating different combinations of independent options as though they were mutually-exclusive (Section 9.4).
- Discuss how HTA organisations and decision-makers, such as NICE and by the World Health Organisation (WHO), currently deal with interactions (Section 9.5).
- Propose changes to the NICE technology appraisal process that could enable more efficient use of healthcare resources and discuss sources of evidence that could be used to guide assumptions about interactions within decision-analytical models (Section 9.6).

9.2 Definition of mutually-exclusive and independent alternatives

Although most researchers and teachers of health economics have a reasonable working appreciation of what treatments can be considered “independent” and which are “mutually-exclusive”, there is a substantial variation in the definitions used in the literature[77] (Table 9.1). Eleven textbooks or reviews defined "mutually-exclusive" literally, stating that treatments are mutually-exclusive if patients cannot receive both. However, there are relatively few treatments that are impossible to implement together (rare examples including different levels of the same treatment or different approaches to irreversible surgery). Seven definitions recommended that all treatments given to the same patient population are considered mutually-exclusive. However, 10 of the published definitions made specific reference to non-additive effects (i.e. interactions) as the defining feature that determines whether or not treatments are considered to be mutually-exclusive or independent.

[77] Definitions were identified by searching Google scholar (first 100 hits only) and Medline (through PubMed) on 5th November 2013 for terms including ‘”mutually-exclusive" cost-effective’, ‘independent option cost-effective’ and ‘independent alternative cost-effective’. Key health economic textbooks and the reference lists from identified papers and books were also reviewed. Fourteen references with definitions were identified.

Table 9.1. Published definitions of "mutually-exclusive" and "independent".

Publication	Definition of "mutually-exclusive"	Definition of "independent"	Type*
Briggs 2000 (188)	"Mutually exclusive programmes involve the same group of patients and, therefore, one or the other must be chosen. Patients can be screened yearly or biennially, but not both. Similarly, patients can receive either a 40 or 80mg regimen for a particular drug."	"Independent programmes can be implemented either singly or jointly, so in the case of the ICD [implantable cardioverter defibrillator], the device could be implanted into patients with a high ejection fraction or those with a low ejection fraction, or into both types of patient. Such a decision should be based on the incremental cost effectiveness of the ICD for these 2 types of patients compared with alternative (medical) management."	L P
Drummond 2005 (125)	"The various alternatives within each programme are assumed to be mutually-exclusive, in that if a patient receives one of the treatments in each programme, they will not receive the others" (pages 127-130)	"The treatment used in one patient group is assumed to be independent of the treatments used in other groups. That is, the costs and health effects of a treatment in one patient group are not affected by the treatment alternative chosen in any other patient group." (pages 127-30) "The costs and health effects in one patient group are not affected by the treatment alternative in any other patient group" (page 12)	L I P
Elbasha 2004 (299)	"Programs in a cluster are said to be locally mutually exclusive if an individual receives a service of one program, and he/she cannot obtain the service of another program from the same cluster [22]. Examples of locally mutually exclusive programs are cancer detection programs defined by the frequency of screening. An individual cannot be screened both annually and biennially. However, it is assumed that two individuals can receive the services of two different programs from the same cluster. For example, one individual is annually screened, whereas the other is screened every 2 years. That is, programs are not globally mutually exclusive."		L
Evans 2005 (190)	"If one is chosen, another cannot be." "Mutually exclusive interventions must replace an existing intervention" *However, the authors do recommend evaluating all combinations of interventions allowing for interactions*	"Can be done at the same time in a population, with or without interactions." "Can be added to existing interventions"	L I-
Gray 2011 (16)	"Clearly, it would be possible to implement both primary *and* secondary prevention; these are independent policies affecting different patient groups, and so the appropriate approach would not be to calculate the incremental cost-effectiveness of primary versus secondary prevention, but rather to calculate the incremental cost-effectiveness of statins in each relative to the next best alternative for that group of patients. The resulting ICERs could then be compared to see which offered the best value for money. Similarly, policies aimed at different age groups are independent, and a direct incremental calculation of the cost-effectiveness in one age group versus another would not be appropriate. However, different doses of statin are a different matter. Giving a 20 mg/day dose cannot be considered independently of giving the same person a 40 mg/day or 80 mg/day dose; these are mutually-exclusive alternatives." (page 15)		L P

Publication	Definition of "mutually-exclusive"	Definition of "independent"	Type*
Heshmat 2001 (187)	"…Two programs, A and B, are mutually-exclusive if implementing program A (B) means that program B (A) cannot also be implemented. Mutually-exclusive programmes can be viewed as programmes for the same population (e.g. two alternative drugs for ulcer patients)."	"Two programs, A and B, are defined as independent if the costs and effectiveness of programme A (B) are not affected by whether program B (A) is implemented or not. The two programmes are viewed as applying to two different populations; an example being the treatment of ulcer patients and the treatment of arthritis patients...."	L I P
Hunink 2001 (172)	"In many health applications of CEA, however, the choices involve two or more interventions that, by definition, cannot be selected at the same time – not just a yes/no choice for each alternative. The choice among four strategies for breast cancer screening, each involving a different screening interval, is such a choice among mutually-exclusive, or *competing choices*. Another is the choice between two different surgical procedures. This situation contrasts with the choice facing a public health agency of whether to implement an anti-smoking campaign and/or a skin cancer screening programme, since the decision to perform one service does not inherently preclude a decision to perform the other one […] The basic difference between these situations is that in competing choice situations such as that among breast cancer screening methods, the alternatives are not independent. In other words, the choice of the first alternative influences the benefit to be gained (or the cost incurred) by the second. In a noncompeting situation – such as the public health agency with a financial budget or the clinician with a time budget – the benefits of each programme can be added. In the competing choice situation, they cannot." (page 277) "If you select Program M1 you cannot select Program M2. For example, M1 and M2 might be two dosages of the same drug; if you chose to administer M1, which requires giving 5 mg of Drug A per day, you could not at the same time implement M2, 10 mg of Drug A." (page 280)		L I
Johannesson 1996 (170)	"Two programmes, A and B, are mutually-exclusive if implementing programme A (B) means that programme B (A) cannot also be implemented. If the programmes A and B can both be implemented physically, but implementing A (B) means that the costs and effectiveness of B (A) change, then the programmes should be defined as mutually-exclusive. In this case, we have three alternatives: carrying out only A, carrying out only B, or carrying out both A and B. Mutually-exclusive programmes can be viewed as programmes for the same population, e.g. two alternative drugs for ulcer patients." (page 135)	"Two programmes, A and B, are defined as independent if the costs and effectiveness of programme A (B) are not affected by whether programme B (A) is implemented or not. The two programmes are viewed as applying to two different populations; an example would be the treatment of ulcer patients and the treatment of arthritis patients…" (page 135)	L I P
Johannesson 1997 (128)	"Two programmes are mutually-exclusive if the costs and/or effects of one programme are affected by whether the other programme is implemented or not, e.g. two alternative drug therapies to lower blood pressure."	"Two programmes are independent if the costs and effects of a programme are not affected by whether the other programme is implemented or not, e.g. a treatment for ulcer and a treatment for cancer."	I
Karlson 1996 (169)	"The treatments that are available to one patient group are mutually-exclusive, which means that a patient can only receive one of the treatment alternatives"	"The treatment used in one patient group is assumed to be independent of the treatments used in other groups. That is the costs and health effects of a treatment in one patient group are not affected by the treatment alternative chosen in any other patient group"	L I P

Publication	Definition of "mutually-exclusive"	Definition of "independent"	Type*
Phillips 2009 (189)	"A distinction must be made between those interventions that are completely **independent –** that is, where the costs and effects of one intervention are not affected by the introduction or otherwise of other interventions – and those that are mutually-exclusive **–** that is, where implementing one intervention means that another cannot be implemented, or where the implementation of one intervention results in changes to the costs and effects of another."		I L
Weinstein 1996 (129)	"Many applications of cost-effectiveness analysis in health care involve comparisons among competing alternatives for the same condition. [...] In such competing choice situations, the basic cost effectiveness paradigm [for independent treatments] no longer applies without modification, because the alternatives (e.g. different drugs for the same condition) are no longer independent. As an obvious example, the benefits of giving two antihypertensive drugs to the same group of patients are not additive. The paradigm needs to be modified to incorporate the possibility of mutually-exclusive competing choices for the same condition."	"The costs and effectiveness of any programme are independent of which other programmes are adopted."	I P
WHO 2003 (185)	"Interventions are said to be mutually-exclusive if only one alternative can be selected. Mutually-exclusive interventions have also been called "competing" or "incompatible" interventions. […] Interventions are said to be independent if choosing one does not prevent the choice of any other intervention." (endnote 3) "Many interventions interact in terms of either costs or effects at the population level. The health impact of undertaking two interventions together is not necessarily additive, nor are the costs of the joint production. To understand whether they are efficient uses of resources independently or in combination requires assessing their costs and health effects independently and in combination. […] Interventions that interact should be evaluated as a group […] The case of mutually exclusive options is similar, i.e. interventions which by definition cannot be implemented simultaneously in the same population. An example is population-based annual and biannual breast cancer screening. These interventions must be evaluated as part of the same set which will ensure that only one of the interventions appears in an optimal mix." (page 21) The term 'mutually-exclusive' is used to describe the combinations of treatments, such as DOTS alone, BCG 50% coverage alone and DOTS + BCG 50% coverage (page 21) Discusses "Clusters of interventions that are interrelated either because they interact on either costs or effects or because they are mutually-exclusive" (page 25)		L I+

* Categories of interaction:
L: Literal definition – treatments are mutually-exclusive if patients cannot receive both.
P: Treatments for the same patient group are mutually-exclusive.
I: Treatments are mutually-exclusive if the costs or effects depend on the other treatment, although the word "interaction" is not mentioned.
I+: Interactions are mentioned as a defining feature of mutually-exclusive treatments.
I-: Definition states that independent treatments may interact.

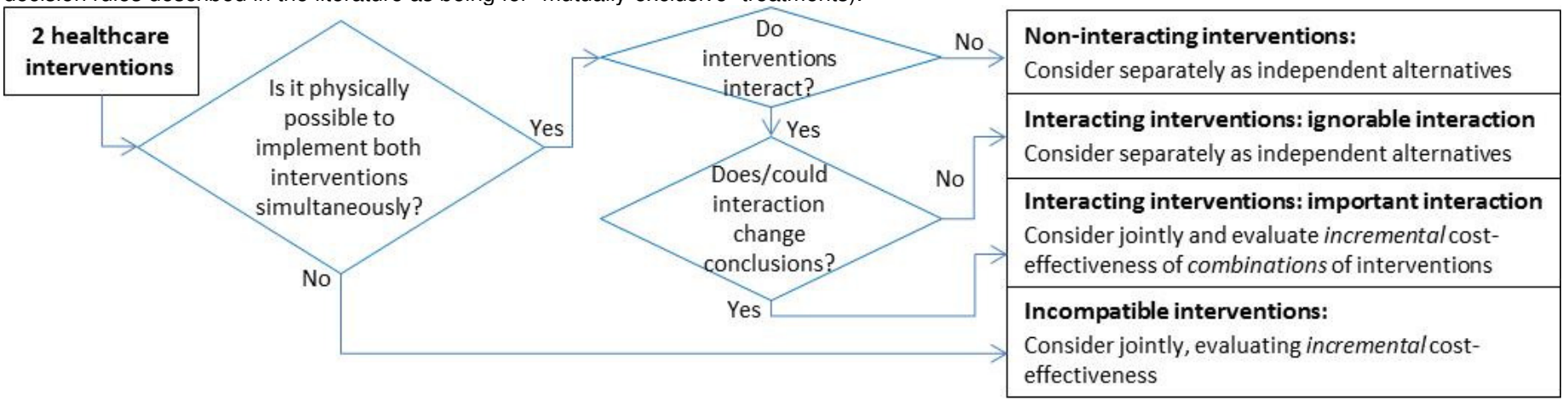
This suggests that whenever treatments interact we should treat the combinations of inter-related treatments (e.g. *0*, *a*, *b* and *ab*) as mutually-exclusive options, comparing the treatment combinations incrementally and selecting the single strategy that maximises NMB for that population (Table 3.7). Conversely, the definitions imply that if there is no interaction between treatments, we should treat A and B as independent treatment options and compare their ICERs against the ceiling ratio separately, adopting those treatments with ICERs below our ceiling ratio (Table 3.6). However, only the two WHO definitions (185, 190) mentioned the word “interaction” and these use slightly different terminology. They define mutually-exclusive treatments as those that cannot be implemented together, and distinguish them from technologies that can be given together but interact and from independent technologies that do not interact. They recommend that combinations of interacting technologies are considered as a “cluster” of inter-dependent interventions and evaluated incrementally *like* mutually-exclusive alternatives.

Within the literature to date, the terms “independent” and “mutually-exclusive” are used inconsistently, potentially causing confusion. I therefore propose using the terms *incompatible*, *interacting* and *non-interacting* (Figure 9.1) and avoidance of the term “mutually-exclusive”.

9.3 Situations when interactions are likely

As discussed in Section 3.4, interactions between treatments can arise due to pharmacological, behavioural or biological mechanisms (category 1 in the taxonomy shown in Table 3.3), through treatments having multiplicative effects on costs, QALYs or the incidence of clinical events (category 2), and through diminishing or increasing marginal returns on utility (category 3).

Figure 9.1. Flow diagram illustrating proposed terminology and decision rules. Given a decision problem concerning use of two or more interventions within a defined population or setting, it is first necessary to assess whether it is physically possible to implement both interventions at the same time. Depending on the decision problem, this question may be posed at the level of individual patients (e.g. is it possible to give both interventions to the same patient?), or at the level of healthcare systems (e.g. is it possible to implement both interventions within the same hospital/region?). The next step is to assess whether the interventions interact: i.e. whether use of one treatment affects the incremental costs and health benefits gained from the other. In practice, almost all interventions given to the same patient population and many of those used in the same healthcare system will interact to some degree, although it is unlikely to be good use of health economists' time or healthcare funds to conduct overly-complicated analyses informing joint decisions regarding numerous treatments with small interactions. It is therefore likely to be necessary to ignore interactions that are unlikely to change the conclusions. If interactions are non-existent or ignorable, separate decisions can be made, treating the options as independent. If the treatments are incompatible or have important interactions, it will be necessary to make a joint decision and evaluate incremental cost-effectiveness using the methods described in Table 3.7 (i.e. the decision rules described in the literature as being for “mutually-exclusive” treatments).



Additional mechanisms can introduce interactions outside the context of factorial trials, such as patient pathways (category 4). For example, the costs and benefits of screening will always depend on the prevalence of the condition and on what interventions are used to diagnose and treat the cases identified (87) (category 4a, example F in Appendix D). At the extreme, screening is of negligible value if there is no effective treatment (87), or if preventative treatments have eradicated the disease in this population. Similarly, the costs and benefits of preventative interventions will depend on the costs and effectiveness of subsequent diagnosis and treatment: e.g. prevention may have little value if there is a cheap, highly effective cure. The costs, benefits and range of options for second-line therapy may also be affected by what treatment was given first-line: particularly if drug resistance or toxicity develop during treatment.

Interactions between treatments can also arise from interactions between diseases (category 4b): e.g. if having heart failure changes the risk, case-fatality, cost or HRQoL associated with stroke (300), interventions reducing the risk of heart failure will also indirectly affect the cost-effectiveness of treatments to prevent and/or treat stroke, even if they do not directly influence the incidence of stroke in patients without comorbid heart failure (example L). The cost of delivering an intervention (e.g. knee replacement) may also be affected by how comorbid conditions are managed (category 4c, example M). Introducing new interventions increasing the cost of treatment for any common condition will also increase the cost accrued in the years of life gained from any life-saving treatment if future costs (301) are included in the analysis (category 4d, example J).

Although the above types of interaction arise between treatments given to the same patients, interactions are also possible between treatments given to different patient groups within the same healthcare system (category 5). For example, freeing up hospital beds through conducting a surgical procedure as a day-case may increase costs and, potentially, health benefits of other forms of surgery: e.g. if patients undergoing other procedures stay in

hospital longer once more beds are available, or if reducing bed-occupancy has a non-linear effect on the cost/bed-day (example O). Equipment purchased for one intervention can also be used for other patients, which may affect their costs and health gains (example N). Complex interventions and changes to how services are delivered or organised will also change the costs and benefits of numerous interventions: e.g. setting up or reorganising a primary care service to offer measles vaccination will change the feasibility, cost and outcomes of tuberculosis treatment that is then delivered in the same centre. Similarly, nurse-led measures to encourage women to check for breast cancer may increase the chance that the same nurses promote awareness and routine checks for other diseases. Conversely, adding an additional intervention into a healthcare package (e.g. primary care services) may affect compliance and/or waiting times and introduce interactions between the new service and existing ones delivered by the same healthcare professionals.

However, in practice interactions between treatments given to different patients are likely to be smaller and occur less commonly than interactions between treatments for the same patients and it may therefore be reasonable to exclude them. This trend explains the apparent contradiction between the definitions described in Section 9.2: since economically-important interactions are substantially more likely between treatments given to the same patients than treatments given to different patients, patient group is a useful rule of thumb to identify the situations in which treatments should be considered to have important interactions. Since interactions cannot be ruled out whenever two treatments are given to the same patients, there is an argument for considering all treatments given to the same patient group as interacting.

9.4 Implications of using the "wrong" decision rule

Since interactions appear to determine which decision-making rule is recommended and are likely to arise commonly, I evaluated the implications of using the "wrong" decision rule: i.e. making separate decisions about technologies that interact and making a joint decision between different combinations of independent technologies. This was explored using league tables of 14 hypothetical interventions, of which six pairs of treatments interacted through the mechanisms described in Sections 3.4 and 9.3 (Table 9.2). The data generation procedure is described in Appendix D.

9.4.1 Making independent decisions about each treatment

I first evaluated each intervention as though they were independent of all other treatments in the league table (Table 9.3). Estimating costs and effectiveness separately for interacting treatments highlights the first difficulty with this approach: namely that the incremental costs and benefits are sensitive to the proportion of patients who receive the concomitant treatment, which is not specified in the research question.[78] For simplicity, the costs and benefits of each treatment were estimated under the assumption that 50% of eligible patients received the concomitant treatment and 50% did not. For example, the incremental cost of statins vs. no statins (£1.96 billion) was estimated as the average of statins+antihypertensive (£3.42 billion) and statins alone (£2.42 billion) minus the average of antihypertensive alone (£1.42 billion) and no treatment (£500 million), assuming that 50% of patients received antihypertensives.

[78] In practice, the proportion of patients receiving the concomitant treatment could be anywhere between 0% and 100%, depending on what proportion of patients in the population of interest are indicated for the concomitant treatment and what proportion of eligible patients receive this treatment in practice.

Table 9.2. Total costs and QALYs for the interventions included in league tables. The data generation procedures are described in Appendix D.

	Cost/1,000 population	QALYs/1,000 population	Description
Patients at low risk of cardiovascular disease			
No statin and no antihypertensive (A0)	£500,000	7,833	Statins and antihypertensives are both assumed to decrease the odds of cardiovascular events, with no interaction between the two types of drug on a log-odds scale. As result, there is a sub-additive effect on the probability of cardiovascular events and on QALYs and a super-additive effect on costs.
Antihypertensive only (A1)	£1,416,429	7,893	
Statin only (A2)	£2,412,222	7,926	
Statin + antihypertensive (A3)	£3,422,405	7,954	
Interaction	£93,754	-31	We can eliminate the interaction by evaluating cost-effectiveness separately for patients at differing risk.
Older smokers with osteoarthritis			
No smoking cessation and no knee replacement (B0)	£120,000	144	Knee replacement doubles patients' utility for all remaining life-years, while smoking cessation (if successful) increases life expectancy by 50%. As result, there is a large super-additive interaction for QALYs, but additive effects on cost (since both interventions are assumed to be one-off). The actual interaction could be substantially larger than this if stopping smoking before knee replacement surgery reduced operative mortality and morbidity.
Smoking cessation only (B1)	£111,000	158	
Knee replacement only (B2)	£420,000	288	
Knee replacement + smoking cessation (B3) vs. B1	£411,000	317	
Interaction	£0	14.4	We can eliminate the interaction by stratifying by life expectancy.
Patients with anxiety			
Once-daily Drug A for anxiety (C0	£2,400	4.8	There are two drugs that can be used to treat anxiety (A and B), which can each be used once-daily or twice-daily. An interaction for cost arises because the cost of twice-daily treatment is always twice as high as the cost of once-daily treatment, regardless of drug costs, which means that doubling dosing frequency increases costs more for the more costly drug (B). It is essential to make a joint decision on dosing frequency and which drug to use.
Once-daily Drug B for anxiety (C1)	£2,960	4.9	
Twice-daily Drug A for anxiety (C2)	£4,800	5.6	
Twice-daily Drug B for anxiety (C3)	£5,920	54.7	
Interaction	£560	0	
Patients with pain only			
No treatment (D0)	£0	1.57	Either of the two painkillers will improve patients' responses level on the pain domain by 1 level. Given individually, they can improve pain from extreme to moderate, but given together they can eliminate pain altogether. As result, there is a sub-additive interaction for QALYs, since the coefficients for level 3 pain are more than twice as high as those for level 2 pain, in addition to level 3 pain attracting the N3 term. Effects on costs are additive. A joint decision is needed to allow for the interaction.
Painkiller 1 only (D1)	£18,000	6.89	
Painkiller 2 only (D2)	£48,000	6.89	
Painkillers 1 & 2 (D3)	£66,000	8.12	
Interaction	£0	-4.09	
Patients with comorbid pain and depression			
No treatment (E0)	£0	-0.05	The painkiller reduces patients' pain from extreme to moderate (i.e. from level 3 to level 2 on EQ-5D), while the antidepressant has the same effect on the anxiety/depression
Antidepressant only (E1)	£5,000	1.05	

	Cost/1,000 population	QALYs/1,000 population	Description
Painkiller only (E2)	£18,000	1.70	domain. As result, there is a large super-additive interaction for utilities and QALYs due to the EQ-5D tariff, which subtracts 0.269 from utility for any patients with level 3 on any domain: both anxiety/depression and pain must be improved to above level 3 to remove the N3 term. A joint decision is needed to allow for the interaction. Similar effects would also be seen for other domains and for combinations of treatments that improve patients to perfect health. Even larger interactions would be seen if improved pain relief also improved anxiety/depression or vice versa.
Painkiller + antidepressant (E3)	£23,000	4.59	
Interaction	£0	1.79	
Women at risk of cervical cancer			
No cervical screening and old cervical cancer treatment (F0)	£3,700,000	789,520	Treating the patients diagnosed with cervical cancer with the new treatment rather than the comparator treatment increases costs and improves quality-adjusted survival for patients at all cancer stages and therefore increases the cost and reduces the benefits of screening. Conversely, screening patients for cervical cancer will increase the chance that cancers are detected at an early stage; total costs and QALYs vary depending on the stage of cancer at which patients are diagnosed. Additionally, the incremental costs and incremental QALYs associated with the new treatment compared with its comparator also vary with stage. As result of this, if we consider cervical cancer as a whole (rather than making separate decisions for different stages), we get a sub-additive interaction for costs and a super-additive interaction for QALYs. However, providing that screening doesn't affect the costs and QALYs of patients within each cancer stage, the cost-effectiveness of treating cervical cancer will be independent of screening if we make separate decisions about treatment for each cancer stage. Once we have made the decision about the best cancer treatment, we can then make a decision on the cost-effectiveness of screening that is contingent on patients receiving the most cost-effective treatment strategy.
Cervical screening and old cervical cancer treatment (F1)	£4,600,000	791,024	
No cervical screening and new cervical cancer treatment (F2)	£5,550,000	792,280	
Cervical screening and new cervical cancer treatment (F3)	£5,900,000	794,536	
Interaction	-£550,000	752	
Patients with ovarian cancer and patients with benign prostatic hypertrophy (BPH)			
No treatment for either ovarian cancer or BPH (G0)	£29,000	328	Since these interventions are given to different, non-overlapping patient populations with different diseases managed by different staff in different healthcare facilities, they are completely independent and have additive effects on costs and QALYs It is therefore reasonable to make independent decisions on the two treatments
Ovarian cancer treatment only (G1)	£41,500	378	
BPH treatment only (G2)	£145,000	332	
Ovarian cancer treatment + BPH treatment (G3)	£157,500	382	
Interaction	£0	0	

Given a £3.5 billion budget and making independent decisions about each of the 14 interventions using the league table shown in Table 9.3 would lead us to adopt the 11 treatments with lowest ICERs and partially-adopt antihypertensive treatment,[79] gaining 5.254 million QALYs compared with adopting no treatments.

9.4.2 Treating all treatments as interacting

This resource allocation was contrasted with an analysis allowing for interactions between the same seven pairs of treatments (Table 9.4). Although Table 9.4 was generated using the same data, incremental costs, QALYs and cost-effectiveness differ from those in Table 9.3 since each combination of treatments is considered separately. For example in cluster A, Table 9.3 shows the incremental costs and QALYs for statin vs. no statin averaged across patients with/without antihypertensives and for antihypertensive vs. no antihypertensive averaged across patients with/without statin. Conversely, Table 9.4 shows the incremental cost and QALYs for antihypertensive (without statin) vs. no treatment, statin (without antihypertensive) vs. antihypertensive and statin+antihypertensive vs. statin.[80]

For illustrative purposes, one of the clusters comprised two treatments that can safely be assumed to be completely independent: one treatment for ovarian cancer and another for benign prostatic hypertrophy (BPH). Since there is no interaction between these treatments, the cost, effectiveness and cost-effectiveness of ovarian cancer treatment is the same regardless of what intervention is given for BPH (and vice versa, Tables 9.3 and 9.4).

[79] When using the league tables, I assume that the last treatment adopted (in this case antihypertensive) is divisible and has constant returns to scale.
[80] For example, the incremental cost of statin only vs. antihypertensive only (£996 million) equals the cost of statin only (£2.42 billion; Table 9.2, Appendix D) minus the cost of antihypertensive only (£1.42 billion).

Table 9.3. League table assuming that all treatments are independent. Treatments that would be adopted are shown in green, treatments that would be partially-adopted in black and treatments not adopted in red. Adopting A1 in 6% of patients as well as the treatments in green gains 5,254 QALYs.

Treatment	Total Δ cost '000s	Total Δ QALYs '000s	ICER	Cumulative cost '000s	Cumulative QALYs '000s
Smoking cessation (B1)	-£9,000	21.60	Dominant	-£9,000	22
Ovarian cancer treatment (G1)	£12,500	50.00	£250	£3,500	72
Cervical cancer screening (F1)	£625,000	1,880	£332	£628,500	1,952
Cervical cancer treatment (F2)	£1,575,000	3,136	£502	£2,203,500	5,088
Knee replacement (B2)	£300,000	151.20	£1,984	£2,503,500	5,239
Antidepressant (E1): Patients with comorbid pain and depression	£5,000	2.00	£2,504	£2,508,500	5,241
Twice-daily treatment for anxiety vs. once-daily (C2)	£2,680	0.80	£3,350	£2,511,180	5,242
Painkiller 1 (D1): Patients with pain only	£18,000	3.28	£5,496	£2,529,180	5,245
Painkiller (E2): Patients with comorbid pain and depression	£18,000	2.65	£6,792	£2,547,180	5,248
Drug B for anxiety vs. Drug A (C1)	£840	0.08	£10,500	£2,548,020	5,248
Painkiller 2 (D2): Patients with pain only	£48,000	3.28	£14,656	£2,596,020	5,251
Antihypertensive (A1)	£963,305	43.90	£21,944	£3,559,325	5,295
Statin (A2)	£1,959,099	76.97	£25,454	£5,518,425	5,372
BPH treatment (G2)	£116,000	4.00	£29,000	£5,634,425	5,376

Table 9.4. League table allowing for interactions between clusters of potentially-interacting treatments. Treatments that would be adopted are shown in green, treatments that would be partially-adopted in black and treatments not adopted in red. Dominated options (B2, C2, D2, E2, F2, G1, G2) are omitted. Adopting G3 in 70% of patients as well as the treatments in green gains 5,313 QALYs.

Treatment	Total Δ cost '000s	Total Δ QALYs '000s	ICER	Cumulative cost '000s	Cumulative QALYs '000s
Smoking cessation only (B1) vs. B0	-£9,000	14.40	Dominant	-£9,000	14
Ovarian cancer treatment only (G1) vs. G0	£12,500	50.00	£250	£3,500	64
Cervical screening and new cervical cancer treatment (F3) vs. F0	£2,200,000	5,016	£439	£2,203,500	5,080
Knee replacement + smoking cessation (B3) vs. B1	£300,000	158.40	£1,894	£2,503,500	5,239
Twice-daily Drug A for anxiety (C2) vs. C0	£1,840	0.72	£3,000	£2,505,340	5,240
Painkiller 1 only (D1) vs. D0: Patients with pain only	£18,000	5.32	£3,383	£2,523,340	5,245
Antidepressant only (E1) vs. E0: Patients with comorbid pain and depression	£5,000	1.10	£4,545	£2,528,340	5,246
Painkiller + antidepressant (E3) vs. E1: Patients with comorbid pain and depression	£18,000	3.55	£5,075	£2,546,340	5,249
Twice-daily Drug B for anxiety (C3) vs. C2	£2,960	0.80	£14,000	£2,549,300	5,250
Antihypertensive only (A1) vs. A0	£916,429	59.52	£15,396	£3,465,729	5,310
Ovarian cancer treatment + BPH treatment (G3) vs. G1	£116,000	4.00	£29,000	£3,581,729	5,314
Statin only (A2) vs. A1	£995,794	33.07	£30,113	£4,577,522	5,347
Statin + antihypertensive (A3) vs. A2	£1,010,182	28.27	£35,730	£5,587,705	5,375
Painkillers 1 & 2 (D3) vs. D1: Patients with pain only	£48,000	1.23	£39,024	£5,635,705	5,376

Given the same £3.5 billion budget, a league table allowing for interactions between treatments would lead us to adopt the 10 treatment-combinations with lowest ICERs and partially adopt G3, gaining a total of 5.313 million QALYs compared with adopting no treatments: 58,086 (1.1%) more than we would gain by ignoring interactions. This QALY gain is realised by adopting treatment combinations with lower ICERs: e.g. using Table 9.4 we fully adopt antihypertensive treatment and partially implement BPH treatment, but do not adopt painkiller 2.

I took the approach of considering combinations of independent options as though they interacted to the extreme by considering all 16,384 (2^{14}) possible combinations of the 14 interventions as mutually-exclusive alternatives and adopting the combination with highest NMB at a £20,000/QALY ceiling ratio. This approach gave identical treatment adoption decisions to the approach used in Table 9.4, but greatly complicated the analysis (not shown).

Making independent decisions on interventions that interact can also give misleading conclusions when decisions are made using a ceiling ratio representing the shadow price of a QALY, rather than a league table. For example, given a £20,000/QALY ceiling ratio, if we evaluated cluster D as interacting, we would adopt drug 1 (D1, costing £3,383/QALY vs. D0), in preference to drug 2 (D2, strongly dominated by D1) or combination therapy (D3, costing £39,024/QALY vs. D1); by contrast if the two drugs are evaluated independently, both would be adopted, since both cost <£20,000/QALY vs. no treatment.

However, the two approaches give identical conclusions for interventions that have additive effects on costs or QALYs, with no interaction (e.g. cluster E). The approaches give different adoption decisions only when there are *qualitative* interactions for NMB at the ceiling ratio of interest: i.e. when the INB for A changes sign depending on whether B is also given (Section 3.7.3, Appendix E). This means that the conclusions for interventions that are unambiguously

above or below our ceiling ratio (e.g. clusters B, C, F and G) are unaffected (since they do not have a qualitative interaction on NMB), whereas those for clusters A and D are.

More generally, taking interactions between treatments into account and making a joint decision on all interventions under consideration always enables us to maximise the health gains from our budget (Appendix E). By contrast, ignoring interactions and making independent decisions about interventions that interact can lead to suboptimal resource allocation decisions that do not maximise health. Furthermore, interactions are the reason why treating mutually-exclusive options as independent gives misleading conclusions, as the costs and benefits of Treatment A are assumed to be unaffected by whether or not we also give Treatment B, implicitly assuming zero interaction between A and B.

However, in practice quantifying costs and benefits of all possible combinations of all treatments and quantifying the extent to which costs and effects are affected by all possible concomitant treatments would require additional data or assumptions and make economic evaluation unfeasibly complex and expensive. As such, it may be appropriate to assume that treatments are independent in certain circumstances.

9.5 Current decision-making methods

9.5.1 Current practice: developed countries

At present, HTA organisations such as NICE, Scottish Medicines Consortium (SMC), Pharmaceutical Benefits Advisory Committee (PBAC) and the Common Drug Review (CDR) currently make separate resource allocation decisions on different interventions used to treat the same conditions. For example, in order to expedite access to new treatments, NICE increasingly conducts single technology appraisals (STAs) evaluating one drug at a time, in preference to multiple technology appraisals (MTAs) evaluating a whole treatment class

simultaneously. For example, NICE made separate recommendations on cervical cancer screening (TA69) and cervical cancer treatment (TA183), which did not explicitly discuss interactions between treatments. The decision about whether to allow for interactions between interventions is left up to individual analysts, with no explicit guidance given in the "scope" document, in which NICE defines the decision problem.

Making separate decisions on treatments for the same condition has several limitations. Firstly, considering treatments one at a time means that the study question does not explicitly address what, if any, concomitant interventions are given alongside treatment and means that interactions between interventions are likely to be ignored, leading to suboptimal decision-making (Section 9.4).

Secondly, even in situations where combination therapy is inappropriate, evaluating drugs one-at-a-time means that Treatment A is not always considered as a comparator when B is being evaluated, which may mean that dominance is overlooked and that ICERs are calculated relative to a comparator that does not lie on the cost-effectiveness frontier. For example, NICE evaluated tenofovir and entecavir in separate STAs and recommended both treatments for hepatitis B (219, 220), despite entecavir being substantially more costly, with no evidence of higher efficacy (221).

Thirdly, recommendations and guidelines based on separate decisions frequently do not explicitly state whether treatments are recommended alone or in combination with other treatments. For example, NICE commonly recommends treatments "as an option for" the condition in question, without explicitly discussing what if any concomitant treatment is assumed or recommended. Recommendations "as an option" appear to be around twice as common for decisions within STAs than those in MTAs.[81]

[81] Based on an analysis of HTAinSite decisions published before 31st December 2011. Data were used with permission from HTAinSite (https://www.htainsite.com).

However, making separate decisions on different treatments will simplify decision-making and enable decisions to be made relatively quickly by different teams. Furthermore, NICE's preference for recommending several options for the same condition may reflect a reluctance to recommend one treatment over another when there is uncertainty around point estimates and no head-to-head evidence demonstrating statistically-significant superiority. Recommending several treatments also enables patients and clinicians to choose the most appropriate treatment based on clinical judgement and the patient's needs, decreases the number of politically-damaging "no" decisions and avoids conflict with the pharmaceutical industry.

9.5.2 Current practice: WHO

By contrast, the WHO framework for generalised cost-effectiveness analysis (GCEA) models clusters of inter-related treatments simultaneously, and explicitly allows for interactions between interventions (183, 185). WHO intends GCEA to provide estimates of relative costs and effectiveness "that are meant to contribute through multiple channels to a more informed debate on resource allocation priorities" (183, 185) and which are broadly applicable across a wide range of countries. Different combinations of interventions are compared against a "null" (comprising costs and health benefits if the entire cluster of interventions was withdrawn), to assess the cost-effectiveness of current interventions and ensure that results can be generalised to other populations. The WHO population model (PopMod) also directly allows for different types of interactions between diseases (300).

The WHO-CHOICE programme has now evaluated 15 large clusters of interventions for up to 14 different regions (http://www.who.int/choice/results/en). However, the evidence on interactions is frequently weak and relies upon assumptions of multiplicative effects (185,

302). Indeed, WHO explicitly aim to simply identify the “order of magnitude differences in cost-effectiveness of different interventions” and categorise interventions into “good buys” and “bad buys” to guide decision-making processes that also consider equity considerations and other constraints.

9.6 A proposed alternative

I have shown that it is always correct to evaluate the incremental cost-effectiveness of combinations of treatments jointly (even if there is no interaction). However, making separate decisions on interacting treatments and failing to consider interactions will fail to maximise health gains from the budget. In practice, small interactions are likely to arise between nearly all treatments given to the same patients and many of those used in different patients in the same healthcare system, just as interactions in factorial trials are never exactly equal to zero. One option for ensuring efficient resource allocation would therefore be to consider *all* combinations of all possible interventions incrementally in one enormous joint decision problem, although this is unlikely to be feasible in practice, or represent good use of research funding or time. In particular, many interactions are highly unlikely to change treatment adoption decisions: either because the interaction is very small, or because one or more treatment will be cost-effective (or will never be cost-effective) no matter how large the interaction is. Instead, the flow chart in Figure 9.1 could be followed to allow for those interactions likely to change the conclusions of the analysis and ignore small, unimportant interactions. Several changes could be made to the current NICE technology appraisal process to allow for potentially-influential interactions in decision-making.

The single most important change is to consider the likelihood, type and magnitude of interactions between interventions at all stages in the appraisal process. In particular, considering interactions at the pre-scoping or horizon-scanning stage will enable

identification of interventions that are likely to have no major interactions and can be evaluated separately. While all interventions given to overlapping populations are likely to interact to some extent (e.g. through impacts on mortality), such interactions may frequently be small and unlikely to change the conclusions of the analysis: particularly if interventions have only a small impact on total life expectancy or if the overlap between populations is small. It may therefore be appropriate to make a pragmatic decision about whether or not the added complexity of making a joint decision is justified given the strength of interactions (185). Some decisions may also interact in the long-term, but be independent in the short-term. For example, the cost-effectiveness of screening for cervical cancer will depend on the availability and coverage of human papillomavirus vaccination, which will reduce the incidence, while the cost-effectiveness of vaccination will depend on the screening conducted in the future. However given the delay between vaccination at age 12-13 and screening from age 25, it is reasonable to make separate decisions about vaccination and screening today, providing that these are reviewed regularly to allow for changes in the incidence and treatment patterns.

Considering interactions at the pre-scoping or horizon-scanning stage can also enable technologies to be evaluated as MTAs rather than STAs to facilitate simultaneous decisions on multiple treatments, or to be evaluated as clinical guidelines that consider all treatments and care pathways for the condition as a whole. Technologies that are specifically dependent on certain other interventions can also be identified at this stage, which may affect the sequencing of appraisals. For example, the options for cervical cancer treatment could be appraised first and the appraisal on screening strategies could be started after the treatment appraisal was completed; the cost-effectiveness of screening could then be evaluated conditional on having already adopted the most cost-effective strategy for treating the cancers detected (Appendix D, example F). It may also be appropriate to schedule appraisals evaluating first-line treatment after appraisals evaluating second or subsequent-

line treatment, so that the pathway of subsequent treatment can be optimised before first-line treatment is evaluated.[82]

Identifying likely interactions at the scoping stage can also ensure that manufacturers and academic groups are instructed to take these into account in the analysis and address the most appropriate research questions. In particular, many interactions between treatments can be eliminated by allowing for heterogeneity and making stratified decisions specific to patients in different risk groups or stages. For example, the cost-effectiveness of statins will depend on whether or not aspirin and antihypertensives are also administered, since these will all affect patients' cardiovascular risk. However, if we make stratified decisions for patients at different risk levels, the cost-effectiveness of statins *at any given level of risk* may be independent of other preventative treatments. Stratified decisions can also eliminate other types of interaction: e.g. decisions could be stratified by life expectancy or all-cause mortality. Stratifying decisions by stage of cancer can also ensure that the decision about treatment of cancer is conditionally independent of screening. Screening may increase the chance that cancers are detected at an early stage, thereby affecting the costs and benefits of treatment when averaged over all stages. However, if costs, HRQoL and mortality *within* any given stage are independent of screening, we can make stratified decisions about the appropriate treatment for any given stage that are then independent of what screening is offered.

Even after ignoring unimportant interactions, sequencing decisions logically and making stratified decisions, some clusters of treatments will still need to be evaluated jointly. This group is likely to include treatments that have non-additive effects on HRQoL (e.g. examples

[82] Using this strategy, we can make optimal decisions by first comparing all possible last-line therapies in a range of different subgroups who have already failed to respond to/tolerate different sets of treatments previously. Once the last-line treatment is optimised for each group, we can evaluate the penultimate-line treatment, making the assumption that everybody has the optimal last-line treatment (i.e. assuming technical efficiency). Each line of treatment can be evaluated in reverse order, eventually enabling us to assess first-line treatment conditional on patients subsequently receiving the optimal sequence of treatments in their subsequent care. This approach simplifies the decision about first-line treatment compared with the alternative: namely modelling all possible sequences of drugs to identify the best treatment pathway overall, which can require comparisons between hundreds of different treatment strategies (e.g. (221)).

C and D) and interactions between prevention and screening/treatment when there is little delay before onset of disease (e.g. vaccination and prophylactic treatment for influenza). In some such cases, decisions could be stratified by whether or not concomitant treatment is received, although it is likely to be more informative to make a joint decision on both treatments simultaneously, since both are amenable to decision-maker control. Parameters within such joint models should also be linked (such that the efficacy of *ab* equals the efficacy of A, the efficacy of B and an interaction term), so that the efficacy of all treatments varies simultaneously in probabilistic and deterministic sensitivity analyses.

Interactions will also affect decisions about subsequent research. For example, if treatments are independent and have additive effects on costs and QALYs, we could assess the expected value of conducting separate trials on each intervention. By contrast, if treatments interact, EVSI calculations may need to reflect the costs and outcomes of a factorial trial collecting information on A, B and the interaction. Analytical methods to estimate the EVSI from factorial trials have not yet been developed, although numerical Bayesian methods are available (118).

9.6.1 Sources of evidence on interactions

Preliminary assessments of the importance of interactions and modelling work to estimate incremental costs and effects both require evidence on the magnitude and direction of interactions. In some cases, the evidence on interactions may be weak: particularly for rare side-effects. As Section 9.4 illustrates, treatment adoption decisions will often be sensitive to the assumptions about interactions. In principle, making an extremely inaccurate assumption about the magnitude or direction of interactions could give more misleading conclusions than ignoring interactions altogether. However, this is not a good reason to make separate decisions on interacting treatments, since making a joint decision between mutually-

exclusive combinations forces us to explicitly consider the direction and magnitude of interactions. It also prompts collection of evidence on interactions and consideration of the uncertainty around the interaction rather than simply assuming that the interaction is exactly equal to zero.

Adequately-powered factorial RCTs would provide the best evidence on interactions (20). However, most factorial trials are not powered for interactions (8, 20), so cannot conclusively demonstrate whether an observed interaction arose by chance. Factorial trials with economic evaluations can directly assess the magnitude of interactions for costs, QALYs and NMB; these can also be used to evaluate the cost-effectiveness of both treatments simultaneously, taking into account any observed interaction (Chapter 3). Nonetheless, even factorial trials lacking prospective collection of resource use or HRQoL data can inform the interactions used in model-based economic evaluations. If a factorial trial observes a statistically significant interaction, it is clearly appropriate to treat these options as interacting and include the interaction term estimated in that trial as a model parameter and model its effect on costs and effects. However, it may be appropriate to include even non-significant interactions within models and make a joint decision about both treatments evaluated in the factorial trial: particularly as interactions may be larger for costs and QALYs than for clinical endpoints. Factorial trials can also highlight situations where treatments have multiplicative, rather than additive effects.

Information on interactions can also be provided by network meta-analysis (NMA), which synthesises evidence on an entire network of treatments and uses head-to-head RCTs and indirect comparisons to estimate the relative efficacy of each treatment compared with all alternatives (303-305). NMA is widely used (306), extensively researched (304, 305) and accepted by NICE (145), although no guidance relating to factorial trials has been issued. Any available factorial trials would be particularly useful for NMA, since they directly inform interactions and directly compare ≥4 alternatives. NMAs could incorporate evidence from

factorial trials using several alternative models that “lump” (303, 307, 308) interventions together differently. Firstly, the cells in the factorial design could be treated as separate interventions, with the differences between each arm and no treatment being estimated independently (while allowing for correlations between arms (304, 309)). Secondly, the effect of each factor could be modelled additively by implicitly assuming that the effect of combination therapy is the sum of the effects of the individual factors. Thirdly, we could explicitly estimate the interaction by modelling outcomes for combination therapy as a function of the main effects for each factor and an interaction term; given that NMA is normally conducted in a Bayesian framework using MCMC (305, 306, 310), we could also attach informative priors to the interaction term. Even for networks without factorial trials, NMA could estimate the same three models based on a network of two-arm trials that between them present evidence on *0*, *a*, *b* and *ab*. However, in practice treatment effects within NMA often have large SEs, making it unclear whether an interaction is real or occurred by chance.

In principle, subgroup analyses stratifying patients by concomitant treatment could be used to estimate interactions, although these will not give unbiased estimates of the efficacy of any treatment to which patients were not randomly assigned. For example, a trial randomising patients to receive A or placebo that stratified patients into those who received concomitant B and those who did not, will give unbiased estimates of the efficacy of A vs. not-A with/without B and could be used to inform decisions about which patients should receive A. However, this study would not provide an unbiased estimate of the cost-effectiveness of B, or inform a joint decision about whether patients should receive *a*, *b* or *ab*, since selection bias could confound any differences between the groups with and without B (Section 10.3).

There may also be evidence suggesting that treatment has a multiplicative effect on outcomes, such as reducing the rate, odds or probability of clinical events by a certain percentage. Ideally, such evidence would come from factorial trials, or (at a minimum)

subgroup analyses or epidemiological studies. Even in the absence of such studies it may be reasonable to assume, as is commonly conducted in WHO-CHOICE studies (185, 302), that treatments affecting the chance of subsequent events have approximately multiplicative effects. However, results will differ depending on whether treatment is assumed to have a multiplicative effect on the hazard, a multiplicative effect on the odds, or a multiplicative effect on the probability, demonstrating the importance of identifying the appropriate scale.[83] Nonetheless, assuming multiplicative effects on any measure of the likelihood of clinical events may be more realistic than assuming additive effects on costs and QALYs when treatments affect the incidence of any common event, such as mortality, uptake of screening or stopping smoking.

Interactions that arise from the clinical pathway can also be built into the model structure based on expert opinion or guidelines. For example, models assessing the cost-effectiveness of screening can allow for different downstream treatment options, while those on first-line treatment can allow for second-line (and subsequent) treatments given after treatment failure. Expert opinion and/or pilot studies could be used to evaluate the impact of service reorganisation or changes to delivery on the costs and benefits of the interventions delivered through that service.

9.7 Conclusions

In conclusion, interactions determine whether it is appropriate to make separate decisions on different treatments. Making a joint decision on several treatments with appropriate assumptions about interactions will always maximise health gains from the budget, whereas making independent decisions on interacting technologies can (but won't always) lead to

[83] For example, if the odds of death are 0.05 without treatment and A halves the odds of death and B reduces the odds of death by 30%, there will be no interaction for odds on a logarithmic scale (interaction =Ln[0.05]-Ln[0.05*0.5]-Ln[0.05*0.3]+Ln[0.05*0.5*0.3]). However, if we convert these odds into probabilities (where probability = odds/[1+odds]) or rates (where rate =-Ln[1-probability]), there will be an interaction on a log-probability or log-rate scale. See Example Q in Appendix D for more details.

inefficient resource allocation decisions. HTA organisations, such as NICE, could improve decision-making by taking account of interactions when selecting, prioritising and appraising healthcare technologies. In particular, evaluating several technologies simultaneously within a guideline or MTA will encourage analysts and decision-makers to explicitly consider interactions between interventions. Similar observations have been made in other fields where considering concurrent decisions separately leads to irrational decision-making (311).

Chapter 10: Conclusions and recommendations

10.1 Summary of novel contributions from the thesis

This thesis represents the first study exploring and evaluating methods for conducting economic evaluations based on factorial RCTs. Chapter 2 reviewed the literature on clinical analyses of factorial trials, introducing principles and terms such as “inside-the-table” and “at-the-margins” analysis and “super-additive”, “sub-additive” and “qualitative” interactions into health economics.

In Chapter 3, I outlined the challenges facing researchers conducting economic evaluations based on factorial trials and proposed ways to overcome these. In particular, I identified the importance of interactions to the decision rules that are used to make treatment adoption decisions based on economic evaluations. I also demonstrated that the results of factorial trials must be interpreted inside-the-table on a natural scale to address the most relevant study question, evaluate incremental cost-effectiveness appropriately and ensure efficient allocation of healthcare resources. I also compared different regression methods for analysing factorial trials, including GLM, seemingly-unrelated regression, bootstrapping and MCMC.

I also proposed a taxonomy of mechanisms by which treatments may have non-additive effects on costs, QALYs and NMB (Sections 3.4 and 9.3) and predicted that large and qualitative interactions arise frequently for costs, QALYs and NMB. My empirical work confirmed this prediction, with 56% (9/16) of published factorial trials and all three of my case studies observing ≥1 interaction larger than simple effects. However, my review of published studies suggested that large interactions were also common for health effects (Chapter 4). My literature review also suggested that factorial trial-based economic evaluations are relatively rare but increasing in number, and that many studies ignore interactions.

Simulation work evaluating different criteria for selecting which interactions are taken into account within economic evaluations of factorial trials supported my assertion that for economic analyses informing treatment adoption decisions, it is better to include interactions that arose by chance than to risk ignoring genuine interactions that could change the conclusions (Chapter 5), even if this complicates interpretation.

I also applied alternative methods to two full factorial trials (Chapters 6 and 8). Within AFORRD, I compared different methods for analysing within-trial and extrapolated results of a 2x2x2 factorial trial, developing two techniques for combining parameter uncertainty and sampling uncertainty within the UKPDS Outcomes Model. For IVAN, I developed and implemented bootstrap-based methods enabling different components of NMB to be analysed separately, while propagating uncertainty and allowing for correlations between different components. This enabled interactions to be included within the analysis of some components but not others, thereby balancing the risk of bias and inefficiency. Implementing this approach using bootstrapping gave similar results to seemingly-unrelated regression (Section 3.6.4), but is more flexible and naturally allows for heteroskedasticity and non-Gaussian distributions (providing that the observed sample adequately represents the tail of the distribution).

As I proposed in Chapter 3, our treatment adoption decisions will be sensitive to how we deal with interactions whenever there is a qualitative interaction for NMB at the ceiling ratio of interest that changes which treatment has highest NMB. This arose in AFORRD, one published trial and in the KAT metal-backed comparison. Allowing for all interactions would also have changed the conclusions in IVAN at a £30,000/QALY ceiling ratio. Although allowing for interactions avoids bias, including additional interaction terms increased SEs. Generally this increased EVPI and shifted the probability that treatment is cost-effective towards even-odds, although there were notable exceptions (e.g. the KAT mobile bearing

comparison), where EVPI fell despite increased SEs, as the difference in expected NMB between the best and second-best treatments also increased substantially.

My thesis also represents the first methodological work on methods for analysing any endpoints (clinical or economic) from partial factorial trials (Chapter 7). The four analytical methods proposed (at-the-margins analysis, inside-the-table analysis on patients randomised to >1 comparison, subgroup analyses analysing patients as-treated and Bayesian analyses with informative priors) could all be applied to clinical or economic endpoints, although the subgroup approach is not recommended. Chapters 6 and 7 also presented methods for conducting multiple imputation of full and partial factorial trials.

I also identified the existence of mixed interactions, which arise when treatments have opposing effects (Section 2.5.2) and proposed the interaction:effect ratio as a measure of the magnitude of interactions (Section 4.2). Both of these can be used in any analysis of factorial trials (not just economic evaluations); the interaction:effect ratio may be particularly useful for comparing the magnitude of interactions between studies or between analyses conducted on different scales (e.g. to assess whether interactions are smaller on a logarithmic scale). The interaction:effect ratio could also be adapted for subgroup analyses of non-factorial trials by simply dividing the interaction by the effect of treatment in the total population. I also proposed methods to calculate the opportunity cost of ignoring interactions under current and perfect information (Section 3.8) and applied this to simulated data and the three case studies.

In Chapter 9, I also explored the implications of between-treatment interactions beyond factorial trial-based economic evaluation and made policy recommendations. This chapter also highlighted the lack of clear definitions for several terms commonly-used in economic teaching and guidelines. In particular, the term “mutually-exclusive” is confusing and

potentially misleading. To help promote best practice, it may be better to avoid this term and instead refer to treatments as incompatible or interacting.

Within the next section, I make recommendations for the design, analysis and reporting of factorial trial-based economic evaluations. I then close the thesis by discussing the usefulness of factorial trials as a vehicle for economic evaluation, limitations of the thesis and future research priorities.

10.2 Recommendations for economic evaluation of factorial trials

The thesis highlighted 17 aspects of the design, analysis and reporting of factorial trial-based economic evaluations that would improve future studies.

10.2.1 Design

- The aims and methods of economic evaluation are fundamentally different from analysis of clinical trial endpoints (Section 3.2.2) and it is generally not appropriate to replicate the analytical methods used for the primary clinical endpoint in the economic evaluation. In particular, large super-additive or qualitative interactions are particularly likely to arise for costs and QALYs and will inevitably occur for NMB at some ceiling ratio when point estimates lie in the north-east or south-west quadrants (Section 3.4). Furthermore, the conclusions of economic evaluations must be drawn on a natural scale, unlike clinical endpoints, which are frequently interpreted on a logarithmic scale to eliminate interactions (Section 3.6.1). However, it may be beneficial for statisticians and health economists to discuss analytical methods before the statistical analysis plan is finalised to help minimise any unnecessary differences in approach.

- The likelihood of interactions in each component of NMB (e.g. in QALYs, intervention costs, hospitalisations, etc.) should be considered at the start of the study. The taxonomy of interactions described in Section 3.4 may be a useful framework for identifying the potential sources of interactions between any set of treatments. Although it may not be appropriate to exclude interactions solely on the basis of prior beliefs, a priori considerations may identify components of NMB that need to be analysed separately (e.g. those likely to have large interactions and those that may be additive on a non-natural scale). Prior evidence or beliefs may also influence the choice of criterion that determines which interactions will be taken into account, or be quantified as informative priors within a Bayesian analysis.
- When designing factorial trials, care should be taken to ensure that the treatment given to the *ab* group is equal to the sum of the treatments given to the *a* and *b* groups (Chapter 4). Giving an additional intervention (e.g. advice or training) to the control group or to the three active treatment groups or varying how treatment is administered should be avoided where possible, since all estimates of the AB interaction will otherwise be confounded by differences in treatment. If it is essential to give an additional treatment (e.g. for ethical reasons), the paper should justify this decision and discuss what effect this is likely to have had on outcomes and interactions. If the additional treatment is likely to influence costs, QALYs or clinical outcomes, the study should not be described as factorial.

10.2.2 Analysis

- A clear decision rule or combination of rules determining the situations where interactions will be included in the base case analysis should be pre-specified to avoid data dredging and bias (Section 3.7). My simulation work suggests that including all interactions may minimise the opportunity cost of making the wrong decision (Chapter 5). Criteria based

on statistical significance and goodness of fit had higher opportunity cost and are unlikely to provide a useful indication of the importance of interactions for economic evaluation. In general, interactions should be included unless proven to be negligible, rather than excluded unless proven to be important, since it is preferable to conduct an inefficient analysis including all interactions than introduce bias that changes the conclusions.

- Bayesian analyses attaching informative priors to interaction terms may be useful (Section 3.6.3): particularly if there is prior evidence on the magnitude of interactions or main effects. However, caution should be exercised in the absence of prior evidence, to ensure that any subjective priors are given a suitably low weight in the analysis.
- Regression analysis provides a convenient way to evaluate interactions and main effects, although inside-the-table analysis gives equivalent results in the absence of additional covariates. Including covariates within regression facilitates adjustment for baseline imbalance, which may be particularly important for factorial trials: especially those with many small treatment arms, cluster-randomisation and/or imbalance in characteristics (e.g. baseline utility) that are incorporated within estimates of cost and QALYs (Chapter 8). The model specification should be chosen carefully. If factors are thought to have multiplicative effects, GLM may be more appropriate than OLS on transformed data (Section 3.6.1).
- Any economic evaluation of factorial trials should follow best practice guidelines for trial-based economic evaluation (16-19, 49, 142, 312) and take account of the distribution of costs and benefits, missing data, censoring and correlations between costs and benefits.
- If the likely importance of interactions or the appropriate scale of analysis differ between different types of cost or benefit, it may be useful to analyse these components of NMB separately (Section 3.6.2.5 and Chapter 6). However, any such multivariate analysis should take account of correlations between different NMB components. Furthermore, costs and health benefits should be analysed in the same way across all time periods

unless there is a clear reason (e.g. a change in the treatments administered) why interactions should be present in one period and not another (Chapter 6).

- Multiple imputation models (if used) should include all treatment indicators considered in any analysis as predictors of missing data. In general, it will be necessary to include dummy variables for each factor and a full set of interaction terms. Omitting interaction terms from the imputation model could cause subsequent analyses to underestimate interactions.
- Any extrapolation of trial results should be conducted inside-the-table or allow for interaction terms unless interactions are shown to be negligible. For example, if a simulation model (e.g. UKPDS Outcomes Model (210)) is used to estimate lifetime costs and QALYs from short-term intermediate endpoints, model outputs must be analysed separately for each treatment combination (Chapter 6). When results are extrapolated using parametric survival models (313), researchers could estimate a single model with interaction terms and dummies for each factor, or (if the survival function differs between arms) estimate separate survival models for each cell (or several sets of cells). Similarly, event-based CEA (49, 149) could allow for interactions within model(s) predicting the risk of events.
- Partial factorial trials comprise one of the few situations in which excluding all interactions may be appropriate (Chapter 7). In particular, partial factorial trials in which most patients are randomised to only one comparison have very little power to detect interactions. Furthermore, in partial factorial trials where the proportion of patients receiving each treatment mirrors routine clinical practice, at-the-margins analysis may provide a useful estimate of population average costs and benefits, even in the presence of interactions. Nonetheless, all partial factorial trials should also estimate interactions in those patients randomised to >1 comparison. Bayesian analyses using the costs and outcomes for patients randomised to one comparison as an informative prior within analyses of

patients randomised to >1 comparison may facilitate analysis of the whole sample while avoiding bias from omitted interaction terms.

10.2.3 Reporting

- Abstracts and manuscripts describing economic evaluations based on factorial trials should state that the trial is factorial.
- Conclusions should be based on an incremental comparison between treatment-combinations that considers the factors as interacting treatments (Section 3.5, Chapter 9), regardless of whether interactions are included in the analyses that estimate the mean cost and mean QALYs for each arm.
- It is more appropriate to report results for all factors in a single paper (rather than reporting each factor separately), unless there is clear evidence that treatment effects are additive or that average results are meaningful (as is the case for partial factorial trials). In particular, if it is necessary to allow for interactions in the analysis, it is also essential to interpret the results as a joint decision between all cells in the factorial design, which requires presentation of all factors simultaneously. Nonetheless, in some situations it may be useful to present different factors to different audiences (e.g. surgeons and nutritionists), present early results for one factor before another reaches its primary endpoint, or conduct CMA for one comparison and CUA for another. In these situations, authors may choose to present results for *ab* vs. *b* and *a* vs. *0* separately from the results of *ab* vs. *a* and *b* vs. *0*, in addition to a combined paper if journal editors allow.[84]
- The results of an analysis that includes all interactions should always be reported as a sensitivity analysis whenever the base case analysis excludes any interactions, to

[84] Such dual presentation can also be done within the same paper, as has been conducted for the main papers reporting results of IVAN (accepted by *Health Technology Assessment* (260) and *BMJ Open* [Appendix J]), which report results of CMA comparing continuous Avastin vs. continuous Lucentis and discontinuous Avastin vs. discontinuous Lucentis, as well as results of CUA comparing continuous Avastin vs. discontinuous Avastin and continuous Lucentis vs. discontinuous Lucentis. These manuscripts also present a four-way comparison between all four study arms.

evaluate the potential for bias resulting from assuming additive effects. Additional sensitivity analyses showing the results with different assumptions about interactions are also useful to demonstrate whether or not the analysis is sensitive to interactions. Estimating the opportunity cost of ignoring interactions can help evaluate the impact of excluding interactions.

- Regardless of whether the base case analysis includes all interactions, the mean, disaggregated outcomes in each arm allowing for all interactions should always be reported to enable the reader to see the magnitude and direction of interactions.
- Researchers should present CEACs for multiple interventions, which show the probability that each cell of the factorial design has highest NMB and a cost-effectiveness frontier (202). Pair-wise CEACs showing the probability that A is cost-effective compared with not-A are not sufficient for factorial trials unless there is clear evidence of additive effects. Similarly, value of information should be evaluated for the joint decision between all mutually-exclusive alternatives evaluated in the trial, not separately for each factor.

10.3 Advantages and disadvantages of conducting economic evaluations alongside factorial trials

Given the high likelihood of interactions and the difficulties associated with appropriately analysing economic endpoints from factorial trials, it is informative to ask whether factorial trials are useful vehicles for economic evaluation.

Although statistical efficiency (conducting two trials in one) is the most common reason for conducting factorial trials (1), it is unlikely that economic evaluations of factorial trials would achieve substantial efficiency gains. In particular, the numerous mechanisms by which treatments can have non-additive effects on costs or QALYs (Section 3.4) will reduce the

power of at-the-margins analysis and require interaction terms to be considered within analysis models.

However, the other major advantage of factorial trials, namely evaluation of interactions, is particularly important for trial-based economic evaluation. In particular, one of the major criticisms levied at trial-based economic evaluation is that many such studies do not include all relevant comparators (134). Factorial trials help address this limitation, by evaluating ≥4 alternatives simultaneously, thereby more closely mirroring the real-life decision problem. In the words of Yates, “the experimenter who confines himself to experiments on single factors, making a guess at the final levels of the other factors, is merely emulating the tactics of an ostrich” (46) (page 5). Although in principle, decision analytical modelling (134) and evidence synthesis methods (e.g. NMA (145, 305)) can be used to synthesise comparisons between treatments, these methods require assumptions about the size of interactions and/or exchangeability of evidence from different trials that are directly addressed in factorial trials.

Factorial trials also provide the best possible evidence to inform GCEA (183), or a more integrated system of healthcare decision-making that takes account of interactions between interventions (Chapter 9). Although some of this information could be collected from clinical trials without economic evaluations, factorial trial-based evaluation provides an ideal opportunity to directly assess the effect of one treatment on the incremental cost and QALYs gained through use of another treatment. However, large sample sizes are likely to be required to definitively demonstrate such interactions and piggybacked trial-based economic evaluations are likely to be under-powered.

Factorial trials may also help assess transferability: Glick et al recommend that trial-based evaluations collect data on whether cost-effectiveness is affected by concomitant treatments (17) to help evaluate the extent to which results can be generalised to other countries.

In principle, both factorial trials and subgroup analyses can evaluate the impact of concomitant treatment on treatment effects. For example, we could explore interactions between drugs A and B using either:

1) A 2x2 factorial trial randomising patients to A or not-A, and to B or not-B.
2) A two-arm trial randomising patients to A or not-A, in which patients are allowed to receive/continue concomitant treatment with B. A subgroup analysis evaluates whether concomitant B treatment influences the effect of A.

Subgroup analyses are commonly conducted and have been widely researched: both for clinical endpoints (113, 267, 268, 314-316) and economic evaluation (218, 230, 317, 318). In principle, making stratified decisions to recommend treatment in those subgroups for which the benefits outweigh the harms and/or costs can help avoid exposing patients to harmful or ineffective treatments and help maximise the health gains from the healthcare budget (218, 230, 318). However, many people are sceptical of subgroup analyses, raising concerns, such as: inflating type 1 errors by conducting multiple post hoc comparisons (113, 267, 268, 316, 318); inappropriate statistical methods (113, 267, 314, 316); lack of power (113, 268, 316); confounding or selection bias from conducting subgroup analyses that were not used as stratification factors in randomisation (113, 268, 314); biological plausibility (113, 267, 316); and the ethical implications of giving treatment to some patients and not others (218, 318). Numerous studies have identified statistically significant interactions that are not borne out by subsequent research (267, 268, 314) or are scientifically implausible.

I assert that factorial designs overcome many of these limitations and biases associated with subgroup analyses, which enables us to draw much stronger inferences about causation from factorial trials. Firstly, factorial trials evaluate a small number of pre-specified interactions, generally stratifying randomisation by the second factor and evaluating treatment characteristics rather than patient characteristics.

More fundamentally, factorial trials also randomly assign patients to both treatment factors, which (if accompanied by concealment of allocation and ITT analysis) avoids selection bias. For instance, in Trial 1, the only systematic difference between patients receiving B and those receiving not-B is whether or not they were given B: randomisation ensures that all patient characteristics are randomly distributed between the groups and good experimental design ensures that groups are treated identically from randomisation. If we observe an interaction between A and B that cannot reasonably be explained by chance, we can therefore infer that it is *caused* by B. We can therefore confidently use a factorial trial to inform decisions about whether patients should receive A, B or both.

By contrast, for Trial 2, those patients receiving B are highly likely to differ systematically from those not receiving B. In particular, there are likely to be various patient characteristics that increase the chance of patients having B (e.g. age, sex, comorbidities or disease severity) that may also affect costs or outcomes. If we observe an interaction between A and B that cannot reasonably be explained by chance, this could either be due to B, or to these additional patient characteristics that confound the relationship between B and outcome and introduce selection bias (268). While econometric methods (e.g. regression, propensity scores or genetic matching (289, 290)) can reduce selection bias with respect to observed confounders, all analyses of non-randomised factors unavoidably assume that there are no unobserved confounders.

Factorial trials can also inform broader decisions than subgroup analyses, since the concomitant treatment is under the decision-maker's control (Section 3.8). Subgroup analyses can only inform recommendations about A in different populations (e.g. is A cost-effective for people needing concomitant B treatment? Is A cost-effective for those without B?). By contrast, factorial trials can inform the joint decision about which combination of A and B maximises expected NMB.

10.4 Limitations and future research

Limitations of the individual empirical studies are discussed in each chapter. In particular, my review of factorial trials was conducted four years ago and may have missed some studies. The simulation work explores only nine trials using relatively simple distributional assumptions. Both of the AFORRD analyses presented are prone to bias and include "heroic" extrapolation of 16-week trial data to a lifetime and simplistic imputation of missing data. Conversely, neither IVAN nor KAT include extrapolation beyond the trial period. Nonetheless, these studies provide reasonable examples of how methods for analysing factorial trials can be applied in diverse settings.

Unfortunately, time pressures prevented me from exploring in any detail Bayesian methods for assigning informative priors to the interaction term or main effects, which remain the major area where further research on methods for analysing factorial trials is needed. In addition to being particularly useful for partial factorial trials (Section 7.5.1), such analyses could provide a compromise between including or excluding interactions, by down-weighting interactions that were not expected a priori or which have no clinical explanation. Such analyses could be conducted using MCMC (216) or Bayesian bootstrapping (213, 214). However, more research is needed on the most appropriate way to elicit priors, or define sceptical or subjective priors around interaction terms. My preliminary analyses suggest that the sceptical priors suggested previously (115) may be too informative and have undue impact on results (Section 3.6.4).

I did not attempt to estimate EVSI or EVPPI, although this has been conducted previously (115, 118). However, previous research suggests that different analytical methods tend to affect EVSI in the same direction as EVPI (126). Methods for calculating EVSI for planned factorial trials require further work and development of suitable priors for interactions. In principle, analytical methods similar to those developed for two-arm trials (203) are likely to

be tractable but have not yet been developed. Alternatively, EVSI can be calculated using decision-analytical models (13) or MCMC (118).

This thesis also makes several hypotheses and suggests several methods that have not yet been empirically tested. These include several mechanisms for interactions (Sections 3.4 and 9.3), methods for analysing factorial trials within NMA (Section 9.6.1) and methods for extrapolating factorial trials using parametric models (Section 10.2).

Finally, all of the case studies and worked examples have two qualitative levels per factor and all studies other than AFORRD were analysed as 2x2 matrices. In principle, the methods discussed in this thesis can be adapted for any form of factorial trial. Designs larger than 2x2 will have >1 interaction term, although higher-order interactions may be smaller (9-11, 37). The cells omitted from incomplete and fractional factorial trials prevent estimation of higher-order interactions (9-11, 37, 83), although in principle main effects and two-way interactions can be calculated in the same way for these designs. However, trials with quantitative factors (e.g. dose or levels of financial incentives) may introduce unique opportunities and challenges for economic evaluation since they enable identification of the dose with highest NMB and could be used to compare a potentially infinite number of mutually-exclusive alternatives. Further research is needed to identify the best methods for analysing and interpreting such trials.

Despite these limitations, I hope that the research I have undertaken will stimulate interest in factorial trials as vehicles for economic evaluation, increase awareness of the issues raised by this study design and improve the design, analysis and reporting of future studies.

Appendix A: Protocol for conducting the conceptual literature review on issues and methods for statistical analysis of factorial design trials in general

Aims

A) Develop a clear, unambiguous definition of what comprises a "factorial" trial and identify any synonyms for use in later literature searches.
B) Identify specific types or variants of factorial studies that have distinct characteristics and review the methods recommended for the design and/or statistical analysis of these study types.
C) Identify the rationale for conducting factorial trials and the advantages, disadvantages and challenges of this study design.
D) Identify the main approaches used to analyse data from factorial trials, the advantages and disadvantages of each approach and any guidelines on how factorial trials should be analysed or reported.
E) Identify trends on how and when factorial RCTs are commonly used.

Search strategy

The following databases were searched:

- MEDLINE: searched through PubMed 14th January 2010
- EMBASE, Ovid MEDLINE(R), Ovid MEDLINE(R) Daily Update, Ovid OLDMEDLINE(R): searched through Ovid 29th January 2010

Search terms included:

- (Guideline* AND factorial) OR ("systematic review" AND factorial)
- factorial[Title] AND randomized controlled trial
- factorial[Title] AND Review[ptyp]

A full list of searches by database and date are given below.

As this review aims to identify the key references that give general guidance on factorial trials and the term "factorial" has been in widespread use since the 1930s, it is likely that any papers that would be essential to this review would include the word "factorial" in their abstract and/or title. Furthermore, these searches are not designed to be completely exhaustive and including terms that increase sensitivity at the expense of adding hundreds or thousands of irrelevant articles was considered unnecessary. Subsequently, no synonyms of the term "factorial" were used in the searches for this conceptual review.

Other searches were conducted to address specific aims listed above, including:

- Google searches for synonyms of "factorial design study"
- Cochrane Handbook was searched for the term "factorial"

The bibliographies of the most useful articles and articles that cited the most useful articles were also examined to identify other articles that would inform the aims of this review. Key textbooks on randomised controlled trials were also examined for references that addressed the above aims.

Studies identified in the literature review described in Appendix F that were relevant to this review were also included.

Inclusion criteria

Studies were included in the review if they addressed one or more of the questions outlined above.

However, since preliminary searches suggested that there are (on the whole) only three commonly-used methods for analysing factorial design studies (at-the-margin, inside-the-table and regression/ANOVA), of which regression-based approaches have clear advantages and are widely used, this review was not designed to identify every paper discussing use of factorial trials (still less to identify every applied example of a factorial design trial). Furthermore, since the most appropriate regression technique will vary depending on the application, I did not set out to comprehensively identify every type of regression model used to analyse factorial design studies.

The types of study included in the review included (but was not limited to):

- Guidelines on use and/or statistical analysis of factorial design trials
- Systematic reviews that included only factorial design trials
- Review articles on design and analysis of factorial design trials that help address the above aims
- Methods papers on statistical analysis of factorial design trials that are relevant to the above aims
- Any other papers that are relevant to the above aims

Data extraction

Studies were identified and extracted by one reviewer (HD).

Since this review was not intended to comprise a systematic review and was conducted to get a thorough insight into the literature rather than to identify an unbiased selection of studies answering a single study question, the aims, inclusion criteria, search terms and data extraction methods were adapted as required during the literature searches to ensure that any new issues were covered. However, an initial draft of the main aims, inclusion criteria, search terms and data extraction table was pre-specified prior to the start of the literature searches.

Results of searches

Medline searches:

- (Guideline*[title] AND factorial) OR (“systematic review” AND factorial)
    - Searched on 14th January 2010
    - Yielded 33 hits, of which 6 were relevant
- factorial[Title] AND randomized controlled trial
    - Searched on 14th January 2010
    - Yielded 143 hits, of which 27 were relevant
- factorial[Title] AND clinical trial NOT “randomized”
    - Searched on 17th September 2010
    - Yielded 31 hits

Ovid searches:

The following searches of EMBASE, Ovid MEDLINE(R), Ovid MEDLINE(R) Daily Update, Ovid OLDMEDLINE(R) were conducted on 29th January 2010:

- guideline*.ti. and factorial.af. – 46 hits (inc 5 duplicates)
- ("systematic review" and factorial).af. – 88 hits
- factorial.ti. and randomized controlled trial.af. - 208 hits

- factorial.ti. and review.af. – 75 hits

A further Ovid search of MEDLINE databases was conducted on 29th January 2010 using limits: Factorial [in title] plus limit 1 to "review articles". The databases searched were: Ovid MEDLINE(R), Ovid MEDLINE(R) Daily Update, Ovid OLDMEDLINE(R).

Across all Ovid searches, a total of 372 references were identified (including duplicates. The titles and abstracts of these were scrutinised to identify relevant studies

An additional search of the same Ovid databases conducted on 17th September 2010 using the search terms: factorial.ti. and clinical.af and trial.af yielded 226 hits.

Appendix B: List of references identified in the conceptual literature review of factorial trial literature

The below table summarises the references informing the aims of the systematic review. The third column indicates which the aims or sub-aims of this review are addressed by each publication, which are coded as follows:

0) Definition of a factorial trial [Aim A]
1) Rationale for conducting factorial trials and the advantages and disadvantages of this study design [Aim C]
2) Synonyms for "factorial design" [Aim A]
3) Issues raised by factorial design trials and the extent to which these differ from trials with three or more arms that do not have a factorial design [Aim C]
4) Guidelines on statistical analyses of factorial design trials [Aim D]
5) Approaches used to analyse data from factorial trials and their advantages and disadvantages [Aim D]
6) Subsets of factorial studies that have distinct characteristics and the methods recommended for the design and/or statistical analysis of these study types. [Aim B]
7) Trends in the use of factorial design trials in medicine [Aim E]

Reference	Type	Address aims*	Notes
Books and articles on factorial design in general			
Cox 1958 (10), chapters 6 and 7	Book	0,1,5,6	Chapter(s) of book on experimental design
Fisher 1966 (first published 1935) (28)	Book	0,1,5,6	Seminal textbook on experimental design
Finney 1955 (9) Chapter VI	Book chapter	0,1,5,6	Chapter(s) of book on experimental design
Bernaerts et al 2005 (32)	Guideline	6	Guideline for experimental design, discussing methods of selecting which cells to evaluate and how to model fractional factorial designs
Festing 2001 (70)	Guideline	1,6	Guidelines on design and analysis for in vitro or animal experiments. Gives definition of factorial design
Fisher 1926 (27)	Review	0,1	Described the factorial design (without using the term "factorial" and expounded the advantages of this design, as well as justifying the rationale for a 5% significance level and randomised trials.
Yates 1935 (26)	Review	0,1,5,6	Outlines the definition, features, advantages, disadvantages and analytical methods of factorial designs
Yates 1937 (46)	Review	0,1,5,6	Seminal paper outlining analytical methods for factorial trials. This covers definition, advantages and disadvantages of factorial trials and describes methods for calculating main effects and interactions for a very wide range of different factorial designs, but focuses on manual calculation and dealing with confounding
Montgomery 2005 (11) chapters 5, 6 and 9	Textbook	0,1,5,6	Comprehensive overview and outline of statistical methods for design and analysis of factorial trials in wider applications (particularly industry)
Muller 2002 (45) chapters 14 and 15	Textbook	0,1,5,6	Statistics textbook, which summarises methods for ANOVA on factorial experiments or observational data, including random effects
Neter 1996 (36) chapters 19-23	Textbook	0,1,5,6	Statistics textbook, which summarises methods for ANOVA and regression for factorial experiments and multi-factor observational studies.

Appendix B: List of references identified in the conceptual literature review of factorial trial literature

Reference	Type	Address aims*	Notes
Textbooks, reviews and guidelines on conduct of clinical RCTs in general and factorial RCTs in particular			
Chow and Liu 2004 (100)	Book chapter	2,5	Textbook on clinical trials. Describes use of factorial trials to evaluate combination therapies & test for global superiority of the combination, but nothing on other aspects of factorial trials
Fox 2006 (47) Chapter 11	Book chapter	0,1,4,6,7	Chapter of a textbook on RCTs. Gives a comprehensive but accessible overview of factorial trials and their advantages, disadvantages and analysis in clinical RCTs
Gray et al 2008 (95)	Book chapter	1,7	Chapter in a textbook on cancer treatment. Comprises a general discussion of RCTs and the advantages of randomisation and large trials. Emphasises the usefulness and hazards of subgroup analyses and exploration of heterogeneity in large factorial trials. Comments that "any obstacle to simplicity is an obstacle to large [sample] size" but emphasises importance of avoiding unduly complicated eligibility criteria & excessive numbers of outcome measures rather than factorial design
Jadad 1998 (69) Chapter 2	Book chapter	0	Includes a definition of factorial RCTs
Piantadosi 2005 (33)	Book chapter	0,2,5,6,7	Textbook chapter covering many aspects of factorial trials
Portney and Watkins 2009 (67)	Book chapter	0,1,2,5,6	Textbook on clinical research
Redmond 2001 (63) factorial trial chapter	Book chapter	0	Textbook chapter including definition, pros and cons of factorial trials. Summarises design and analysis of factorial trials, but appears to draw heavily from 1st edition of (33)
Cochrane handbook (24)	Guideline	0,1,4	Guidelines on the appropriate conduct of systematic reviews and meta analyses. With respect of factorial trials, this includes the definition and rationale, recommendations on reporting and recommendations on use of factorial trials in systematic reviews and meta-analyses.
CONSORT statement (124)	Guideline	4	Pivotal guidelines for reporting RCTs in general
Food and Drug Administration (FDA) (319)	Guideline	4	Guidance on evidence for submissions to FDA. Comments that "Properly designed factorial studies may be analyzed as a series of pairwise comparisons, representing, within a single study, separate demonstrations of activity of a drug as monotherapy and in combination with another drug".
Lammertse et al 2007 (64)	Guideline	0,1	General guideline for clinical trials on spinal-cord injury, reviewing different trial methods. Defines factorial trials and summarises when they might be used in spinal-cord research
Montgomery et al 2003 (20)	Guideline	0,1,3,4,5	Full guideline on use of factorial trials. This is considered a core part of the evidence underpinning the CONSORT statement http://www.consort-statement.org/resources/database/evidence-underpinning-consort/trial-design-issues/factorial-trials/
Akritas et al 1997 (320)	Methods paper	5	Develop non-parametric methods for testing hypotheses of no main effects, no interaction effects and no simple effects in factorial trials with censored ordinal data
Blyth and Gebski 2004 (8)	Methods paper	1,5	Discuss power calculations indicating the power of factorial trials for main effects and detecting interactions with different sizes of interaction and how this compares with

Appendix B: List of references identified in the conceptual literature review of factorial trial literature

Reference	Type	Address aims*	Notes
			two 2-arm trials and a three-arm trial of the same total sample size
Brittain and Wittes 1989 (30)	Methods paper	0,1,5,6,7	Pivotal review article on factorial trials, which comprehensibly outlines the advantages and disadvantages of factorial studies, and when they should and shouldn't be used
Brown et al 2006 (116)	Methods paper	5	Applied example using Bayesian analysis
Byar et al 1993 (81)	Methods paper	6	Review of incomplete and fractional factorial designs with applied examples
Fletcher et al 1990 (321)	Methods paper	6	Proposes methods for crossover factorial designs: particularly in relation to how the sequences of treatments are chosen
Gail and Simon 1985 (107)	Methods paper	5	Define the concept of qualitative and quantitative, crossover and non-crossover interactions between treatment and patient characteristics. Develop a likelihood ratio test for qualitative interactions
Green 2006 (65)	Methods paper	0,1,4	Methodological study published as a book chapter. This reviews methods for analysing factorial trials (primarily at-the-margins, two-step approach and global tests followed by two-step) and tests them using a simulation study. Outlines literature on factorial trials, advantages and disadvantages.
Green et al 2002 (31)	Methods paper	1,5	Simulation study comparing at-the-margins and two-step approaches with/without adjustment for multiple comparisons on clinical endpoints.
Hafner and Ruberg 1996 (103)	Methods paper	1,4,6,7	Propose using factorial trials to evaluate frequency and magnitude of dosing, rather than basing frequency on pharmacokinetic properties and then assessing dose in parallel group trials
Holms and Berrettoni 1969 (322)	Methods paper	6	Outlines methods for ANOVA in replication-free factorial experiments
Hung 1990 (101)	Methods paper	1,5	Extends the statistical tests for showing whether there is a dose combination that is superior to its ingredients to trials with baseline imbalance.
Hung 1990 (97)	Methods paper	1,5	Describes methods for response surface mapping and testing global superiority of combination therapy
Hung 1992 (98)	Methods paper	4	Describes a quadratic polynomial model for dose-ranging factorial studies with several doses of A and several doses of B
Hung 1993 (3)	Methods paper	5,6	Clear algebraic description of the two-stage approach for analysing 2x2 factorial trials and demonstration of the key features
Hung 1996 (82)	Methods paper	5	Expands on tests for whether combination therapy is more effective than its component doses
Hung 2000 (84)	Methods paper	4,5,6	Looks at evaluation of combination therapy in multiple doses in unbalanced factorial design trials. Presents two methods for conducting global tests on whether combination therapy is more effective than monotherapy in dose-ranging unbalanced factorial trials and their pros and cons, as well as power calculations.
Hung et al 1989 (102)	Methods paper	1,5	Evaluates three statistical procedures for testing whether *ab* combination therapy is superior to *a* or *b* within large factorial designs evaluating several dose combinations.
Hung et al 1993 (99)	Methods paper	1,7,4	Comments on advantages of using factorial studies to evaluate different doses of combination therapy

Appendix B: List of references identified in the conceptual literature review of factorial trial literature

Reference	Type	Address aims*	Notes
			components. Proposes a test statistic to evaluate whether at least one dose combination is more effective than its components, without appealing to any dose-response model.
Hung et al 1995 (4)	Methods paper	1,5,6,7	Outlines how to calculate main effects, simple effects and interaction manually. Outlines a two-stage estimation/testing procedure, whereby interactions are tested first, to determine whether the effect of each monotherapy should be estimated using main effects or simple effects. Comments on appropriateness of factorial design and interaction term in cases where 1 treatment has no effect.
Lindblad 1996 (105)	Methods paper	4,5,7	Evaluates effect of non-compliance on power in factorial studies (both in terms of patients not taking study medication and in terms of patients taking the same medicines or supplements on top of study medication).
Miller and Seaman 1998 (117)	Methods paper	5	Describes Bayesian approach to assessing superiority of dose combinations and a hierarchical prior structure for factorial trials that can incorporate informative and non-informative priors
Natarajan et al 1996 (73)	Methods paper	5,6,7	Simulation study presenting methods for power calculations for survival data
Ng 1991 (5)	Methods paper	1,5	Calculates type I error rates for two-stage procedure, inside-the-table and at-the-margins approach at different sizes of interaction term and different significance levels used for the interaction
Patel and Tsai 1996 (109)	Methods paper	4	Analysis of paired ordered categorical data in factorial studies: suggests methods to evaluate ordinal outcome measures when have before and after for each patient in a factorial study
Peters et al 2003 (323)	Methods paper	6	Methods paper on cluster randomised trials using an example that is factorial. Compare three approaches for analysing the study at the cluster level and five at individual level.
Peterson and George 1993 (74)	Methods paper	5	Presents sample size and trial duration calculations for detecting interactions in survival endpoints in factorial trials and for estimating main effects in the presence of an interaction
Pool et al 1997 (324)	Methods paper	5,7	Example of a 4x4 factorial trial that models dose relationships using a quadratic function
Simon and Freedman 1997 (42)	Methods paper	4,5,1	Considers use of Bayesian analyses and informative priors within the design and analysis of factorial trials
Slud 1994 (40)	Methods paper	1,5,6	Models for survival outcomes in factorial studies
Stewart 1996 (112)	Methods paper	5	Applies response surface methods to dose-ranging factorial trials
Welton et al 2008 (115)	Methods paper	5	Applied example using Bayesian analysis and informative priors
Xiang et al 1994 (94)	Methods paper	1,4	Discusses estimators and power calculations for testing for interactions in factorial studies with a 2xk design that use a failure time outcome: these comprise the Mantel-Haenszel method, a method that is based on the ratio of observed to expected results and maximum likelihood estimation
Ottenbacher 1991 (106) and	SR, methods paper	4,5	Study systematically reviewing factorial trials published in three medical journals and comments on how they interpreted the interaction term from ANOVA. Identified 83

Appendix B: List of references identified in the conceptual literature review of factorial trial literature

Reference	Type	Address aims*	Notes
subsequent letter by Janosky 1992 (325)			studies with statistically significant interactions, although it is not clear what proportion of these were factorial trials and which were two-arm trials with subgroup analyses.
Bria et al 2006 (41)	Other	1,5	Letter highlighting the dangers of factorial design trials (or rather the dangers of an at-the-margins approach) when there are interactions in the context of an article on pacletaxel vs. docetaxel
Campbell et al 2007 (96)	Other	1	HTA monograph describing an applied example of a failed clinical trial that discusses the reasons for slow recruitment
Last 1988 (72)	Other	0	Epidemiology dictionary defining factorial RCT
Lubsen and Pocock 1994 (2)	Other	1,4,5,7	Editorial discussing the pros & cons of factorial designs in the context of the ISIS-4 trial. Gives guidance on analysis and reporting of factorial trials. Highlights the disadvantages of the at-the-margins approach, including a great outline of what is meant by an interaction effect or lack thereof.
Allore and Murphy 2008 (61)	Review	1,6	Describes “standardly tailored design” of study, whereby the individual components of the intervention are assigned to different patients “as needed”, as well as the partial factorial design, incomplete factorial design, fractional factorial design and modified reciprocal control design
Buring and Hennekens 1990 (85)	Review	1	Review summarising the advantages of factorial trials in the context of discussing the rationale for the design of the Physicians’ Health Study.
Byar 1990 (37)	Review	6,0,1,7,5	Reviews factorial design trials (particularly 2^p designs) and fractional factorial trials. Proposes “reciprocal control designs”
Byar and Piantadosi 1985 (35)	Review	1,5,6,7	Pivotal review article on factorial trials. Discusses how factorial trials are analysed, there are advantages and disadvantages and when they are most appropriate. Discusses interpretation of interaction terms and whether adjustment for multiple comparisons is necessary. Gives examples showing sample size and power from different study designs evaluating the same treatments
Couper et al 2005 (7)	Review	1,5,6,7	Comprehensive review of the issues around factorial trials, including definition, pros and cons, design issues and the main methods for analysis
Korettila 2005 (39)	Review	1,7	Review commenting on the advantages and disadvantages of factorial trials and trends in their use.
Nair et al 2008 (83)	Review	6	Thorough examination of fractional factorial designs, including reference to a trial in health
Sheikh et al 2002 (38)	Review	1	Outlines study designs in general and their pros and cons, including a brief discussion of factorial trials
Simon 1991 (90)	Review	1	Reviews advancements in trial methods during the 1980s, including the advantages and disadvantages of factorial trials
Stampfer et al 1985 (34)	Review	1,5,6,7	Reviews use of factorial trials and their advantages and disadvantages, with particular reference to Physicians’ Health Study
Stolberg et al 2004 (75)	Review	1	Briefly discuss advantages of factorial design
Wing et al 1987 (104)	Review	1,6	Non-systematic review of a small number of factorial crossover trials in hypertension.
Prescott et al 1999 (89)	Methods paper	7,1	Systematic review of factors limiting the quality, number and progress of RCTs. Factorial design is mentioned as one of the factors for consideration in research design,

Appendix B: List of references identified in the conceptual literature review of factorial trial literature

Reference	Type	Address aims*	Notes
			although factorial design isn't mentioned elsewhere. Identified one study (Foley & Moertel 1991) that found that 27% of clinicians said that complex studies were a barrier to recruitment.
Hopewell et al 2010 (21)	SR	7	Systematic review of all RCTs indexed in Pubmed 2000-2006. Identified only 10 factorial trials out of 616, suggesting that factorial trials account for only 1.6% of trials. By contrast, 16% of trials (97/616) had 3 arms and 8% (51/616) had 4 or more arms without being classed as factorial.
McAlister et al 2003 (1)	SR	2,3,4,5,7	Systematic review of factorial RCTs. Identified 44 factorial trials with binary endpoints. Pivotal reference introducing the terminology inside-the-table and at-the-margins and making recommendations on appropriate analysis and reporting of factorial trials.
Silagy and Jewell 1994 (22)	SR	7	Review of RCTs published in British Journal of General Practice between 1953 and 1991, which found that 4% of the 90 RCTs identified were factorial.
Bowling 2009 (66)	Textbook	0,2	Summarises definition and advantages of factorial trials, focussing on generalisability.
Earl-Slater 2002 (76)	Textbook	0,1	Textbook giving a definition of a factorial trial and discussing advantages/disadvantages
Meinhert 1986 (68)	Textbook	0,2,6	Glossary of key terms
Torgerson and Torgerson 2008 (71)	Textbook	0,1,5,7	Defines what factorial trials are, their advantages and disadvantages and how they are typically analysed
Wald 2004 (87) pages 22-4	Textbook	1,5	Textbook on epidemiology, which includes a brief discussion of the advantages, disadvantages and analysis of factorial trials.

Abbreviations: SR, systematic review.

* Numbers relate to which the aims or sub-aims of this review are addressed by each publication:
  0) Definition of a factorial trial [Aim A]
  1) Rationale for conducting factorial trials and the advantages and disadvantages of this study design [Aim C]
  2) Synonyms for “factorial design” [Aim A]
  3) Issues raised by factorial design trials and the extent to which these differ from trials with three or more arms that do not have a factorial design [Aim C]
  4) Guidelines on statistical analyses of factorial design trials [Aim D]
  5) Approaches used to analyse data from factorial trials and their advantages and disadvantages [Aim D]
  6) Subsets of factorial studies that have distinct characteristics and the methods recommended for the design and/or statistical analysis of these study types. [Aim B]
  7) Trends in the use of factorial design trials in medicine [Aim E]

Appendix C: Types of factorial trial

Incomplete and fractional factorial designs

Fractional factorials (also known as fractional replication (9)) comprise a specific form of factorial experiment in which a defined fraction (e.g. a half or a quarter) of the treatments are omitted (6, 9, 11, 37, 61, 83) (Table C1); this design is generally used when the total number of treatments required for a full factorial design exceeds the number of experimental units available or the number of treatment conditions that is feasible (11, 37, 83). Incomplete factorial trials[85] comprise a related design in which one or more factor combinations are omitted from the design without the design corresponding to any fractional factorial design (61, 81, 82) (Table C1). These designs are often chosen due to practical or ethical concerns about specific treatment combinations, such as the ethics of denying patients treatment (81) or the toxicity of a triple combination (37, 47, 61, 81).

Table C1. Illustration of incomplete and fractional factorial trial designs for a 2^4 design

Treatment combination	Full factorial	Incomplete factorial: omitting 0 and abcd	Half fractional factorial
0	✓	X	✓
a	✓	✓	X
b	✓	✓	X
c	✓	✓	X
d	✓	✓	✓
ab	✓	✓	✓
ac	✓	✓	✓
ad	✓	✓	X
bc	✓	✓	✓
bd	✓	✓	X
cd	✓	✓	X
abc	✓	✓	X
abd	✓	✓	✓
acd	✓	✓	✓
bcd	✓	✓	✓
abcd	✓	X	X

Symbols: ✓, treatment directly tested in the trial; X, combination omitted from the trial.

Although all four cells in a 2x2 factorial design are required to assess interactions (24, 33) (making it inappropriate to class three-arm trials as incomplete factorials), it is feasible to omit some treatments and still estimate main effects and some (but not all) interaction effects from factorial designs with more than two factors (9, 33, 36) or designs in which at least one factor is investigated at three or more quantitative levels if some additional assumptions are made. In particular, incomplete and fractional factorials tend to assume that higher-order interactions are *generally* negligible (9-11, 37, 83) (the "sparsity of effects principle"). Providing that the treatments included in the experiment are chosen carefully to balance the factors (9, 32), main effects and lower-order interactions can be estimated efficiently without the need for all treatments to be tested directly (9, 32, 61). As result, incomplete and fractional designs qualify as factorial trials under my definition.

Fractional factorials are rare in medicine (61) but are commonly used in agriculture and engineering, being particularly useful for studies with more than four factors (9, 35), such as early research screening multiple treatments to identify those influencing response (11).

[85] Several authors refer to incomplete factorial designs as "partial factorial designs" (30, 33, 47, 68); however, I follow the terminology of Byar (81) by using the term "incomplete factorial" to refer specifically to those trials omitting certain treatment combinations from the factorial design and reserve the term "partial factorial" to refer specifically to trials such as KAT (59) or Women's Health Initiative (62) in which some but not all patients are randomised to more than one comparison.

However, they may be a useful approach for factorial crossover trials: particularly when the number of factor combinations is larger than the number of treatment phases that can feasibly be conducted (104, 321). The recent CORONIS trial used a 2x2x2x2x2 fractional factorial design (326).

Partial factorial designs

Within my thesis, I define partial factorial trials as those in which some but not all patients are randomised to more than one comparison in a factorial manner, while other patients are randomised to just one comparison or to a different combination of factors. Several high-profile clinical trials use a partial factorial design (59, 61, 62, 263). However, some such studies (e.g. the UKPDS (263)) do not refer to their design as “partially factorial” and the term is sometimes used to describe incomplete factorial designs (30, 33, 47, 68). Reports of the CORONIS trial (in which each patient was randomised in three of the five comparisons) described it is designed as an "non-regular fractional factorial" (326). Partial factorial designs are often used when economic, geographic, or clinical constraints restrict the comparisons to which patients can be randomised (61). For example, some partial factorial trials conduct the second randomisation only for certain countries (264), while others randomise patients with specific comorbidites, clinical characteristics or laboratory findings to a second (or even third) comparison (62, 265) or recruit patients only to those comparisons for which their surgeon is in equipoise (59) or for those patients who choose to take part (62).

Although these trials raise unique issues, such as the fact that unbiased estimation of interaction effects can be estimated only on a subset of patients (Chapter 7), my literature review identified no publications discussing the appropriate methods for analysing partial factorial trials. The implications of a partial factorial design for analysis of clinical and economic endpoints is discussed further in Chapter 7 with reference to the KAT trial (59, 60).

Unbalanced factorial designs

In unbalanced factorial trials, different factor combinations have different sample sizes (47, 84). Small imbalances frequently arise due to appropriate randomisation methods, missing data (36, 45) and withdrawals (36). However, marked variation in sample size may be planned from the outset to allocate a larger sample size to the dosing regimen most likely to be marketed (47, 84), the least costly treatment (36, 62) or to maximise the power of pre-specified comparisons between pairs of treatments (36). Unbalanced designs complicate analysis (Section 2.6), but nonetheless allow assessment of interactions and assess ≥2 factors, thereby qualifying as factorial trials. Treatment means may also vary in importance, in which case it may be appropriate to assign weights to each treatment mean (which may or may not equal their relative sample size) for estimation and hypothesis testing (36).

Large factorial designs

2^k designs with large numbers of factors are commonly used in the early stages of research for screening multiple factors (11). Increasing the number of factors obviously increases the number of questions that can be addressed with a single experiment, but also increases trial complexity and the risk of non-compliance or interactions that can introduce substantial bias into estimated treatment effects (30, 35). Although most clinical trials use a 2x2 design (Chapter 4), some trials have evaluated 5-6 factors (327, 328).

Studies with more than two levels are often used to estimate the magnitude and shape of dose response relationships (10, 11), which is often done using response surface techniques (10, 11, 26, 46, 97, 98, 100, 112). However, including more than two levels per factor makes assessment of interactions more complicated: particularly for qualitative factors present at

three or more levels (11). Including a centre point in a 2^k design may enable curvature estimation without substantially increasing study complexity (11). Most clinical trials use two levels, although 5x3 (329) and 4x3 (330, 331) designs have also been used.

Large factorial experiments in agriculture are often conducted without replication (i.e. with just one observation per treatment) by assuming that higher-order interactions are negligible, which is often used for large factorial trials when the number of subjects is limited (9-11, 26, 28, 36, 46, 322). However, this design is unlikely to be appropriate in medicine due to variability between patients and the limited usefulness of very large factorial trials. Confounding, randomised block designs and split plot designs are also often used in other fields for large factorial designs in which it is not feasible to test all factor combinations in a single homogenous block; in these designs, each replication is conducted over two or more blocks, with the cells arranged such that all important comparisons are conducted within each block and higher-order interactions are confounded (9, 10, 28, 46).

Related non-factorial designs

Several other designs have been proposed that share some characteristics with factorial trials but do not fulfill my definition of factorial trials as they do not permit interactions to be assessed.

In reciprocal control trials (or “balanced incomplete block designs” (332, 333)), patients are randomised to two groups and receive interventions targeted at different endpoints, with each group acting as the control for the other (37, 61). For example, Zelen proposed a trial in which 50% of patients are randomised to undergo breast cancer screening while the other receive cervical cancer screening; if we assume no interactions, we can assess the effect of breast cancer screening on breast cancer mortality and the effect of cervical cancer screening on cervical cancer mortality by comparing the two groups. This design allows multiple questions to be addressed in the same study, but does not allow interactions to be assessed. Although interactions may be less common for treatments targeting different endpoints (37), estimates from such designs may nonetheless be confounded by unforeseen interactions that may introduce bias without any opportunity to investigate such interactions. In particular, some patients (e.g. smokers or those with genetic predisposition) are at risk of having both cancers in a lifetime, so small interactions may occur as these diseases represent competing risks of death, and ensuring early diagnosis of one cancer may indirectly increase the chance of the other being detected or eradicated during systemic treatment. This design has been used to assess the effect of computerised guidelines for angina and asthma (332, 333) and could be combined with a factorial design to address three or more questions simultaneously (37).

Allore also proposed a standardly tailored design as an alternative to a factorial design (61); in such a study, a heterogeneous patient cohort with a mixture of risk factors is randomised to one of two groups. One group receives usual care, while the other receives pre-specified treatments for all of the risk factors that they display (e.g. losartan for all hypertensive patients, atorvastatin for those with hypercholesteraemia and travaprost for those with ocular hypertension). Such a trial assesses the effect of the entire package of measures in a broad population but cannot assess interactions or the effects of individual treatments.

Appendix D: Examples of different types of interaction

This appendix describes worked examples illustrating different mechanisms by which interactions can arise. Examples A-G are used within the league tables shown in Section 9.4; for these examples, additional tables showing the incremental costs, QALYs and cost-effectiveness inside-the-table and at-the-margins are included. Each example with interactions is cited in the taxonomy of interactions in Table 3.3; however, Example G (which has additive effects) is omitted from Table 3.3, since it is only used for illustration in Section 9.4.

Within each of the examples shown, interactions are calculated as $\mu_0 + \mu_{ab} - \mu_a - \mu_b$ (the outcomes for *0* – outcomes for *a* – outcomes for *b* + outcomes for *ab*). For simplicity, no discounting is applied in any of the examples, although this would not affect the conclusions.

All examples are hypothetical and should not be interpreted as reflecting actual clinical practice, efficacy or cost.

Example A: Interaction type 2a (multiplicative effects on the odds of clinical events)

Statins and antihypertensives are both assumed to decrease the odds of cardiovascular disease (CVD) events, with no interaction between the two types of drug on a log-odds scale. As result, as result, there is a sub-additive effect on the probability of CVD events and on QALYs and a super-additive effect on costs.

Table D-A1. Data inputs

Odds ratio of CVD: statins	0.4
Odds ratio of CVD: antihypertensive	0.6
Baseline odds of CVD over 10-year time horizon	0.2
QALY loss from CVD	1
Cost per case of CVD	£3,000
QALYs without CVD	8
Drug cost over the 10-year time horizon: statins	£2,190
Drug cost over the 10-year time horizon: antihypertensives	£1,100
Population	1,000,000

Table D-A2. Outcomes for each treatment combination

	Probability of CVD*	Population Cost†	Population QALYs‡	Population NB§
No treatment	17%	£500,000,000	7,833,333	£156,166,666,667
Antihypertensive only	11%	£1,416,428,571	7,892,857	£156,440,714,286
Statin only	7%	£2,412,222,222	7,925,926	£156,106,296,296
Statin + antihypertensive	5%	£3,422,404,580	7,954,198	£155,661,564,885
Interaction	3%	£93,753,787	-31,251	-£718,779,030

* Probability = odds/(1+odds)
† Cost = probability of CVD*cost per case of CVD + drug cost, all multiplied by the population size
‡ QALYs = Population*(QALYs without CVD – QALY loss from CVD*probability of CVD)
§ At a £20,000/QALY ceiling ratio.

Table D-A3. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	Incr QALYs	ICER
Antihypertensive (A1)	£963,305	43.90	£21,944
Statin (A2)	£1,959,099	76.97	£25,454

Table D-A4. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Antihypertensive only (A1) vs. A0	£916,429	59.52	£15,396
Statin only (A2) vs. A1	£995,794	33.07	£30,113
Statin + antihypertensive (A3) vs. A2	£1,010,182	28.27	£35,730

We can eliminate the interaction by evaluating cost-effectiveness separately for patients at differing cardiovascular risk.

Example B: Interaction type 2b (multiplicative effects between quality and length of life)

Knee replacement doubles patients' utility for all remaining life-years, while smoking cessation interventions (if successful) increase life expectancy by 50%. As result, there is a large super-additive interaction for QALYs, but additive effects on cost (since both interventions are assumed to be one-off). The actual interaction could be substantially larger than this if stopping smoking before knee replacement surgery reduced operative mortality and morbidity.

Table D-B1. Data inputs

Life expectancy: 70-year-old smoker	8
Life expectancy: 70-year-old ex-smoker	12
Utility without knee replacement	0.3
Utility with knee replacement	0.6
Cost of knee replacement	£5,000
Lifetime cost of smoking-related diseases - smoker	£2,000
Lifetime cost of smoking-related diseases - ex-smoker	£1,000
Cost of smoking cessation intervention	£50
Effectiveness of smoking intervention: % of smokers who stop	20%
Population	60,000

Table D-B2. Outcomes for each treatment combination

	Population Cost†	Population QALYs‡	Population NB§
No treatment	£120,000,000	144,000	£2,760,000,000
Smoking cessation only	£111,000,000	158,400	£3,057,000,000
Knee replacement only	£420,000,000	288,000	£5,340,000,000
Knee replacement + smoking cessation	£411,000,000	316,800	£5,925,000,000
Interaction	£0	14,400	£288,000,000

† Cost = intervention cost + cost of smoking-related diseases for smokers*proportion not quitting + cost of smoking-related diseases for ex-smokers*proportion quitting, all multiplied by population size
‡ QALYs = Population*relevant utility*(life expectancy for smokers*proportion not quitting + life expectancy for ex-smokers*proportion quitting)
§ At a £20,000/QALY ceiling ratio.

Table D-B3. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	Incr QALYs	ICER
Smoking cessation (B1)	-£9,000	21.60	Dominant
Knee replacement (B2)	£300,000	151.20	£1,984

Table D-B4. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Smoking cessation only (B1) vs. B0	-£9,000	14.40	Dominant
Knee replacement only (B2)	£309,000	129.60	Dominated
Knee replacement + smoking cessation (B3) vs. B1	£300,000	158.40	£1,894

We can eliminate the interaction by stratifying by life expectancy.

Example C: Interaction type 2c (multiplicative effects on drug cost)

There are two drugs that can be used to treat anxiety (A and B), which can each be used once-daily or twice-daily. An interaction for cost arises because the cost of twice-daily treatment is always twice as high as the cost of once-daily treatment, regardless of drug costs, which means that doubling dosing frequency increases costs more for the more costly drug (B).

Table D-C1. Data inputs

Cost of once-daily Drug A/year	£300
Cost of once-daily Drug B/year	£370
Cost of twice-daily Drug A/year	£600
Cost of twice-daily Drug B/year	£740
QALYs with once-daily Drug A	0.60
QALY difference between A and B	0.01
QALY difference between twice-daily and once-daily	0.10
Population	8,000

Table D-C2. Outcomes for each treatment combination

	Population Cost†	Population QALYs‡	Population NB§
Once-daily Drug A	£2,400,000	4,800	£93,600,000
Once-daily Drug B	£2,960,000	4,880	£94,640,000
Twice-daily Drug A	£4,800,000	5,600	£107,200,000
Twice-daily Drug B	£5,920,000	5,680	£107,680,000
Interaction	£560,000	0	-£560,000

† Cost = intervention cost multiplied by population size
‡ QALYs = Population*QALYs per patient
§ At a £20,000/QALY ceiling ratio.

Table D-C3. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	Incr QALYs	ICER
Drug B for anxiety vs. Drug A (C1)	£840	0.08	£10,500
Twice-daily treatment for anxiety vs. once-daily (C2)	£2,680	0.80	£3,350

Table D-C4. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Once-daily Drug B for anxiety (C1) vs. C0	£560	0.08	Ex dominated
Twice-daily Drug A for anxiety (C2) vs. C0	£1,840	0.72	£3,000
Twice-daily Drug B for anxiety (C3) vs. C2	£2,960	0.80	£14,000

It is essential to make a joint decision on dosing frequency and which drug to use.

Example D: Interaction type 3 (decreasing marginal effect on utility)

Either of the two painkillers will improve patients' response level on the pain domain by 1 level. Given individually, they can improve pain from extreme to moderate, but given together they can eliminate pain altogether. As result, there is a sub-additive interaction for QALYs, since the coefficients for level 3 pain are more than twice as high as those for level 2 pain, in addition to level 3 pain attracting the N3 term. Effects on costs are additive.

Table D-D1. Data inputs

Painkiller 1 cost/4 months	£600
Painkiller 2 cost/4 months	£1,600
EQ-5D utility with no drug (11232)	0.157
EQ-5D utility with painkiller 1 (112**2**2)	0.689
EQ-5D utility with painkiller 2 (112**2**2)	0.689
EQ-5D utility with 2 painkillers (112**1**2)	0.812
Population	30,000

Table D-D2. Outcomes for each treatment combination

	Population Cost†	Population QALYs‡	Population NB§
No treatment	£0	1,570	£31,400,000
Painkiller 1 only	£18,000,000	6,890	£119,800,000
Painkiller 2 only	£48,000,000	6,890	£89,800,000
Painkillers 1 & 2	£66,000,000	8,120	£96,400,000
interaction	£0	-4,090	-£81,800,000

† Cost = intervention cost multiplied by population size
‡ QALYs = Population*relevant utility*0.33 years
§ At a £20,000/QALY ceiling ratio.

Table D-D3. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	Incr QALYs	ICER
Painkiller 1 (D1): Patients with pain only	£18,000	3.28	£5,496
Painkiller 2 (D2): Patients with pain only	£48,000	3.28	£14,656

Table D-D4. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Painkiller 1 only (D1) vs. D0: Patients with pain only	£18,000	5.32	£3,383
Painkiller 2 only (D2) vs. D1: Patients with pain only	£30,000	0.00	Dominated
Painkillers 1 & 2 (D3) vs. D1: Patients with pain only	£48,000	1.23	£39,024

A joint decision is needed to allow for the interaction. Similar trends would be seen on any other domain.

Example E: Interaction type 3 (increasing marginal effect on utility)

The painkiller reduces patients' pain from extreme to moderate (i.e. from level 3 to level 2 on EQ-5D), while the antidepressant has the same effect on the anxiety/depression domain. As result, there is a large super-additive interaction on utilities and QALYs due to the EQ-5D tariff, which subtracts 0.269 from utility for any patients with level 3 on any domain: both anxiety/depression and pain must be improved to above level 3 to remove the N3 term.

Table D-E1. Data inputs

Painkiller cost/4 months	£900
Antidepressant cost/4 months	£250
EQ-5D utility with no drug (11233)	-0.008
EQ-5D utility with painkiller (112**2**3)	0.255
EQ-5D utility with antidepressant (1123**2**)	0.157
EQ-5D utility with painkiller & anxiety and depression (112**22**)	0.689
Population	20000

Table D-E2. Outcomes for each treatment combination

	Population Cost†	Population QALYs‡	Population NB§
No treatment	£0	-53	-£1,066,667
Antidepressant only	£5,000,000	1,047	£15,933,333
Painkiller only	£18,000,000	1,700	£16,000,000
Painkiller + antidepressant	£23,000,000	4,593	£68,866,667
interaction	£0	1,793	£35,866,667

† Cost = intervention cost multiplied by population size
‡ QALYs = Population*relevant utility*0.33 years
§ At a £20,000/QALY ceiling ratio.

Table D-E3. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	Incr QALYs	ICER
Antidepressant (E1): Patients with comorbid pain and depression	£5,000	2.00	£2,504
Painkiller (E2): Patients with comorbid pain and depression	£18,000	2.65	£6,792

Table D-E4. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Antidepressant only (E1) vs.E0: Patients with comorbid pain and depression	£5,000	1.10	£4,545
Painkiller only (E2) vs.E1: Patients with comorbid pain and depression	£13,000	0.65	Ex dominated
Painkiller + antidepressant (E3) vs. E1: Patients with comorbid pain and depression	£18,000	3.55	£5,075

A joint decision is needed to allow for the interaction. Similar effects would also be seen for other domains and for combinations of treatments that improve patients to perfect health. Even larger interactions would be seen if improved pain relief also improved anxiety/depression or vice versa.

Example F: Interaction type 4a (earlier intervention [screening] affects costs and benefits of later intervention [treatment] and vice versa)

Treating the patients diagnosed with cervical cancer with the new treatment rather than the comparator treatment increases costs and improves quality-adjusted survival for patients at all cancer stages and therefore increases the cost and reduces the benefits of screening. Conversely, screening patients for cervical cancer will increase the chance that cancers are detected at an early stage (e.g. stage I); total costs and quality adjusted survival vary depending on the stage of cancer at which patients are diagnosed. Additionally, the incremental costs and incremental QALYs associated with the new treatment compared with its comparator also vary with stage.

As result of this, if we consider cervical cancer as a whole (rather than making separate decisions for different stages), we get a sub-additive interaction for costs and a super-additive interaction for QALYs.

Table D-F1. Data inputs

No. women in population (millions)	20			
Cost of screening (per woman screened)	£100			
Prevalence of cervical cancer	2%			
QALYs if don't develop cancer	40			
Data inputs by cancer stage				
Cancer stage	I	II	III	IV
Probability diagnosed at this stage if not screened	50%	30%	15%	5%
Probability diagnosed at this stage if screened	79%	15.0%	5.0%	1.0%
QALYs if diagnosed at this stage: comparator treatment	20	10	5	1
QALYs if diagnosed at this stage: new treatment	30	15	7.5	1.5
Cost if diagnosed at this stage: comparator treatment	£5,000	£10,000	£15,000	£30,000
Cost if diagnosed at this stage: new treatment	£7,500	£15,000	£22,500	£45,000

Table D-F2. Outcomes and costs by stage for each treatment combination

Cost-effectiveness of treatment	**Cost per million in population**			**QALYs per million in population**			**ICER: New Tx vs. comparator**
	New Tx†	**Compa-rator†**	**Δ**	**New Tx‡**	**Compa-rator‡**	**Δ**	
Stage I - without screening	£1,500	£1,000	£500	6.00	4.00	2.00	£250
Stage II - without screening	£1,800	£1,200	£600	1.80	1.20	0.60	£1,000
Stage III - without screening	£1,350	£900	£450	0.45	0.30	0.15	£3,000
Stage IV - without screening	£900	£600	£300	0.03	0.02	0.01	£30,000
All stages combined - without screening, but including QALYs for patients without cervical cancer	£5,550	£3,700	£1,850	792.28	789.52	2.76	£670
Stage I - with screening	£2,370	£1,580	£790	9.48	6.32	3.16	£250
Stage II - with screening	£900	£600	£300	0.90	0.60	0.30	£1,000
Stage III - with screening	£450	£300	£150	0.15	0.10	0.05	£3,000
Stage IV - with screening	£180	£120	£60	0.01	0.00	0.00	£30,000
All stages combined - with screening and costs and QALYs for patients without cervical cancer	£5,900	£4,600	£1,300	794.54	791.02	3.51	£370

† Cost = population * prevalence of cervical cancer * proportion of patients diagnosed at this stage * cost if diagnosed at this stage. The all stages combined cost with screening includes the cost of screening (equal to population*cost of screening).
‡ QALYs = population * prevalence of cervical cancer * proportion of patients diagnosed at this stage * QALYs if diagnosed at this stage.

Table D-F3. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
No cervical screening and old cervical cancer treatment	£3,700,000,000	789,520,000	£15,786,700,000,000
Cervical screening and old cervical cancer treatment	£4,600,000,000	791,024,000	£15,815,880,000,000
No cervical screening and new cervical cancer treatment	£5,550,000,000	792,280,000	£15,840,050,000,000
Cervical screening and new cervical right cancer treatment	£5,900,000,000	794,536,000	£15,884,820,000,000
Interaction	-£550,000,000	752,000	£15,590,000,000

† Costs taken from Table D-F2.
‡ QALYs taken from Table D-F2
§ At a £20,000/QALY ceiling ratio.

Table D-F4. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	**Incr QALYs**	**ICER**
Cervical cancer screening (F1)	£625,000	1,880	£332
Cervical cancer treatment (F2)	£1,575,000	3,136	£502

Table D-F5. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Cervical screening and old cervical cancer treatment (F1) vs. F0	£900,000	1,504	Ex dominated
No cervical screening and new cervical cancer treatment (F2) vs. F0	£950,000	1,256	Ex dominated
Cervical screening and new cervical cancer treatment (F3) vs. F0	£2,200,000	5,016	£439

However, providing that screening doesn't affect the costs and QALYs of patients within each cancer stage, the cost-effectiveness of treating cervical cancer will be independent of screening if we make separate decisions about treatment for each cancer stage. Once we have made the decision about the best cancer treatment, we can then make a decision on the cost-effectiveness of screening that is contingent on patients receiving the most cost-effective treatment strategy.

Example G: No interaction

Since these interventions are given to different, non-overlapping patient populations with different diseases managed by different staff in different healthcare facilities, they are completely independent and have additive effects on costs and QALYs. It is therefore reasonable to make independent decisions on the two treatments.

Table D-G1. Data inputs

	Total cost/patient	Total QALYs/patient
Ovarian cancer treatment	£7,500	30
Ovarian cancer no treatment	£5,000	20
BPH treatment	£15,000	29
BPH no treatment	£500	28.5
Number of cases of ovarian cancer	5,000	
Number of cases of BPH	8,000	

Table D-G2. Outcomes for each treatment combination

	Population Cost†	Population QALYs‡	Population NB§
No treatment	£29,000,000	328,000	£6,531,000,000
Ovarian cancer treatment only	£41,500,000	378,000	£7,518,500,000
BPH treatment only	£145,000,000	332,000	£6,495,000,000
Ovarian cancer treatment + BPH treatment	£157,500,000	382,000	£7,482,500,000
Interaction	£0	0	£0

† Cost = Cost of ovarian cancer*number of cases of ovarian cancer + cost of BPH*number of BPH cases

‡ QALYs = QALYs with ovarian cancer*number of ovarian cancer cases + QALYs with BPH*number of BPH cases

§ At a £20,000/QALY ceiling ratio.

Table D-G3. Incremental outcomes per thousand population treating the options as independent

	Incr Cost	Incr QALYs	ICER
Ovarian cancer treatment (G1)	£12,500	50.00	£250
BPH treatment (G2)	£116,000	4.00	£29,000

Table D-G4. Incremental outcomes per thousand population treating the options as interdependent or mutually-exclusive

	Incr Cost	Incr QALYs	ICER
Ovarian cancer treatment only (G1) vs. G0	£12,500	50.00	£250
BPH treatment only (G2) vs. G1	£103,500	-46.00	Ex dominated
Ovarian cancer treatment + BPH treatment (G3) vs. G1	£116,000	4.00	£29,000

Example H: Interaction type 1 (direct pharmacological, behavioural or biological mechanisms)

Drug A is given orally six times a day at regular intervals, but cannot be given within an hour of administering Drug B. Drug B is given three times a day 30 minutes after meals as an injection. Giving both drugs together reduces compliance and therefore the cost and health gains by 10%.

Table D-H1. Data inputs

QALYs with no treatment	6
QALYs gained from A alone	0.8
QALYs gained from B alone	1.5
Compliance when treatments are given individually	95%
Compliance when treatments are given together	85%
Drug cost A over 10 years	£300
Drug cost B over 10 years	£5,000
Population	5,000

Table D-H2. Outcomes for each treatment combination

	Population Cost†	Population QALYs‡	Population NB§
No treatment	£0	30,000	£600,000,000
A only	£1,425,000	33,800	£674,575,000
B only	£23,750,000	37,125	£718,750,000
A+B	£22,525,000	39,775	£772,975,000
Interaction	-£2,650,000	-1,150	-£20,350,000

† Cost = population*compliance*drug cost
‡ QALYs = Population*(QALYs without treatment + compliance*QALY gain)
§ At a £20,000/QALY ceiling ratio.

Since poor compliance affects only the gain in QALYs and added cost for combination therapy, this type of interaction is not multiplicative and isn't removed by working on a logarithmic scale.

Example I: Interaction type 2c (multiplicative effects on cost)

Two interventions altering the way in which inpatient stays for hysterectomy managed both reduce length of stay by a certain percentage. Administering both interventions together therefore reduces length of stay (and therefore cost) by less than the sum of individual effects.

Table D-I1. Data inputs

Cost per bed day	£300
Length of stay (standard open procedure)	7
% reduction in length of stay from laparoscopic vs. open procedure	25%
% reduction in length of stay from early discharge scheme	30%
QALYs open hysterectomy	0.8
QALYs laparoscopic hysterectomy	0.83
QALY gain from early discharge	0.02
Incremental cost of primary care with early discharge scheme	£500
Population	500,000

Table D-I2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
Open hysterectomy without early discharge scheme	£1,050,000,000	400,000	£6,950,000,000
Laparoscopic hysterectomy without early discharge scheme	£787,500,000	415,000	£7,512,500,000
Open hysterectomy with early discharge scheme	£735,000,000	410,000	£7,465,000,000
Laparoscopic hysterectomy with early discharge scheme	£236,250,000	425,000	£8,263,750,000
Interaction	-£236,250,000	0	£236,250,000

† Cost = cost per bed day multiplied by length of stay multiplied by population size. The early discharge scheme also costs an additional £500*population size in addition to the length of stay.
‡ QALYs = QALYs per patient*population.
§ At a £20,000/QALY ceiling ratio.

Example J: Interaction type 4d (impact of future costs)

Future costs introduce an interaction between a smoking cessation intervention at age 65 and the introduction of unlimited free nursing care. Introducing free nursing care is assumed to increase the cost of care by £5,000 in every year of life. Patients who have a longer life expectancy will therefore accrue a larger cost from the free nursing care. Since the smoking cessation intervention increases life expectancy, those patients accrue higher costs in the years of life gained and (if these future costs are included in the analysis), the incremental cost of smoking cessation is higher with free nursing care than without. In this example, there is also a small super-additive interaction for QALYs due to the same effect observed in example B.

Table D-J1. Data inputs

Incremental cost per year of life of unlimited free nursing care to NHS	£5,000
Utility with unlimited free nursing care	0.75
Utility without unlimited free nursing care	0.7
Lifetime cost of smoking-related diseases - smoker	£2,000
Lifetime cost of smoking-related diseases - ex-smoker	£1,000
Cost of smoking cessation intervention	£50
Effectiveness of smoking intervention: % of smokers who stop	20%
Life expectancy smoker	18
Life expectancy ex-smoker	22
Population	3,000,000

Table D-J2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
No treatment	£6,000,000,000	37,800,000	£750,000,000,000
Smoking cessation intervention only	£5,550,000,000	39,480,000	£784,050,000,000
Free nursing care only	£270,000,000,000	40,500,000	£540,000,000,000
Smoking cessation and free nursing care	£287,550,000,000	42,300,000	£558,450,000,000
interaction	£18,000,000,000	120,000	-£15,600,000,000

† Cost = cost of smoking cessation intervention (if relevant) + annual cost of nursing care*weighted average life expectancy + weighted average cost of smoking-related diseases, all multiplied by population size. Life expectancy and the cost of smoking-related diseases were averaged over the proportion of people who quit smoking and the proportion who do not.
‡ QALYs = Population*relevant utility*(life expectancy for smokers*proportion not quitting + life expectancy for ex-smokers*proportion quitting).
§ At a £20,000/QALY ceiling ratio.

Example K: Interaction type 2d (between immediate mortality and extensions of life expectancy)

Measles mumps and rubella (MMR) vaccination reduces the risk of infant mortality but is (in this simplified example) assumed to have no major impact on subsequent HRQoL or life expectancy. However, the benefits of saving a life in childhood depend on the number of QALYs that will be accrued in the remaining life expectancy. These QALYs are assumed to be substantially increased by hypothetical interventions that completely prevent cardiovascular disease and therefore greatly increases life expectancy. There is therefore a large super-additive interaction for QALYs. In practice, such effects are likely to introduce interactions for cost due to the impact of future costs, although these are omitted from this example.

Table D-K1. Data inputs

Odds of mortality from MMR by age 5 without vaccination	2.00%
Odds ratio of mortality from MMR by age 5 with vaccination	0.05
Quality-adjusted life expectancy at age 5 given current treatment	60
Quality-adjusted life expectancy at age 5 with a new treatment that prevents all cardiovascular disease	90
Cost of MMR vaccine	£10
Lifetime net cost of treatment preventing cardiovascular disease (drug cost minus savings from disease prevented)	£200,000
Population (babies per year)	600,000

Table D-K2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
No treatment	£0	35,294,118	£705,882,352,941
MMR vaccine only	£6,000,000	35,964,036	£719,274,719,281
CVD prevention only	£120,000,000,000	52,941,176	£938,823,529,412
MMR + CVD prevention	£120,006,000,000	53,946,054	£958,915,078,921
interaction	£0	334,959.2	£6,699,183,170

† Cost = intervention cost multiplied by population size
‡ QALYs = (odds of MMR mortality/(1+ odds of MMR mortality))*quality-adjusted life expectancy*population.
§ At a £20,000/QALY ceiling ratio.

Within this example, it is essential to make a plausible assumption about the subsequent life expectancy when evaluating the cost-effectiveness of MMR. However, the only impact of the MMR vaccine on the cardiovascular intervention is to increase the number of patients who can benefit from it. An evaluation looking at patients who have already survived to adulthood would not be dependent on whether or not MMR is given.

Example L: Interaction type 4b (interactions between diseases)

Drug A decreases the risk of developing heart failure, but has no direct effect on the risk of stroke. Drug B decreases the risk of stroke, but has no direct effect on the risk of heart failure.

However, patients with heart failure are at increased risk of stroke and the two events have multiplicative effects on cost and HRQoL, such that the HRQoL impact of heart failure and stroke is less than that of the two diseases separately, while the additional cost of managing patients with both heart failure and stroke is greater than the effect of each disease separately.

The interaction between these two diseases was modelled using a Markov model, to estimate the number of patients in each of the five disease states (no disease, stroke, heart failure, stroke + heart failure and dead). 25 annual cycles were modelled with a half cycle correction. Development of stroke and heart failure were assumed to be permanent.

Table D-L1. Data inputs

Annual Odds of heart failure	0.03
Odds ratio for heart failure Drug A	0.4
Annual Odds of stroke (without heart failure)	0.05
Odds ratio for stroke Drug B	0.5
Odds ratio for stroke: in patients with heart failure vs. those without	2
Annual cost with neither stroke nor heart failure	1,000
Annual cost of stroke alone	£15,000
Annual cost of heart failure alone	£4,000
Annual cost of stroke and heart failure	£19,000
Utility with neither stroke nor heart failure	0.8
Utility with stroke alone	0.32
Utility with heart failure alone	0.56
Utility with stroke and heart failure	0.08
Annual odds of death without either heart failure or stroke	0.02
Odds ratio of death after developing heart failure	4
Odds ratio of death after stroke	7
Drug cost/year: A (heart failure)	£200
Drug cost/year: B (stroke)	£300
Population	500,000

Table D-L2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
No treatment	28,703,311,049	4,410,346	£59,503,618,166
Drug A (reduces risk of heart failure)	32,132,300,059	4,864,927	£65,166,240,290
Drug B (reduces risk of stroke)	25,237,151,910	5,300,440	£80,771,641,099
Drug A + Drug B	26,267,367,316	5,854,271	£90,818,057,439
interaction	-£2,398,773,604	99,251	£4,383,794,216

† Annual cost = sum product of the annual costs of each disease state and the number of patients each disease state. Drug A was assumed to be given to living patients who had not yet developed heart failure, while drug B was assumed to be given to living patients who had not yet had a stroke. Annual costs were summed over all years, with a half cycle correction applied to cycles zero and 25.
‡ QALYs = sum product of the utility of each disease state and the number of patients in each disease state. Annual QALYs were summed over all years, with a half cycle correction applied to cycles zero and 25.
§ At a £20,000/QALY ceiling ratio.

Example M: Interaction type 4c (Effect of comorbid conditions [knee replacement and obesity surgery] on treatment costs)

The cost of conducting knee replacement surgery is substantially higher for morbidly obese patients, since specialised beds, instruments, etc. are required. For the population of people with morbid obesity and knee osteoarthritis, there is therefore a mixed interaction for cost between knee replacement surgery and interventions for obesity (e.g. bariatric surgery) that are conducted before knee surgery.

Table D-M1. Data inputs

Cost of knee replacement: non-morbidly obese patients	£6,000
Cost of knee replacement: morbidly-obese patients	£10,000
Cost of obesity surgery	£8,000
Utility with morbid obesity and knee osteoarthritis	0.2
Utility improvement from obesity surgery	0.15
Utility improvement from knee replacement	0.3
Life expectancy: non-morbidly obese patients (years)	20
Life expectancy: morbidly obese patients (years)	20
Proportion of patients ceasing to be morbidly obese after obesity surgery	60%
Additional annual cost of managing morbidly obese patient compared with non-morbidly obese patient	£1,500
Population: number of patients with knee osteoarthritis and morbid obesity	20,000

Table D-M2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
No knee replacement and no obesity surgery	£600,000,000	80,000	£1,000,000,000
Obesity surgery only	£400,000,000	116,000	£1,920,000,000
Knee replacement only	£800,000,000	200,000	£3,200,000,000
Obesity surgery and knee replacement	£552,000,000	236,000	£4,168,000,000
interaction	-£48,000,000	0	£48,000,000

† Cost = cost of any of obesity surgery, plus the proportion of patients ceasing to be morbidly obese multiplied by the cost of knee replacement, plus the proportion of patients remaining morbidly obese multiplied by the cost of knee replacement and the lifetime cost of managing obesity (all multiplied by the population size).
‡ QALYs = proportion of patients ceasing to be morbidly obese multiplied by life expectancy for non-morbidly obese patients multiplied by the HRQoL with/without knee replacement, plus proportion of patients to be morbidly obese multiplied by life expectancy for morbidly obese patients multiplied by the HRQoL with/without knee replacement (all multiplied by the population size).
§ At a £20,000/QALY ceiling ratio.

Example N: Interaction type 5 (Interactions arising from shared capital expenditure)

This hospital currently has no MRI machine. They could introduce facilities for MRI scans of patients with head injuries and/or they could introduce facilities for MRI scans of patients with suspected cancer. Within this simplified example, the MRI machine has capacity for at least 5,000 patients and would not be used for any other patients, nor would patients be referred elsewhere. However, similar principles will apply even if these simplifications were relaxed.

Because the machine can be used for both services, but only needs to be purchased once, the interaction for cost is exactly equal to the cost of the MRI machine. By contrast, in this example there is no interaction for QALYs.

Table D-N1. Data inputs

Cost of purchasing MRI machine	£500,000
Variable costs per scan	£50
QALYs with head injury without MRI scan	30
QALYs with suspected cancer without MRI scan	15
QALY gained from MRI scan for head injuries	1
QALY gained from MRI scan for suspected cancer	0.5
Incremental cost of scanning for head injuries excluding cost of scan	-£50
Incremental cost of scanning for suspected cancer excluding cost of scan	£300
Number of scans for head injuries per MRI machine	3,000
Number of scans for cancer per MRI machine	2,000

Table D-N2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**	**Population NB§**
No MRI	£0	120,000	£2,400,000,000
MRI for head injuries only	£500,000	123,000	£2,459,500,000
MRI for cancer only	£1,200,000	120,008	£2,398,950,000
MRI for both head injuries and cancer	£1,200,000	123,008	£2,458,950,000
interaction	-£500,000	0	£500,000

† Cost = cost of MRI machine (if relevant) + number of scans for head injuries*(variable cost per MRI scan + incremental cost of scanning for head injuries) + number of scans for cancer*(variable cost per MRI scan + incremental cost of scanning for cancer).
‡ QALYs = Number of head injuries*number of QALYs accrued with head injuries + number of suspected cancer cases*number of QALYs for patients with suspected cancer
§ At a £20,000/QALY ceiling ratio.

Example O: Interaction type 5 (Interactions arising from shared ward facilities)

A hospital needs to decide whether the offer hip replacement and whether to conduct knee arthroscopy as a day case (one day hospital stay) or with a three-day hospital stay. This hypothetical ward is only used for these two types of patients.

Table D-O1. Data inputs

Total fixed cost of running ward per year excluding nursing staff and variable costs	£100,000
Healthcare assistant cost per 1095 bed days/year (1:3 ratio with 3 shifts/day and annual cost per HCA of £23,000)	£69,000
Qualified nurse cost per 2190 bed days/year (1:6 ratio with 3 shifts/day and annual cost per nurse of £30,000)	£90,000
Variable cost per patient on ward	£150
Number of bed days available per year	7300
QALYs without hip replacement	11
QALYs with hip replacement	22
QALYs with day case knee arthroscopy	11
QALYs with knee arthroscopy with inpatient stay	11
Number of patients with hip replacement	950
Number of patients with knee arthroscopy	500
Length of stay for hip replacement	6
Length of stay for arthroscopy if not done as a day case	3

Table D-O2. Outcomes for each treatment combination

	Bed days used/year*	Total cost/ bed day¥	Population Cost†	Population QALYs‡	Population NB§
No hip replacement; arthroscopy with inpatient stay	1500	£369	£553,000	15,950	£318,447,000
Hip replacement; arthroscopy with inpatient stay	7200	£281	£2,023,000	26,400	£525,977,000
No hip replacement; day case arthroscopy	500	£668	£334,000	15,950	£318,666,000
Hip replacement; day case arthroscopy	6200	£276	£1,714,000	26,400	£526,286,000
interaction	0	-£304	-£90,000	0	£90,000

* Bed days used/year = number of hip replacements*length of stay for hip replacement + number of knee arthroscopies*length of stay for knee arthroscopy (where this is 1 for a day case)
¥ Total cost/bed day = variable cost per patient + (fixed cost/ward + Healthcare assistant cost*(bed days÷ 1095, rounded up) + qualified nurse cost*(bed days÷ 2190, rounded up)
† Cost = cost/bed day*bed days used.
‡ QALYs = Number of hip replacement candidates*QALYs accrued + number of arthroscopy candidates*QALYs accrued
§ At a £20,000/QALY ceiling ratio.

The impact of non-compliance within the healthcare facility would be similar to that shown in Example H.

Modelling the impact of complex interventions and changes to service delivery could be achieved using similar methods to those used to evaluate complex interventions in general. However, as with or complex interventions, the methods will be substantially more complicated than the examples shown here.

Example P: Interaction type 3 (ceiling effects for HRQoL introducing diminishing marginal returns for QALYs)

Treatments A and B are evaluated in a 10-year trial. The treatments have no effect on length of life, but in the absence of any ceiling effect, treatment A would increase utility by 0.15 and B would increase utility by 0.2, with no interaction. The treatments have additive effects on cost. However, since utility cannot be increased above 1, patients cannot accrue more than 10 QALYs in this 10-year trial (no discounting is applied for simplicity, although if it were applied it would simply decrease QALYs by the same proportion in all patients).

In this example, the ceiling effect does not come into play for the average patient, since the simple effects would suggest that the mean utility for the average patient receiving treatment *ab* is 0.95 (0.6+0.15+0.2). However, when we take account of the distribution of QALYs by randomly sampling from normal distributions with SD equalling 0.2, the proportion of patients accruing 10 QALYs (the maximum) is substantially higher for the *ab* group, introducing a large sub-additive interaction for QALYs. Although the interaction for cost (and the interaction for QALYs in the absence of the ceiling effect) would be zero if infinite numbers of patients were sampled, the use of Monte Carlo simulation means that there is a small interaction for cost within this sample of 500 patients.

Table D-P1. Data inputs

Time horizon of the study (years)	10
Maximum utility	1
Mean utility no treatment	0.6
Increase in utility from treatment A	0.15
Increase in utility from treatment B	0.2
SD around utility	0.2
Cost of A	£2,000
Cost of B	£3,000
Cost without treatment	£1,000
SD around cost without treatment	£300
Population	10,000

Table D-P2. Outcomes for each treatment combination

	Population Cost†	**Population QALYs‡**			**Population NB§**
		If there were no ceiling effect	**% patients with 10 QALYs**	**With ceiling effect**	**With ceiling effect**
No treatment	£9,855,731	59,837	3%	59,646	£1,183,070,245
A only	£30,050,502	75,348	9%	74,395	£1,457,858,878
B only	£40,022,721	79,724	15%	78,179	£1,523,555,682
A and B	£60,093,652	95,507	40%	89,603	£1,731,971,226
interaction	-£123,839	273	19%	-3,325	-£66,373,089

† Cost was estimated for each of 500 simulated patients by sampling values from a gamma distribution representing cost without treatment with mean £1,000 and SD £300 (alpha = 11.11, beta = 90). Treatment costs were added to the cost without treatment and were assumed to be fixed. Population cost was calculated by averaging across the 500 trial participants and multiplying the average by the population size.
‡ QALYs without a ceiling effect were calculated for each of 500 simulated patients by sampling values from normal distribution with a mean equal to the group mean for that treatment arm and a SD of 0.2. Population QALYs were calculated by averaging over the 500 trial participants and multiplying the average by the population size.
§ At a £20,000/QALY ceiling ratio.

Similar trends would be seen for life-extending treatments within a trial of fixed duration, even if the effect of treatment on mortality were not additive.

Example Q: Interaction type 2e (non-multiplicative scale effects: in this case between odds, risk and hazard)

As discussed, interactions can also arise from scale effects other than multiplicative effects. One example of this is the difference between treatments being additive on a log-odds scale, additive on a log-risk scale and additive on a log-hazard scale. This example is not translated into costs and QALYs, since costs and QALYs need to be interpreted on a natural scale in which all of these examples will have large interactions arising from the multiplicative effect, which swamp any effect of differences between odds, risk and hazard.

In each case, the interaction on a natural scale is calculated as $\mu_0 + \mu_{ab} - \mu_a - \mu_b$, while the interaction on a logarithmic scale is calculated as $\ln(\mu_0) + \ln(\mu_{ab}) - \ln(\mu_a) - \ln(\mu_b)$.

For treatments that have no interaction on a log-odds scale, there will be an interaction on a log-risk scale and a log-hazard scale, as well as interactions on the natural scale of odds, risks and hazards.

Table D-Q1. Outcomes for an example where treatments have multiplicative effects on odds. In this example, the odds of events with no treatment are 0.1, the odds ratio for A vs. not-A is 0.5 and the odds ratio for B vs. not-B is 0.3.

	Odds*	Risk†	Hazard‡
0	0.100	0.091	0.095
a	0.050	0.048	0.049
b	0.030	0.029	0.030
ab	0.015	0.015	0.015
Interaction on a log scale	0.000	-0.032	-0.016
Interaction on a natural scale	0.035	0.029	0.032

* Defined by the data inputs
† Risk = odds/(1+ odds)
‡ Hazard or rate = -Ln(1-risk)

For treatments that have no interaction on a log-risk scale, there will be an interaction on a log-odds scale and a log-hazard scale, as well as interactions on the natural scale of odds, risks and hazards.

Table D-Q2. Outcomes for an example where treatments have multiplicative effects on risk. In this example, the probability of events with no treatment is 0.1, the relative risk for A vs. not-A is 0.5 and the relative risk for B vs. not-B is 0.3.

	Odds*	Risk†	Hazard‡
0	0.111	0.100	0.105
A	0.053	0.050	0.051
B	0.031	0.030	0.030
Ab	0.015	0.015	0.015
Interaction on a log scale	0.039	0.000	0.019
Interaction on a natural scale	0.043	0.035	0.039

* Odds = risk/(1-risk)
† Defined by the data inputs
‡ Hazard or rate = -Ln(1-risk)

For treatments that have no interaction on a log-hazard scale, there will be an interaction on a log-odds scale and a log-risk scale, as well as interactions on the natural scale of odds, risks and hazards.

Table D-Q3. Outcomes for an example where treatments have multiplicative effects on hazard, or rate. In this example, the hazard or rate of events with no treatment is 0.1, the hazard ratio or relative rate for A vs. not-A is 0.5 and the hazard ratio or relative rate for B vs. not-B is 0.3.

	Odds*	Risk†	Hazard‡
0	0.105	0.095	0.100
a	0.051	0.049	0.050
b	0.030	0.030	0.030
ab	0.015	0.015	0.015
Interaction on a log scale	0.018	-0.017	0.000
Interaction on a natural scale	0.039	0.032	0.035

* Odds = risk/(1-risk)
† Risk = $1 - e^{-rate}$
‡ Defined by the data inputs

Appendix E: Proof that joint decisions inside-the-table give efficient treatment-adoption decisions

This appendix proves that inside-the-table analysis and making a joint decision about interacting technologies always gives the right conclusion about which treatment maximises NMB, whereas at-the-margins analysis will lead to adoption of a suboptimal treatment whenever there are qualitative interactions for NMB at the ceiling ratio of interest.

The proof begins from the following starting points:

A. The aim of economic evaluation is to allocate decisions to maximise the health gained from the budget.
B. If we are only interested in the current treatment adoption decision, inference is irrelevant and we achieve aim (A) by adopting the single treatment with highest expected NMB.
C. Inside-the-table analysis or regression with interaction terms gives unbiased estimates of the effect of treatment A when given alone: i.e. $\hat{\delta}_{a-0} = \mu_a - \mu_0$ is an unbiased and consistent estimate of δ_{a-0} and $\hat{\delta}_{a-0}$ will approach δ_{a-0} as the sample size approaches infinity.

Logical proof

Analyses allowing for interactions give unbiased estimates of the difference between *0* and *a* (3, 4) and, therefore, unbiased estimates of the expected NMB for each treatment. By implication, this means that allowing for interactions and considering the four treatment combinations as mutually-exclusive alternatives must give the correct ranking of treatment combinations by NMB and therefore correctly identify the treatment with highest NMB.

By contrast, excluding interactions will introduce bias into NMB estimates whenever treatment effects are not genuinely additive (3, 4, 8, 20, 30). While quantitative interactions in NMB will change the magnitude of differences between treatments, qualitative interactions will change the ranking of treatments. Furthermore, qualitative interactions that change which treatment has highest expected NMB will change the conclusions of our economic evaluation and ignoring such interactions will cause us to adopt a treatment that does not actually have highest NMB. If there are no qualitative interactions for NMB, this approach will give the same conclusions as allowing for interactions and making a joint decision between the four mutually-exclusive alternatives. However, if there are qualitative interactions that change which treatment maximises NMB, this approach will lead to adoption of a treatment that doesn't have highest NMB.

Mathematical and graphical proofs

Maximising NMB and health gains from the budget require us to adopt a treatment that lies on the cost-effectiveness frontier, but which has an ICER relative to its next best non-dominated alternative that is below our ceiling ratio. For a 2x2 factorial trial, if we make a joint decision about treatments A and B, we will treat each of the cells as a separate decision option and can calculate six different ICERs:

$$\mathrm{ICER}_{\text{a-0}} = \left(\mathrm{Cost}_{\text{a}} - \mathrm{Cost}_{0}\right) \div \left(QALYs_a - QALYs_0\right) \qquad INB_{a\text{-}0} = NMB_a - NMB_0 \quad \text{(E1a)}$$

$$\mathrm{ICER}_{\text{b-0}} = \left(\mathrm{Cost}_{\text{b}} - \mathrm{Cost}_{0}\right) \div \left(QALYs_b - QALYs_0\right) \qquad INB_{b\text{-}0} = NMB_b - NMB_0 \quad \text{(E1b)}$$

$$\mathrm{ICER}_{\text{a-b}} = \left(\mathrm{Cost}_{\text{a}} - \mathrm{Cost}_{\text{b}}\right) \div \left(QALYs_a - QALYs_b\right) \qquad INB_{a\text{-b}} = NMB_a - NMB_{\text{b}} \quad \text{(E1c)}$$

$$\text{ICER}_{\text{ab-a}} = (\text{Cost}_{\text{ab}} - \text{Cost}_{\text{a}}) \div (QALYs_{ab} - QALYs_{a}) \qquad INB_{ab\text{-a}} = NMB_{ab} - NMB_{\text{a}} \quad \text{(E1d)}$$

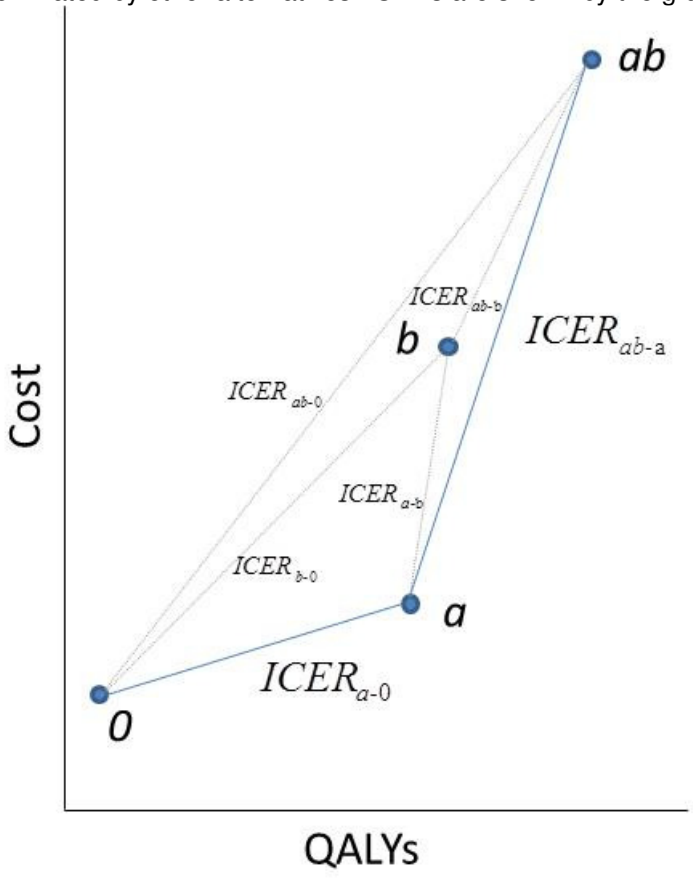
$$\text{ICER}_{\text{ab-b}} = (\text{Cost}_{\text{ab}} - \text{Cost}_{\text{b}}) \div (QALYs_{ab} - QALYs_{b}) \qquad INB_{ab\text{-b}} = NMB_{ab} - NMB_{\text{b}} \quad \text{(E1e)}$$

$$\text{ICER}_{\text{ab-0}} = (\text{Cost}_{\text{ab}} - \text{Cost}_{0}) \div (QALYs_{ab} - QALYs_{0}) \qquad INB_{ab\text{-0}} = NMB_{ab} - NMB_{0} \quad \text{(E1f)}$$

In any given study, up to three of these six ICERs will lie on the cost-effectiveness frontier and the others will be irrelevant due to dominance or not comparing treatments against the next most effective treatment lying on the frontier. For example, in Figure E1, the cost-effectiveness frontier runs from *0* to *a* to *ab*, with *b* being extendedly dominated by a combination of *a* and *ab*. The relevant ICERs are therefore $ICER_{a\text{-}0}$ and $ICER_{ab\text{-a}}$. We will adopt the treatment that has highest NMB, which (in this example) will equal:

- *0* if $ICER_{a\text{-}0} > \text{Rc}$
- *a* if $ICER_{a\text{-}0} < \text{Rc}$ and $ICER_{ab\text{-a}} > \text{Rc}$
- *ab* if $ICER_{ab\text{-a}} < \text{Rc}$

Figure E1. Cost-effectiveness plane showing the four treatment options in the situation where there is a super-additive interaction for cost and a sub-interaction for QALYs and conclusions are drawn inside-the-table, allowing for interactions and considering the four cells as mutually-exclusive alternatives. The blue line shows the cost-effectiveness frontier, which joins the treatments that are not dominated by other alternatives. ICERs are shown by the gradient of lines joining two treatments.



By contrast, if we make independent decisions about treatments A and B or use at-the-margins analysis or regression without an interaction term, we implicitly assume that the increase in cost from adding A is the same regardless of whether B is also given. We therefore consider two separate decision problems, each containing two decision options (A

and not-A within the decision for factor A and B and not-B within the decision for factor B). We calculate the INB and ICER for A vs. not-A as:

$$\mathrm{ICER}_{\mathrm{AvsnotA}} = \frac{0.5\left(\mathrm{Cost}_{\mathrm{ab}} + \mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{b}} - \mathrm{Cost}_{0}\right)}{0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right)} \quad \text{(E2a)}$$

$$\mathrm{INB}_{\mathrm{AvsnotA}} = 0.5\left(\mathrm{NMB}_{\mathrm{ab}} + \mathrm{NMB}_{\mathrm{a}} - \mathrm{NMB}_{\mathrm{b}} - \mathrm{NMB}_{0}\right) \quad \text{(E2b)}$$

This implicitly assumes that:

$$\begin{aligned}\mathrm{ICER}_{\mathrm{AvsnotA}} &= \left(\mathrm{Cost}_{\mathrm{ab}} - \mathrm{Cost}_{\mathrm{b}}\right) \div \left(\mathrm{QALYs}_{\mathrm{ab}} - \mathrm{QALYs}_{\mathrm{b}}\right) \\ &= \left(\mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{0}\right) \div \left(\mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{0}\right) \\ &= \frac{0.5\left(\mathrm{Cost}_{\mathrm{ab}} + \mathrm{Cost}_{\mathrm{a}}\right) - 0.5\left(\mathrm{Cost}_{\mathrm{b}} + \mathrm{Cost}_{0}\right)}{0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}}\right) - 0.5\left(\mathrm{QALYs}_{\mathrm{b}} + \mathrm{QALYs}_{0}\right)} \\ &= \frac{0.5\left(\mathrm{Cost}_{\mathrm{ab}} + \mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{b}} - \mathrm{Cost}_{0}\right)}{0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right)}\end{aligned} \quad \text{(E3)}$$

and that:

$$\begin{aligned}\mathrm{INB}_{\mathrm{AvsnotA}} &= \mathrm{NMB}_{\mathrm{ab}} - \mathrm{NMB}_{\mathrm{b}} = \mathrm{NMB}_{\mathrm{a}} - \mathrm{NMB}_{0} \\ &= 0.5\left(\mathrm{NMB}_{\mathrm{ab}} + \mathrm{NMB}_{\mathrm{a}} - \mathrm{NMB}_{\mathrm{b}} - \mathrm{NMB}_{0}\right)\end{aligned} \quad \text{(E4)}$$

The assumptions within equations (E3) and (E4) are true only if there is no interaction: i.e. if $\mathrm{NMB}_{0} - \mathrm{NMB}_{\mathrm{a}} - \mathrm{NMB}_{\mathrm{b}} + \mathrm{NMB}_{\mathrm{ab}} = 0$.

Graphically, at-the-margins analysis estimates one single ICER for A vs. not-A (the blue line in Figure E2), which joins a point halfway between *0* and *b* (i.e. the coordinates $\left(0.5\left(\mathrm{Cost}_{\mathrm{b}} + \mathrm{Cost}_{0}\right),\ 0.5\left(\mathrm{QALYs}_{\mathrm{b}} + \mathrm{QALYs}_{0}\right)\right)$), with a point halfway between *a* and *ab* (i.e. the coordinates $\left(0.5\left(\mathrm{Cost}_{\mathrm{ab}} + \mathrm{Cost}_{\mathrm{a}}\right),\ 0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}}\right)\right)$). The gradient for this single ICER line will be somewhere between the gradient of the line joining *a* and *0* and the gradient of the line joining *b* and *ab* (although there is no simple relationship between these ICERs as the ratios are not transitive).

Instead of identifying the cost-effectiveness frontier and adopting the most effective treatment with an ICER below Rc, within at-the-margins analysis, we will simply adopt A if $\mathrm{INB}_{\mathrm{AvsnotA}} > 0$ (and otherwise adopt not-A). Alternatively, in terms of the ICER:

$$\begin{aligned}&\text{if } 0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right) > 0 \text{: Adopt A if } \mathrm{ICER}_{\mathrm{AvsnotA}} < Rc \\ &\text{if } 0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right) < 0 \text{: Adopt A if } \mathrm{ICER}_{\mathrm{AvsnotA}} > Rc\end{aligned} \quad \text{(E5)}$$

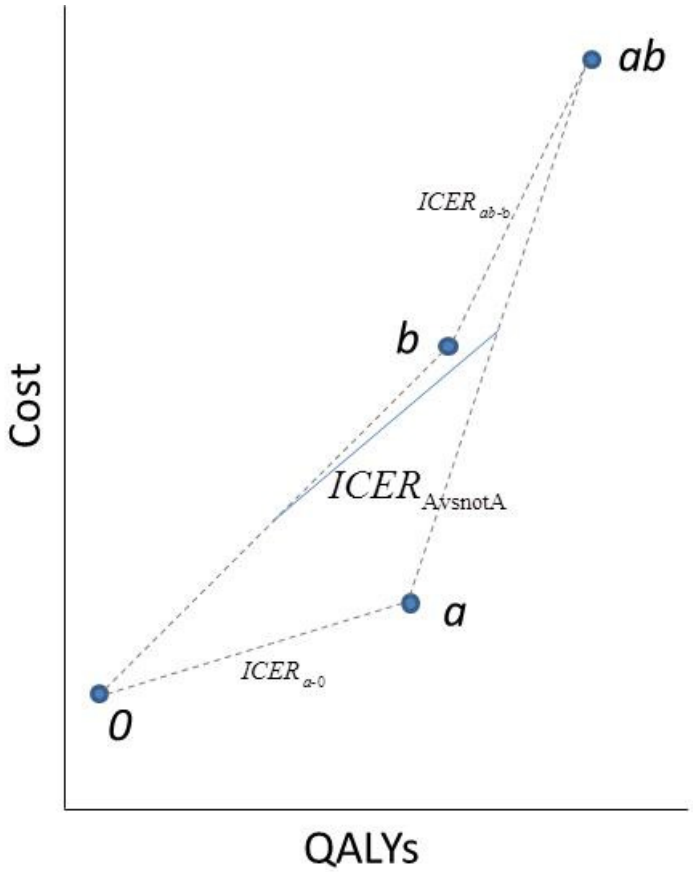
At-the-margins analysis and inside-the-table analysis (or regression without an interaction term and regression with an interaction term) will give the same results if A and B have perfectly additive effects (i.e. zero interaction) for both costs and QALYs:

$$\begin{aligned}&\text{if } \mathrm{QALYs}_{0} - \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} + \mathrm{QALYs}_{\mathrm{ab}} = 0 \\ &\text{and } \mathrm{Cost}_{0} - \mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{b}} + \mathrm{Cost}_{\mathrm{ab}} = 0\end{aligned} \quad \text{(E6)}$$

Zero interaction for costs and QALYs translates into zero interaction for NMB, since:

$$\begin{aligned}&\mathrm{NMB}_{0} - \mathrm{NMB}_{\mathrm{a}} - \mathrm{NMB}_{\mathrm{b}} + \mathrm{NMB}_{\mathrm{ab}} = \left(\mathrm{Rc} \cdot \mathrm{QALYs}_{0} - \mathrm{Cost}_{0}\right) \\ &- \left(\mathrm{Rc} \cdot \mathrm{QALYs}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{a}}\right) - \left(\mathrm{Rc} \cdot \mathrm{QALYs}_{\mathrm{b}} - \mathrm{Cost}_{\mathrm{b}}\right) + \left(\mathrm{Rc} \cdot \mathrm{QALYs}_{\mathrm{ab}} - \mathrm{Cost}_{\mathrm{ab}}\right) \\ &= \mathrm{Rc} \cdot \left(\mathrm{QALYs}_{0} - \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} + \mathrm{QALYs}_{\mathrm{ab}}\right) - \left(\mathrm{Cost}_{0} - \mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{b}} + \mathrm{Cost}_{\mathrm{ab}}\right)\end{aligned} \quad \text{(E7)}$$

Figure E2. Cost-effectiveness plane showing the four treatment options in the situation where there is no interaction for either costs or QALYs, but where results are analysed at-the-margins, ignoring interactions. The blue line shows the gradient of $\mathrm{ICER}_{\mathrm{AvsnotA}}$, which joins a point halfway between *0* and *b* with a point halfway between *a* and *ab*.



Furthermore, given no interaction, the at-the-margins ICER for A vs. not-A ($\mathrm{ICER}_{\mathrm{AvsnotA}}$) equals $\mathrm{ICER}_{\mathrm{a\text{-}0}}$ and $\mathrm{ICER}_{\mathrm{ab\text{-}b}}$:

if $y_0 - y_a - y_b + \mathrm{y}_{\mathrm{ab}} = 0$, then $\mathrm{y}_{\mathrm{ab}} = y_a + y_b - y_0$ (E8)

For at-the-margins analysis, substituting (E8) within equation (E2a) (and replacing *y* with Cost and QALYs):

$$\mathrm{ICER}_{\mathrm{AvsnotA}} = \frac{0.5\left(\mathrm{Cost}_{\mathrm{ab}} + \mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{b}} - \mathrm{Cost}_{0}\right)}{0.5\left(\mathrm{QALYs}_{\mathrm{ab}} + \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right)} \quad \text{(E2a)}$$

$$= \frac{0.5\left(\left[\mathrm{Cost}_{\mathrm{a}} + \mathrm{Cost}_{\mathrm{b}} - \mathrm{Cost}_{0}\right] + \mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{\mathrm{b}} - \mathrm{Cost}_{0}\right)}{0.5\left(\left[\mathrm{QALYs}_{\mathrm{a}} + \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right] + \mathrm{QALYs}_{\mathrm{a}} - \mathrm{QALYs}_{\mathrm{b}} - \mathrm{QALYs}_{0}\right)} \quad \text{(E9a)}$$

$$= \frac{0.5\left(2\mathrm{Cost}_{\mathrm{a}} - 2\mathrm{Cost}_{0}\right)}{0.5\left(2\mathrm{QALYs}_{\mathrm{a}} - 2\mathrm{QALYs}_{0}\right)} \quad \text{(E9b)}$$

$$= \left(\mathrm{Cost}_{\mathrm{a}} - \mathrm{Cost}_{0}\right) \div \left(QALYs_a - QALYs_0\right) = \mathrm{ICER}_{\mathrm{a\text{-}0}} \quad \text{(E1a)}$$

For inside-the-table analysis, substituting (E8) within equation (E1e):

$$\mathrm{ICER}_{\mathrm{ab\text{-}b}} = \left(\mathrm{Cost}_{\mathrm{ab}} - \mathrm{Cost}_{\mathrm{b}}\right) \div \left(QALYs_{ab} - QALYs_b\right) \quad \text{(E1e)}$$

$$= \left(\left[\text{Cost}_a + \text{Cost}_b - \text{Cost}_0\right] - \text{Cost}_b\right) \div \left(\left[\text{QALYs}_a + \text{QALYs}_b - \text{QALYs}_0\right] - QALYs_b\right) \quad \text{(E10)}$$

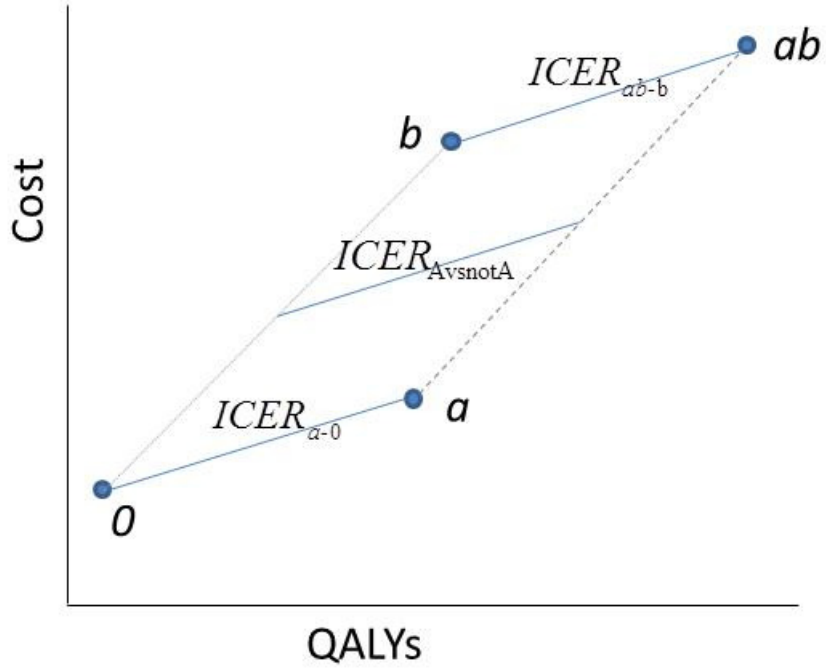
$$= \left(\text{Cost}_a - \text{Cost}_0\right) \div \left(QALYs_a - QALYs_0\right) = \text{ICER}_{a\text{-}0} \quad \text{(E1a)}$$

Therefore:

$$\text{ICER}_{\text{AvsnotA}} = \text{ICER}_{a\text{-}0} = \text{ICER}_{ab\text{-}b} \quad \text{(E11)}$$

Graphically, if there is no interaction the gradient of the blue at-the-margins ICER is equal to the gradient of the line joining the midpoint between *a* and *0* and the midpoint between *ab* and *b*:

Figure E3. Cost-effectiveness plane showing the four treatment options in the situation where there is no interaction for either costs or QALYs. The blue lines show the three ICERs ($\text{ICER}_{\text{AvsnotA}}$, $\text{ICER}_{a\text{-}0}$ and $\text{ICER}_{ab\text{-}b}$) that are equal when there is zero interaction for NMB.



Appendix F: Protocol for the systematic review of economic evaluations of factorial trials

Aims

- To identify all examples of full economic evaluations conducted on patient-level data from factorial design RCTs that were published before 31st December 2009.
- To review the characteristics, study design and statistical methods used in these studies, identify any issues raised by their conduct or results, assess the strengths and weaknesses of each approach and identify data suitable for simulations.

Inclusion criteria

- Described the methods and/or results of a full economic evaluation quantifying the costs *and* benefits of interventions designed to improve health or affect healthcare systems. CEA, CUA, CCA and CBA were classed as full economic evaluations.
    - CMAs and costing analyses were excluded as they consider only one dimension of outcome (i.e. costs but not benefits) and therefore do not raise the same issues as full economic evaluations that consider both costs and outcomes.
    - Although the primary focus was on analyses estimating ICERs, CBA of factorial design studies were included (although it was considered highly unlikely that any such studies would be identified).
    - CCA were included since some such studies do calculate ICERs and therefore raise similar issues to CEAs and CUAs.
- Used patient-level data from a factorial RCT.
    - Although studies with randomisation at the individual level were only included if they used patient-level data (as opposed to model-based economic evaluations based on secondary data from factorial trials), economic evaluations on cluster-randomised trials were included if they used data either at level of each cluster (e.g. for each GP surgery) or at the level of individual patients.
    - Based on the conceptual literature review described in Chapter 2, factorial RCTs were defined as studies in which two or more factors (each with two or more levels) are deliberately varied at the same time, with different groups of subjects being randomly allocated to receive different combinations of levels for each factor in such a way that interactions between two or more factors can be evaluated. Three-arm trials (e.g. studies comparing 0 vs. A vs. B, or A vs. B vs. A+B) were excluded from the systematic review since they do not permit evaluation of interactions. However, incomplete or fractional factorial design studies of larger or higher-order designs and partial factorial designs were included. Studies in which the interaction between factors was, to some extent, confounded by treating the group assigned to no treatment or combination therapy differently were included only if the authors described their study as factorial (Section 4.3).
- Published at least brief details of the methods and/or results of the factorial trial-based economic evaluation before 31st December 2009.
    - Full protocols of studies for which results were not yet available but for which a full economic evaluation was planned were included in the review if they were published in a journal article. Protocols available only on the NIHR HTA website were excluded. Protocols that mentioned collection of costs as well as clinical outcomes but did not explicitly discuss assessment of cost-effectiveness were excluded.
    - Papers that described the clinical results of a factorial study and stated that a full economic evaluation was planned were identified but were not included in the count of studies meeting inclusion criteria unless the economic paper was

subsequently identified. Basic details on these studies were extracted into a separate table to aid identification of the full economic evaluation. Additional targeted literature searches were conducted to find the economic evaluations to which these clinical papers referred (see below). However, no specific searches were conducted to identify clinical papers of this type and papers describing studies with economic evaluations described elsewhere were only identified if they were picked up by the main literature searches AND if the paragraph(s) that included the term "cost-effect*" or "economic" explicitly mentioned that an economic evaluation was planned (as opposed to simply making general statements that "cost-effectiveness of X should be assessed" or "X may be cost-effective"). Clinical trials identified in searches were also flagged and economic papers were sought if the reviewer (HD) already knew that an economic evaluation had been done.

Exclusion criteria

- Studies were not excluded from the review on the basis of language, providing that at least some text (e.g. an abstract) was available in English.
- Since the search strategy focussed on databases of studies relevant to medicine and the focus of this thesis is on economic evaluation of healthcare interventions, studies unrelated to health or healthcare were excluded.

Search strategy

Table F1: Search strategy for the conceptual review on issues and methods for factorial design trials

Database	Search terms for study design				Search terms for economic evaluation
○ MEDLINE (Ovid) ○ EMBASE (Ovid) ○ Econlit (Ovid) ○ Journals@Ovid full textx ○ Bmj.com (full text of all Highwire journals) ○ Sciencedirect.com	○ Factorial			AND	cost-effect* OR economic [in either title or abstract (or key words if available)]
	○ "2x2" ○ "2 x 2" ○ "2 by 2" ○ "two x two" ○ "two by two" ○ "2 x 3" ○ "3 x 3" ○ "3 x 2" ○ "2 x 2 x 2"				
	○ Matrix	AND	Randomised OR randomized		
○ CRD (NHS EED and HTA only) ○ Tufts CEA registry (https://research.tufts-nemc.org/cear/Default.aspx) ○ NIHR HTA publications list (http://www.hta.ac.uk/)	○ Factorial ○ "2x2" ○ "2 x 2" ○ "2 by 2" ○ "two x two" ○ "two by two" ○ "2 x 3" ○ "3 x 3" ○ "3 x 2" ○ "2 x 2 x 2" ○ Interaction [not used for searches of CRD due to excessive number of irrelevant hits]			-	-

In cases where the primary literature searches identified a clinical paper that explicitly mentioned in the abstract (or in the paragraph(s) of the paper that used the words "cost-

effect*" or "economic") that a full economic evaluation was planned, additional searches were conducted to identify the paper that described the economic evaluation. In the first instance, I searched Medline using EITHER the study name/acronym, OR the generic name of one of the treatments used, the name of the disease and "cost" as search terms. If this did not identify the study, searches were repeated without the name of the disease and with the word "economic" used in place of "cost" and the studies citing the clinical paper were then reviewed. If the study was not identified on this basis, it was assumed that the economic evaluation was not yet published or that the authors had decided not to conduct an economic evaluation and the study was excluded from the review.

Any studies identified in the conceptual review on factorial studies (Appendix A) or while reviewing the core references relating to issues raised by economic evaluation of factorial design studies that met the inclusion criteria for this review were also included.

The full text of studies identified in electronic searches that did not include an abstract was only examined in cases where the available details (e.g. title, authors and/or journal) suggested that the article was likely to meet the inclusion criteria. In particular, full text was not sought for articles with titles that made reference to a clinical trial but not to any economic analysis unless the abstract, authors or journal suggested that an economic evaluation was indeed conducted.

Data extraction

Studies were identified and extracted by one reviewer (HD).

Results of searches

Website	Databases searched	Terms searched	Hits
Ovid	Journals@Ovid, Econlit, EMBASE, Ovid MEDLINE(R), Ovid MEDLINE(R) Daily Update, Ovid OLDMEDLINE(R)...	factorial.af. and cost-effect*.ab.	296
		factorial.af. and cost-effect*.ti.	61
		factorial.af. and economic evaluation*.ti.	10
		factorial.af. and economic evaluation.ab.	19
		"2x2".af. and cost-effect*.ab.	41
		"2x2".af. and cost-effect*.ti.	9
		"2x2".af. and economic evaluation*.ti.	1
		"2x2".af. and economic evaluation*.ab	1
		"2 x 2".af. and cost-effect*.ab.	51
		"2 x 2".af. and cost-effect*.ti	17
		"2 x 2".af. and economic evaluation*.ab	4
		"2 x 2".af. and economic evaluation*.ti	4
		"2 by 2".af. and cost-effect*.ab.	382
		"2 by 2".af. and cost-effect*.ti	95
		"2 by 2".af. and economic evaluation*.ab	24
		"2 by 2".af. and economic evaluation*.ti	16
		"two x two".af. and cost-effect*.ab.	0
		"two x two".af. and cost-effect*.ti	0
		"two x two".af. and economic evaluation*.ab	0
		"two x two".af. and economic evaluation*.ti	0
		"two by two".af. and cost-effect*.ab. and trial.af	33
		"two by two".af. and cost-effect*.ti. and trial.af	9
		"two by two".af. and economic evaluation*.ab. and trial.af	2
		"two by two".af. and economic evaluation*.ti. and trial.af	0
		"two by two".af. and cost-effect*.ab.	46
		"two by two".af. and cost-effect*.ti.	11
		"two by two".af. and economic evaluation*.ab.	7

Website	Databases searched	Terms searched		Hits
		"two by two".af. and economic evaluation*.ti.		1
		"2 x 3".af. and cost-effect*.ab. and trial.af		1
		"2 x 3".af. and cost-effect*.ti. and trial.af		0
		"2 x 3".af. and economic evaluation.ab. and trial.af		0
		"2 x 3".af. and economic evaluation.ti. and trial.af		0
		"2 x 3".af. and cost-effect*.ab		4
		"2 x 3".af. and cost-effect*.ti.		0
		"2 x 3".af. and economic evaluation.ab.		0
		"2 x 3".af. and economic evaluation.ti		0
		"3 x 3".af. and cost-effect*.ab.		5
		"3 x 3".af. and cost-effect*.ti.		0
		"3 x 3".af. and economic evaluation.ab.		0
		"3 x 3".af. and economic evaluation.ti		0
		"3 x 2".af. and cost-effect*.ab.		7
		"3 x 2".af. and cost-effect*.ti		0
		"3 x 2".af. and economic evaluation.ab.		0
		"3 x 2".af. and economic evaluation.ti.		0
		"2 x 2 x 2".af. and cost-effect*.ab.		1
		"2 x 2 x 2".af. and cost-effect*.ti.		0
		"2 x 2 x 2".af. and economic evaluation.ab.		0
		"2 x 2 x 2".af. and economic evaluation.ti.		0
		Matrix.af. and cost-effect*.ab. and randomized.ab		79†
		Matrix.af. and cost-effect*.ab. and randomised.ab		25†
		Matrix.af. and cost-effect*.ti. and randomized.ab		22†
		Matrix.af. and cost-effect*.ti. and randomised.ab		5†
		Matrix.af. and economic evaluation*.ab. and randomized.ab		9†
		Matrix.af. and economic evaluation*.ab. and randomised.ab		5†
		Matrix.af. and economic evaluation*.ti. and randomized.ab		4†
		Matrix.af. and economic evaluation*.ti. and randomised.ab		4†
Bmj.com – advanced search		cost-effective (exact phrase in title or abstract) factorial (exact phrase anywhere in article)		5
		cost-effectiveness (all words in title or abstract) factorial (all words anywhere in article)		5
		economic evaluation (all words in title or abstract) factorial (all words anywhere in article)		1
		cost-effective (as phrase) in title or abstract and matrix randomised (all words) in full text.		6†
		cost-effective (as phrase) in title or abstract and matrix randomized (all words) in full text.		4†
		economic evaluation (as phrase) in title or abstract and matrix randomised (all words) in full text.		1†
		economic evaluation (as phrase) in title or abstract and matrix randomised (all words) in full text.		1†
		economic evaluation (all words in title or abstract)	2x2 phrase anywhere in article	0
			2 x 2 phrase anywhere in article	0
			2 by 2 phrase anywhere in article	0
			Two by two phrase anywhere in article	0
			Two x two phrase anywhere in article	0
			2 x 3 phrase anywhere in article	0
			3 x 3 phrase anywhere in article	0
			3 x 2 phrase anywhere in article	0
			2 x 2 x 2 phrase anywhere in article	0
		Cost-effective (all words in title or	2x2 phrase anywhere in article	0
			2 x 2 phrase anywhere in article	1

Website	Databases searched	Terms searched		Hits
		abstract)	2 by 2 phrase anywhere in article	0
			two x two phrase anywhere in article	0
			two by two phrase anywhere in article	3
			2 x 3 phrase anywhere in article	1
			3 x 3 phrase anywhere in article	0
			3 x 2	1
			2 x 2 x 2	0
		Cost-effectiveness (all words in title or abstract)	2x2 phrase anywhere in article	0
			2 x 2 phrase anywhere in article	1
			2 by 2 phrase anywhere in article	0
			two x two phrase anywhere in article	0
			two by two phrase anywhere in article	2
			2 x 3 phrase anywhere in article	0
			3 x 3 phrase anywhere in article	0
			3 x 2 phrase anywhere in article	0
			2 x 2 x 2 phrase anywhere in article	0
Science direct – journals and books, all years		Expert search TITLE-ABSTR-KEY("cost-effective" OR "cost-effectiveness" OR "economic evaluation") and (factorial OR "2x2" OR "2 x 2" OR "2 by 2" OR "two x two" OR "two by two" OR "2 x 3" OR "3 x 3" OR "3 x 2" OR "2 x 2 x 2")		457
		matrix cost-effective and TITLE-ABSTR-KEY(randomised)		5†
		matrix cost-effective and TITLE-ABSTR-KEY(randomized)		93†
		matrix "economic evaluation" and TITLE-ABSTR-KEY(**randomized**)		30†
		matrix "economic evaluation" and TITLE-ABSTR-KEY(**randomised**)		7†
CRD (NHS EED, HTA and DARE		factorial OR "2x2" OR "2 x 2" OR "2 by 2" OR "two x two" OR "two by two" OR "2 x 3" OR "3 x 3" OR "3 x 2" OR "2 x 2 x 2"		499
		matrix		82†
Tufts Advanced search searching articles. No limits		Factorial		8
		2x2		0
		2 x 2		1
		2 by 2		0
		Two by two		0
		Two x two		0
		2 x 3		0
		3 x 3		0
		3 x 2		0
		2 x 2 x 2		0
		Matrix		2†
		Interaction		3
NIHR HTA publications list (http://www.hta.ac.uk/)		Factorial		15
		2x2		7
		2 x 2		‡
		2 by 2		‡
		Two by two		4
		Two x two		4
		2 x 3		‡
		3 x 3		‡
		3 x 2		‡
		2 x 2 x 2		‡
		Interaction		60

Website	Databases searched	Terms searched	Hits
		Matrix	21
HESG papers (searching titles only). No quotations used		factorial	0
		2x2	0
		2 x 2	0
		Two by two	0
		Two x two	0
		2 x 3	0
		3 x 3	0
		3 x 2	0
		2 x 2 x 2	0
		Matrix	0†

† Searched on 3rd September 2010
‡ The query contained only ignored words

Across all searches, a total of 1,671 references were identified (including duplicates). The titles and abstracts of these were scrutinised to identify relevant studies.

Appendix G: Data extraction table for the systematic review of studies conducting economic evaluations of factorial design studies

Table G1. Completed studies

Author/study acronym	**ACUITY (236, 334-336)**	**ADAPT (239, 337, 338)**	**AFIST II (245, 339)**
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation	2008	2003	2004
Year of publication for earliest paper presenting ***methods*** of economic evaluation	2008	2003	2004
Year of publication for earliest paper presenting ***results*** of economic evaluation	2008	2009	2004
Type of factorial design (e.g. partial/fractional)	Full factorial	Full factorial	Full factorial
Total sample size (across all arms)	13819	316	160
Size of factorial matrix (e.g. 2x2)	2x2 + 1	2x2	2x2
Disease area	Acute coronary syndromes	Knee osteoarthritis	Open-heart surgery for coronary artery disease
Sponsor	Manufacturer	US public funding body	Manufacturer and other funding body
Country	US, Europe, Australia, NZ and Canada. Economic evaluation focusses on US patients	US	US
Treatments (or treatment combinations) compared	Factor 1: bivalirudin vs. unfractionated heparin or enoxaparin. Factor 2: "upstream" routine glycoprotein (GP) Iib/IIIa inhibitor (given to all pts immediately) vs. deferred selective GP Iib/IIIa (given immediately before PCI in those pts who underwent PCI). Arm 5: bivalirudin with no GP IIb/IIIa.	Factor 1: Exercise programme 3-times a week vs. no exercise programme. Factor 2: 18 sessions of dietary counselling vs. no dietary counselling. Patients in the control group had a healthy lifestyle group session monthly for first 3 months, with phone contact thereafter	Factor 1: Amiodarone vs. placebo. Factor 2: pacing vs. no pacing.
Type of comparisons	A1 vs. A2 vs. B1 vs. B2 (≥2 dosing regimens of ≥2 interventions)	*0* vs. *a* vs. *b* vs. *ab*	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated	Pharmaceutical	Nutritional supplement/diet + exercise	Pharmaceutical and surgery
Power calculation allowed for	Appears to have been powered for	Powered for main effects. "The main	Not stated. Interactions not

Author/study acronym		ACUITY (236, 334-336)	ADAPT (239, 337, 338)	AFIST II (245, 339)
interaction effect and/or powered to test for interaction?		at-the-margins comparisons between bivalirudin alone vs. bivalirudin+GPIIb/IIIa vs. heparin+GPIIb/IIIa. Power calculation adjusts for multiple comparisons as 2 primary analyses involve comparison with common treatment arm.	effects of weight reduction and physical activity interventions were felt to be additive or nearly additive. If the effect of one factor is additive by a factor p then the joint effect of weight reduction and lifestyle physical activity intervention would have a joint effect of 10% × (1 p). Sample size calculations were made to allow for 75–100% additivity."	mentioned
Type of economic evaluation (e.g. CUA)		CCA	CEA	CEA
Other notable aspects of study design	Cluster randomised	No	No	No
	Cross-over	No	No	No
	Intentionally unbalanced	No	No	No
	Interventions target different diseases	No	No	No
	Other	Trial powered for sequential tests of non-inferiority and superiority		
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	At-the-margins	Linear regression/ANCOVA/ANOVA	Linear regression/ANCOVA/ANOVA
			Not stated	No mention of interaction term
		Main analysis treated the study as a 3-arm trial and compared bivalirudin+GPIIb/IIIa vs. heparin+GPIIb/IIIa and then compared bivalirudin alone vs. heparin+GPIIb/IIIa. Present subgroup analyses by timing of GP IIb/IIIa; comparions between upstream and deferred GP IIb/IIIa are secondary endpoint.	Unclear how ANCOVA was conducted or whether it controlled for both treatments. Results presented for each of the 4 groups; text reports results of comparison between each of the 3 active groups vs. healthy lifestyle sessions	"Power analysis was for a 2x2 fixed effects analysis of variance"; also describe methods for pair-wise comparisons
	Base case economic evaluation	Inside-the-table (treating each cell separately)	Inside-the-table (treating each cell separately)	Inside-the-table (treating each cell separately)

Author/study acronym		ACUITY (236, 334-336)	ADAPT (239, 337, 338)	AFIST II (245, 339)
	Sensitivity analyses of economic evaluation	None	None	At-the-margins
				Costs and cost-effectiveness presented on an at-the-margins basis
Presentation of economic results		Present only costs not ICERs.	Each active group is compared with healthy lifestyle	Costs are compared on an at-the-margins basis, but cost-effectiveness results are presented on inside-the-table basis and at-the-margins basis
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Conducted bootstrapping and present CEAC-like graphs plotting the cummulative proportion of bootstrap replicates against estimated incremental cost; these are presented for bivalirudin monotherapy vs. heparin+upstream GPIIb/IIIa and for bivalirudin monotherapy vs. heparin+deferred GPIIb/IIIa	Present SD around healthcare costs, but otherwise appear to give no indication of the precision around incremental costs or cost-effectiveness.	Cost-effectiveness results are presented on inside-the-table basis and at-the-margins basis as the proportion of bootstrap replicates lying in each quadrant relative to placebo group
Statistical significance of differences of each comparison individually		No significant difference between bivalirudin alone, bivalirudin+GPIIb/IIIa and heparin +GPIIb/IIIa.	Exercise+diet significantly superior to healthy lifestyle. No significant differences between exercise only or diet only groups and the healthy lifestyle group	Clinical paper concludes that A is significantly superior to non-A but that pacing had no significant benefits. However, abstract of economic paper says that A+pacing is significantly superior to all three of the other treatment arms.
Significance of interaction effect		Unclear. Present interactions or subgroup analyses for heparin+GPIIb/IIIa vs. "bivalirudin alone" that include subgrouping by upstream vs. deferred GPIIb/IIIa; unclear whether this is a typo or how interactions were assessed for this comparison, although Figure 3 of Stone 2007 suggests that there is a non-significant qualitative interaction between upstream/downstream and heparin+GPIIb/IIIa vs. bivalirudin alone for composite ischaemia in pts having PCI. Pinto et al note that there	Not stated	Interactions not mentioned, although discussion of an "apparent synergistic effect" was noted in the discussion of the economic paper. The differing conclusions of clinical and economic papers and the probability of pacing being in the NW or SW quadrants also imply that there may be a large *qualitative* interaction for efficacy and cost

Author/study acronym	ACUITY (236, 334-336)	ADAPT (239, 337, 338)	AFIST II (245, 339)
	was no significant interaction for 30-day cost savings with bivalirudin monotherapy compared with heparin + upstream/deferred GPIIb/IIIa		
Results of comparing different methods for statistical analysis	N/A	N/A	Probability that treatment is cost-effective was substantially higher in at-the-margins analysis than inside-the-table analysis. Notably, the probability that pacing increased efficacy compared with placebo was 79% based on at-the-margins analysis and 41% based on inside-the-table analysis, suggesting a qualitative interaction

Author/study acronym		ASCOT 1 (231, 249, 271-275, 340)	ATEAM (246, 341)	Bankhead et al 2001 (116, 342)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2005	2008	2006
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2005	2008	2006
Year of publication for earliest paper presenting ***results*** of economic evaluation		2005	2008	2006
Type of factorial design (e.g. partial/fractional)		Partial factorial	Full factorial	Full factorial
Total sample size (across all arms)		19,257	579	1158
Size of factorial matrix (e.g. 2x2)		2x2	2x4	2x2
Disease area		Cardiovascular disease	Back pain	Breast cancer screening
Sponsor		Manufacturer	MRC	MRC
Country		UK, Ireland and Scandinavia	UK	UK
Treatments (or treatment combinations) compared		Factor 1 (all patients): amlodipine adding perindopril as required vs. atenolol adding bendroflumethiazide. Factor 1 (pts with low cholesterol): atorvastatin vs. placebo	Factor 1: Normal care vs. massage vs. 6 lessons in Alexander technique vs. 24 lessons in the Alexander technique. Factor 2: exercise prescription from a doctor plus behavioural counselling from a nurse vs. no exercise prescription and counselling.	Factor 1: letter from GP reminding about screening vs. no letter. Factor 2: flag in women's notes prompting discussion of screening, vs. no flag.
Type of comparisons		A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Pharmaceutical	Physiotherapy and related techniques	Logistical intervention targetted at health service
Power calculation allowed for interaction effect and/or powered to test for interaction?		No	"Powered for only moderately large interactions"	Not stated
Type of economic evaluation (e.g. CUA)		CEA and CUA	CEA and CUA	CEA
Other notable aspects of study design	Cluster randomised	No	No	No
	Cross-over	No	No	No

Author/study acronym		ASCOT 1 (231, 249, 271-275, 340)	ATEAM (246, 341)	Bankhead et al 2001 (116, 342)
	Intentionally unbalanced	No	No	No
	Interventions target different diseases	No	No	No
	Other	Lipid comparison was prematurely terminated two years before the blood pressure comparison. Results of each comparison published entirely separately		
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	At-the-margins	Linear regression/ANCOVA/ANOVA	Logistic regression
		No mention of interaction term	Interaction term included regardless of significance	No interaction term included
		Analysed using survival data. Each comparison was reported entirely separately		Adjusted for practice with FE model. "The adjusted odds ratios (ORs) for attendance and their associated 95% CIs are presented, and the interaction between the two interventions was also investigated. Also, crude relative risks (RRs) and 95% CIs for the effects of the two interventions were calculated, with the interaction effect for RRs (95% CIs) estimated from simple stratified analyses." Table 4 states that the OR for main effects are based on a model that did not contain an interaction term.
	Base case economic evaluation	Event-based CEA	Inside-the-table (treating each cell separately)	Logistic regression
		Interaction term included regardless of significance		Interaction term included regardless of significance
		Model-based economic evaluation using IPD, in which risk of event was based on survival analysis model including dummies for 3 of the 4 treatment groups (amlodipine-based therapy plus atorvastatin, atenolol-	Present outcomes for the 8 arms separately and calculate ICERs vs. specific comparisons. These comprised the "main" results.	Logistic regression model in WinBUGS, adjusting for clustering by practice and an interaction term. However, interaction term was assigned an informative prior that it is unlikely (with 95% probability) to

Author/study acronym		ASCOT 1 (231, 249, 271-275, 340)	ATEAM (246, 341)	Bankhead et al 2001 (116, 342)
		based therapy plus atorvastatin or amlodipine-based therapy plus placebo, with atenolol plus placebo as the reference category) (249).		exceed the magnitude of the main effects
	Sensitivity analyses of economic evaluation	At-the-margins	At-the-margins	Linear regression/ANCOVA/ANOVA
		2008 paper used within-trial and event-based model approach to compare atenolol and amlodipine (272), while 2005 paper used within trial approach to compare atorvastatin vs. no statin (271), although neither considered factorial nature of trial.		Economic evaluation in primary clinical paper presented only point estimate for ICER (mean incremental cost/mean incremental effect) for flag or letter vs. control.
Presentation of economic results		2009 paper presented regression coefficients and costs separately for each of the 4 arms and calculated ICERs along the frontier (249). 2005 and 2008 papers presented results separately for one pair-wise comparison (271, 272).	Present outcomes for each of the 8 groups separately "to provide the most relevant information for policy makers" and conducted at-the-margins analysis "in line with the convention for a factorial design [referencing Fox et al 2006]". The authors compared each intervention group with "the most appropriate comparator group" [which generally comprised the group that would be expected to be the next most effective]	Brown et al presented incremental cost and additional numbers of attendences for each active treatment arm vs. control group and presented incremental outcomes along the frontier, although the original paper gave only CER vs. control.
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		2009 paper conducted PSA and presented CEACs for multiple comparisons (249). 2005 and 2008 papers presented CEACs for single comparisons from PSA and/or bootstrapping (271, 272).	Present pariwise CEACs for a number of pair-wise comparisons between individual arms (based on bootstrapping)	Presented incremental costs & effects vs. control on CE and CEAC frontier and EVPI allowing for multiple comparisons. These were calculated in WinBUGS.
Statistical significance of differences of each comparison individually		Atenolol was non-significantly superior to amlodipine for primary endpoint and significantly superior for mortality. Atorvastatin significantly superior to placebo for primary endpoint but not mortality.	Exercise and lessons in the Alexander technique, but not massage, remained effective at one year compared with control.	Letter had a significant effect. Flag had no significant effect

Author/study acronym	ASCOT 1 (231, 249, 271-275, 340)	ATEAM (246, 341)	Bankhead et al 2001 (116, 342)
Significance of interaction effect	Significant interaction was observed for CHD events (275). Interactions not mentioned in economic papers. Interactions for costs and QALYs were not reported in papers, although I estimated there to be a small quantitative sub-additive interaction for costs and tiny super-additive interaction for QALYs.	Not significant, although authors commented that "the effect of exercise and 24 Alexander technique lessons combined is less than the sum of the two individual effects"	Not significant. "the interaction OR of less than 1 suggests that there would be little to gain from combining the two interventions." Not clear what they mean by interaction OR [presumably exponent of the coefficient for interaction term
Results of comparing different methods for statistical analysis	Comparing CUA results for UK from 2008 and 2009 papers, the two papers produce analogous conclusions (that amlodipine is cost-effective vs. atenolol (272) or that amlodipine+atorvastin maximises NB (249)), although estimated costs & effects for all arms are slightly different due to different methodology, making direct comparisons difficult.	ICERs differed substantially between at-the-margins and inside-the-table analyses - especially for massage, which was dominated by control based on inside-the-table analysis of the non-exercise groups, but cost £10,800/QALY based on at-the-margins approach.	ICER for letter vs. control was £41/attendance in Welton analysis that allowed for uncertainty and included interaction term, vs. £35/attendance in original clinical paper. By taking an incremental analysis, Brown et al concluded that the flag was inefficient (being extendedly dominated by the other 2 interventions), whereas the primary clinical paper concluded only that flag was less cost-effective than the letter based on CERs for each intervention vs. control

Author/study acronym		BARI 2D (343-345)	Barnett/Sorensen (252, 346)	BELLS (248, 347, 348)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2009	2006	2009
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2009	2006	2009
Year of publication for earliest paper presenting ***results*** of economic evaluation		2009	2006	2009
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Total sample size (across all arms)		2368	126	496
Size of factorial matrix (e.g. 2x2)		2x2	2x2	2x2
Disease area		Patients with both type 2 diabetes and coronary artery disease	Drug abuse	Bell's Palsy
Sponsor		Manufacturer and other funding body	US public funding body	NIHR HTA
Country		Multinational	US	UK (Scotland)
Treatments (or treatment combinations) compared		Factor 1: prompt coronary revascularisation or medical therapy. Factor 2: insulin sensitisation therapy or insulin provision therapy	Factor 1: voucher for free methadone treatment vs. no voucher. Factor 2: 6 months' case management services vs. no such services.	Factor 1: prednisolone vs. placebo. Factor 2: aciclovir vs. placebo.
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Pharmaceutical and surgery	Other	Pharmaceutical
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated.	Not stated	Assumed no interaction between factors in power calculation, alhtough power to detect main effects in the presence of a negative interaction was also estimated
Type of economic evaluation (e.g. CUA)		CEA and CUA	CCA	CEA and CUA
Other notable aspects of study design	Cluster randomised	No	No	No
	Cross-over	No	No	No
	Intentionally unbalanced	No	No	No
	Interventions target different	No	No	No

Author/study acronym		BARI 2D (343-345)	Barnett/Sorensen (252, 346)	BELLS (248, 347, 348)
	diseases			
	Other	Population was divided into 2 subgroups basedon whether the responsible physician felt that CABG or PCI was more apporpriate		
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	At-the-margins	At-the-margins	At-the-margins
		No interaction term included		Interaction term excluded as it was not statistically significant
		Protocol states that the two randomised comparisons will be evaluated separately, with no adjustment for multiple comparisons. Methods not clear in main papers presenting results: presentation of results implies at-the-margins analysis, although also comments that: "In addition, the statistical interactions between the cardiac study groups and the glycemic study groups for rates of death and major cardiovascular events were tested overall and within the PCI and CABG strata at a two-sided alpha level of 0.05". Also mention results of what appears to be a inside-the-table analysis.	Defined 4 primary hypotheses, of which (a), (b) and (c) appear to have been based on at-the-margins analysis, and (d) was based on inside-the-table comparison between cell 00 and cell AB.	Tested data for any interaction prior to evaluating primary endpoint
	Base case economic evaluation	At-the-margins	Inside-the-table (treating each cell separately)	At-the-margins
				Interaction term excluded as it was not statistically significant
		Assumed to be at-the-margins analysis based on the way in which results are presented, although this is not explicitly stated	Costs and outcomes are first compared between individual cells of the design. Statistical tests presented comparing costs & effects between pairs of arms	Estimated using two decision-tree models (one for each factor) based largely on trial data that compared costs and outcomes with and without that factor
	Sensitivity analyses of	None	At-the-margins	Inside-the-table (treating each cell separately)

Author/study acronym		BARI 2D (343-345)	Barnett/Sorensen (252, 346)	BELLS (248, 347, 348)
	economic evaluation		Costs also compared on at-the-margins basis+BC55	Estimated using a decision-tree model with 4 treatment strategies (1 for each group in the trial)
Presentation of economic results		Presented costs and benefits for medical therapy vs. revascularisation (stratified by whether PCI or CABG were most suitable) and for insulin sensitisation vs. insulin provision.	No ICERs presented; both costs and effects are presented but not combined. No clear indication of whether inside-the-table or at-the-margins is base case, although abstract focuses on former	For base case, presented outcomes of at-the-margins analysis
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Presented pair-wise probability cost-effective at 2 thresholds, but no CEACs	No ICERs presented	Presented CEACs for pair-wise comparisons for the base case at-the-margins analyses. For sensitivity analysis comparing 4 arms individually, also presented CEACs for multiple comparisons and probability that each treatment is optimal.
Statistical significance of differences of each comparison individually		No significant differences between revascularisation and medical therapy or between insulin sensitisation and insulin provision	Both case management and vouchers significantly increase costs in at-the-margins analysis. Vouchers significantly increased enrollment on methadone, although other differences were not significantly different.	Prednisolone found to significantly increase chance of recovery, although aciclovir had no significant effect.
Significance of interaction effect		Interaction not significant in overall group, although interaction was borderline significance (p=0.07 for those patients who were more suitable for CABG than PCI. Methods to test for interaction unclear.	Interactions not mentioned in Barnett et al and results of clinical analyses including interaction term were not stated	No significant difference
Results of comparing different methods for statistical analysis		N/A	At-the-margins and inside-the-table give same conclusions.	Inside-the-table analysis (which showed prednisolone without aciclovir to strongly dominate other strategies) largely confirmed findings of at-the-margins analysis (which showed prednisolone to dominate no prednisolone and no aciclovir to dominate aciclovir).

Author/study acronym		**Boyle 2007 (235, 349)**	**Brandon et al 2000 (350)**	**Brandon et al 2004 (251)**
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2007	2000	2004
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2007	2000	2004
Year of publication for earliest paper presenting ***results*** of economic evaluation		2007	2000	2004
Type of factorial design (e.g. partial/fractional)		Full factorial	full factorial	full factorial
Total sample size (across all arms)		161	446	431
Size of factorial matrix (e.g. 2x2)		2x2 + 1	2x2	2x2
Disease area		Language impairment in children	Smoking cessation	Smoking cessation
Sponsor		NIHR HTA	US public funding body	US public funding body
Country		UK	US	US
Treatments (or treatment combinations) compared		Factor 1: Direct vs. indirect therapy. Factor 2: individual vs. group therapy. There was also a control group receiving usual care	Factor 1: telephone hotline vs. no telephone hotline. Factor 2: repeated relapse prevention mailings vs. no mailings	Factor 1: 1 mailing of material to prevent relapse of smoking vs. 8 mailings. Factor 2: 1 booklet vs. 8 booklets
Type of comparisons		Other	0 vs. a vs. b vs. ab	Dose-ranging for 1 drug (e.g. freq x dose)
Type(s) of treatment evaluated		Other	Other	Other
Power calculation allowed for interaction effect and/or powered to test for interaction?		Yes (if interaction effect was 0.50, compared with power to detect a 0.40 difference for main effects)	No power calculation	No power calculation
Type of economic evaluation (e.g. CUA)		CEA	CEA	CEA and CUA
Other notable aspects of study design	Cluster randomised	No	no	No
	Cross-over	No	no	No
	Intentionally unbalanced	No	no	No
	Interventions target different diseases	No	no	No
	Other			

Author/study acronym		Boyle 2007 (235, 349)	Brandon et al 2000 (350)	Brandon et al 2004 (251)
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Linear regression/ANCOVA/ANOVA	Linear regression/ANCOVA/ANOVA	Logistic regression
		Interaction term excluded as it was not statistically significant	Interaction term excluded as it was not statistically significant	Interaction term included regardless of significance
		Controlled for covariates	Analysed various outcomes using ANOVA. Found hotline to have no significant effect and no interaction with mailings, so focus on mailings comparison. No explicit statement about whether or not the interaction term was or was not included in the final model, but I am assuming that it was omitted on grounds that it was non-significant.	"Interaction term, Contact x Content, which was entered as a second step in the logistic regression equation"
	Base case economic evaluation	Inside-the-table (treating each cell separately)	At-the-margins	Inside-the-table (treating each cell separately)
			No mention of interaction term	
		Compared total costs using ANOVA. Pair-wise comparisons conducted using t-tests. Means and 95% CIs for each group appear to have been calculated using bootstrapping and analytical methods. Main method of presentation appears to be inside-the-table analysis	Simple comparison of cost/incremental benefits for mailings vs. no mailings.	Each group is compared with control group
	Sensitivity analyses of economic evaluation	At-the-margins	None	None
		Costs, outcomes and cost-effectiveness also compared on an at-the-margins basis		
Presentation of economic results		Gave outcomes separately for each arm and at-the-margin. Compared costs for all direct therapy groups vs. control and calculates **average** cost-effectiveness ratios for each of the 5 groups. In Dickinson, 2009, the authors focus on costing results and present "cost ratios" for pair-wise comparisons equal to the cost in one group divided	Cost and ICER mentioned in text only	Present incremental costs, QALYs, abstinence rates and ICERs compared with control group

Author/study acronym	Boyle 2007 (235, 349)	Brandon et al 2000 (350)	Brandon et al 2004 (251)
	by the cost in a comparator group; some of these ratios appear to compare cells inside-the-table and others appear to be at-the-margins.		
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)	Two-way sensitivity analysis	None	None
Statistical significance of differences of each comparison individually	No significant differences in primary endpoint; direct therapy superior in some secondary endpoints. Significant differences in cost	Hotline not significant. Mailing significantly superior to no mailings on several endpoints.	Giving 8 booklets rather than 1 significantly reduced relapse rates, although frequency of booklet mailing didn't affect outcome
Significance of interaction effect	Not significant	Not significant	Not significant, although Table 3 implies that there is a qualitative interaction, with repeated letters producing lower abstinence rates than control group, but with repeated mailings increasing abstinence rates compared with sending all 8 booklets at once.
Results of comparing different methods for statistical analysis	Inside-the-table analysis suggests that therapist-led group therapy dominates all treatments except control. At-the-margins suggests that group dominates individual therapy, but that therapist-led therapy is more costly and more effective than assistant-led therapy	N/A	N/A

Author/study acronym		**CADILLAC (351, 352)**	**Cantor/Morisky/Green (250, 353, 354)**	**CELL (355, 356)**
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2003	1985	1996
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2003	1985	1996
Year of publication for earliest paper presenting ***results*** of economic evaluation		2003	1985	1996
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Total sample size (across all arms)		2082	400	681
Size of factorial matrix (e.g. 2x2)		2x2	2x2x2	2x3
Disease area		Acute myocardial infarction	Hypertension	Primary prevention of cardiovascular events
Sponsor		Manufacturer	US public funding body	Manufacturer and other funding body
Country		US (Multinational trial, although economic evaluation focusses on US patients)	US	Sweden
Treatments (or treatment combinations) compared		Factor 1: PTCA vs. stenting with MultiLink stent. Factor 2: abciximab vs. no abciximab	Factor 1: Exit interview encouraging compliance immediately after physician appointment (E1) vs. no exit interview (C1). Factor 2: Interview with spouse to encourage compliance (E2) vs. no interview (C2). Factor 3: series of up to 3 group sessions with a social worker (E3) vs. no sessions (C3)	Factor 1: Intensive advice (group sessions) vs. usual advice. Factor 2: pravastatin vs. usual care vs. no drug.
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab	Other
Type(s) of treatment evaluated		Pharmaceutical and surgery	Other	Pharmaceutical + other
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated. Appears to have been powered for a inside-the-table analysis	No power calculation? (no apparent mention of rationale for sample size in Morisky 1983 or Green 1975)	Appears to be powered for inside-the-table analysis, although not entirely clear
Type of economic evaluation (e.g. CUA)		CUA	CEA	CEA
Other notable	Cluster	No	No	No

Author/study acronym		CADILLAC (351, 352)	Cantor/Morisky/Green (250, 353, 354)	CELL (355, 356)
aspects of study design	randomised			
	Cross-over	No	No	No
	Intentionally unbalanced	No	No	No
	Interventions target different diseases	No	No	No
	Other	Designed for non-inferiority wrt the comparison between stent alone and PTCA+abciximab	Interventions given at different times: factor 1 first, then factor 2 a few weeks later and then factor 3	
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Inside-the-table (treating each cell separately)	Inside-the-table (treating each cell separately)	Inside-the-table (treating each cell separately)
		"Two primary hypotheses were prespecified: stenting without abciximab is superior to PTCA without abciximab, and stenting without abciximab is not inferior to PTCA plus abciximab." The majority of the results are presented separately for each of the 4 arms, with p-values for the difference between any of the groups being calculated from one-way ANOVA. However, the OR for all pts receiving stent vs. all patients receiving PTCA was shown in a figure with subgroup analyses.	Appear to have compared each group with the control group	
	Base case economic evaluation	At-the-margins	Inside-the-table (treating each cell separately)	Inside-the-table (treating each cell separately)
			Compare each arm with control	Excludes intensive advice+drug arm as it is dominated (no more effective and more costly) and compares costs & calculates ICERs for drug only, intensive advice only and no treatment (although it is unclear whether the no treatment arm is

Author/study acronym		CADILLAC (351, 352)	Cantor/Morisky/Green (250, 353, 354)	CELL (355, 356)
				based on the usual advice only arm or assuming no cost and no change in risk from baseline). Based on patient numbers, the economic evaluation appears to have excluded patients randomised to placebo and include only men on active drug or no drug.
	Sensitivity analyses of economic evaluation	None	None	None
Presentation of economic results		Presented costs and effects in at-the-margins format	Present CERs for each arm vs. control based on a wide range of outcomes. Don't present ICERs along the frontier, but do present results on CE plane and identify strongly dominated but not extendedly dominated options and present marginal costs and effects excluding strongly dominated options.	Present costs for all 4 arms, plus incremental costs, life-years gained and ICERs for intensive advice vs. no treatment and usual advice+drug vs. no treatment and drug+usual advice vs. intensive advice
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Estimated probability cost-effective for stenting vs. PTCA, but did not present CEACs	None	Scenario analysis only
Statistical significance of differences of each comparison individually		Significant differences between groups when all 4 groups compared [in 1-way ANOVA]. Stenting alone significantly superior to PTCA alone but non-inferior to PTCA+abciximab. In economic evaluation, abciximab was found to non-significantly reduce number of QALYs accrued when mortality difference were ignored, and non-significantly increase QALYs when mortality differences were included.	Authors conclude that the programmes significantly improved appointment keeping, weight control, blood pressure control and mortality.	Pravastatin significantly reduces Framingham risk score. No difference between placebo and no drug. Advice has small but statistically significant effect in pts not having active drug, but no effect in patients on placebo or no drug

Author/study acronym	**CADILLAC (351, 352)**	**Cantor/Morisky/Green (250, 353, 354)**	**CELL (355, 356)**
Significance of interaction effect	Not stated	not stated	Significance of interaction not stated, though a qualitative interaction clearly exists at 12 and 18 months based on Figure 1. Advice has small but statistically significant effect in pts not having active drug, but no effect in patients on placebo or no drug
Results of comparing different methods for statistical analysis	N/A	N/A	N/A

Author/study acronym		COPE II (357)	ERUPT (358)	Exercise Intervention Trial (359)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2009	2008	2002
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2009	2008	2002
Year of publication for earliest paper presenting ***results*** of economic evaluation		2009	2008	2002
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	full factorial
Total sample size (across all arms)		159	2833	4999
Size of factorial matrix (e.g. 2x2)		2x2	2x2	2x2
Disease area		asthma	Dentistry	Breast cancer prevention
Sponsor		Other public funding body	Other UK public funding body	US public funding body
Country		Netherlands	UK (Scotland)	US
Treatments (or treatment combinations) compared		Factor 1: self-treatment of exacerbations (and self-management programme) vs. self-management programme only; Factor 2: COPE-active physiotherapy exercise programme vs. no COPE-active	Factor 1: fee (financial incentive for applying sealant) vs. no fee. Factor 2: 1-day education workshop on evidence-based practice vs. no workshop.	Factor 1: inclusion/exclusion of personal invitation letter. Factor 2: first class stamps vs. bulk mail
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)
Type(s) of treatment evaluated		Physiotherapy and related techniques	Logistical intervention targetted at health service	Other
Power calculation allowed for interaction effect and/or powered to test for interaction?		Appears to have been powered based on no interaction. Not explicitly stated in sample size section, although it is noted at the beginning of the methods that "this design assumes that both interventions do not interact with each other"	Not powered to detect clinically-significant interaction, and appears to have been powered based on at-the-margins analysis "(fee vs. no fee; education vs. no education)"	No power calculation described (primary aim was to randomise women to RCT)
Type of economic evaluation (e.g. CUA)		CEA	CEA	CEA
Other notable	Cluster	No	Yes	no

Author/study acronym		COPE II (357)	ERUPT (358)	Exercise Intervention Trial (359)
aspects of study design	randomised			
	Cross-over	No	No	no
	Intentionally unbalanced	No	No	no
	Interventions target different diseases	Not stated	No	no
	Other	Factors administered sequentially: physiotherapy given in months 2-12, whereas self-management and self treatment training given in month 1		
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	At-the-margins	Linear regression/ANCOVA/ANOVA	Logistic regression
		Interaction term excluded as it was not statistically significant	No mention of interaction term	Interaction term excluded as it was not statistically significant
		Chi-square tests used to compared proportion of patients with an individual value for the exacerbation days/year larger than the median, the 75th and the 90th percentile. However, results are only presented for self treatment vs. control (not for COPE physiotherapy)	Means presented as inside-the-table and at-the-margins, although differences generally only presented for at-the-margins analysis.	
	Base case economic evaluation	At-the-margins	Inside-the-table (treating each cell separately)	Unclear
				Not stated
		Cost-effectiveness evaluated using a decision-analytical model employing patient-level data on resource use. Economic evaluation only presented on self treatment vs. control (not on COPE physiotherapy)	Simply presented fees and outcomes for each group	Appears to be at-the-margins. Simply present ACERs of cost/pt randomised for first-class mail and for bulk rate mail
	Sensitivity analyses of economic evaluation		None	None
			N/A	
Presentation of economic results			Presented ICER for each active treatment group vs. control. Costs not	Simply present ACERs of cost/pt randomised for first-class mail and for

Author/study acronym	COPE II (357)	ERUPT (358)	Exercise Intervention Trial (359)
		reported.	bulk rate mail
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)	Cost-effectiveness planes for pair-wise comparison between self-treatment and no self-treatment	None	None
Statistical significance of differences of each comparison individually	Self-treatment vs. control significant and some outcomes but not others. No results presented for COPE this year therapy. Study doesn't explicitly state which is the primary endpoint, although number of exacerbation days appears to be main endpoint, while being statistically tested in three different ways.	Fee significantly increased sealant use, while education had no significant effect.	First class mail non-significantly increased responses vs. bulk mail. Including personal letter had no effect
Significance of interaction effect	Non-significant? ("Before the final analysis we checked for interactions between both interventions and concluded that no interaction was present")	**Non-significant qualitative interaction for primary endpoint.**	Non-significant
Results of comparing different methods for statistical analysis	N/A	N/A	N/A

Author/study acronym	FRISC II (270, 276-280)	GISSI-3 (360, 361)	GISSI-P (243, 362)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation	2003	1998	2001
Year of publication for earliest paper presenting ***methods*** of economic evaluation	2003	1998	2001
Year of publication for earliest paper presenting ***results*** of economic evaluation	2003	1998	2001
Type of factorial design (e.g. partial/fractional)	Partial factorial	full factorial	full factorial
Total sample size (across all arms)	3469	19394	11324
Size of factorial matrix (e.g. 2x2)	2x2	2x2	2x2
Disease area	Myocardial infarction	Acute myocardial infarction	2ndary prevention of myocardial infarction
Sponsor	Manufacturer and other funding body	manufacturer	Manufacturer
Country	Scandinavia	Italy	Italy
Treatments (or treatment combinations) compared	Factor 1: Dalteparin vs. placebo. Factor 2: Invasive therapy vs. non-invasive. Patients were randomised to both comparisons unless they were contraindicated to invasive therapy or were recruited after the invasive therapy comparison had stopped recruiting	Factor 1: Lisinopril vs. open-label control. Factor 2: nitrates (IV nitrates for first 24 hours, then transdermal GTN) vs. open-label control.	Factor 1: n-3 polyunsaturated fatty acids (PUFA) vs. placebo. Factor 2: vitamin E vs. placebo.
Type of comparisons	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated	Pharmaceutical and surgery	Pharmaceutical	Nutritional supplement
Power calculation allowed for interaction effect and/or powered to test for interaction?	Not clearly stated, although power calculations appear to have been done separately for each comparison, suggesting that was powered for at-the-margins. Authors state that main reason for using factorial design was efficiency not to estimate interactions.	Not stated	Powered for inside-the-table analysis

Author/study acronym		FRISC II (270, 276-280)	GISSI-3 (360, 361)	GISSI-P (243, 362)
Type of economic evaluation (e.g. CUA)		CEA	CEA	CEA
Other notable aspects of study design	Cluster randomised	No	No	No
	Cross-over	No	No	No
	Intentionally unbalanced	No	No	No
	Interventions target different diseases	No	No	No
	Other			
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	At-the-margins	At-the-margins	Inside-the-table and at-the-margins co-primary analyses
		Interaction test conducted as secondary analysis	N/A	N/A
		Comparisons reported in separate paper. Main clinical paper on dalteparin vs. placebo (280) conducted purely at-the-margins with no mention of interaction (although this analysis was done solely on patients with non-invasive treatment). Main clinical paper on intensive vs. non-intensive (279) was done at-the-margins for primary analysis, but also evaluates interactions in secondary analyses (using logistic regression) and presents some results inside-the-table.	Protocol-specified comparisons were L vs. no L, nitrates vs. no nitrates and N+L vs. neither	"2-way" analysis done first, then 4-way analysis comparing 3 active arms with control and combination therapy with 2 monotherapy arms. Analysis done by KM survival curves and Cox regression. Multivariate analyses with interaction terms also investigated.
	Base case economic evaluation	At-the-margins	At-the-margins	Inside-the-table (treating each cell separately)
		No mention of interaction term	N/A	
				Although the paper covers only n3 PUFA not vitamin E, base case results use the results of 4-way comparison as the study was designed to give a comparison

Author/study acronym		FRISC II (270, 276-280)	GISSI-3 (360, 361)	GISSI-P (243, 362)
				between the 4 groups
	Sensitivity analyses of economic evaluation		none	At-the-margins
Presentation of economic results		Results for each comparison presented in separate papers with very little mention of the other comparison, although this weakness was pointed out in a letter by Moller et al (278).	Present at-the-margins comparison of costs and efficacy for lisinopril vs. no lisinopril	Costs for n3 PUFA and control presented for both 2-way and 4-way analysis
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Calculate 95% CI around cost difference, but give no indication of uncertainty around ICERs.	Present 95% around the ICER that are based only on the 95% CI for lives saved, taking no account of costs.	Present CIs around ICERs and life-years but not costs. No clear indication of how 95% CI were derived and no CEACs
Statistical significance of differences of each comparison individually		No significant difference in clinical outcomes or costs between dalteparin and placebo at 3 months. Invasive treatment significantly reduced incidence of MI or death combined at 6 months and 1 year and significantly increased costs at 1 year.	Lisinopril significantly reduced events compared with no lisinopril. Nitrates had no signifiant effect.	n3-PUFA significantly reduced events based on both 2-way and 4-way analysis. Vitamin E had no significant effect
Significance of interaction effect		No significant interaction observed in sample of pts included in invasive vs. non-invasive treatment. Interactions not assessed (and cannot be assessed) in the sample included in the main paper on dalteparin vs. placebo.	Not stated	Not significant, but relatively large: RR for PUFA vs. placebo was 10% by 2-way analysis and 15% by 4-way analysis
Results of comparing different methods for statistical analysis		N/A	None	Results differ by about 50% between inside-the-table and at-the-margins analysis (both in terms of relative risks and ICERs), though this wouldn't change the conclusions at a 40,000 Euro/QALY ceiling ratio

Author/study acronym		Halpern et al 2002 (363)	Hollis 2007 (247)	HPS (149, 364-368)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2002	2007	2005
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2002	2007	2005
Year of publication for earliest paper presenting ***results*** of economic evaluation		2002	2007	2005
Type of factorial design (e.g. partial/fractional)		full factorial	Full factorial	Full factorial
Total sample size (across all arms)		1200	4614	20,536
Size of factorial matrix (e.g. 2x2)		2x2x2	2x3	2x2
Disease area		Physicians	Smoking cessation	Cardiovascular disease
Sponsor		Manufacturer	US public funding body	MRC
Country		US	US	UK
Treatments (or treatment combinations) compared		Factor 1: $5 vs. $10 incentive mailed in envelope with questionnaire. Factor 2: large envelope vs. standard small envelopes. Factor 3: inclusion of a peppermint sweet vs. no sweet.	Factor 1: Brief, moderate or intensive telephone counselling. Factor 2: offer of free nicotine replacement therapy (NRT) patches.	Factor 1: Simvastatin vs. placebo. Factor 2: vitamin supplement vs. placebo.
Type of comparisons		A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Logistical intervention targetted at health service	Pharmaceutical + other	Pharmaceutical + nutritional supplement
Power calculation allowed for interaction effect and/or powered to test for interaction?		Appears to have been powered to detect large interactions	Not stated	Interactions not mentioned in any publication. However, study was highly powered to detect significant effects for each factor in patient subgroups, so is likely to have been powerful enough to detect moderate interaction effects.
Type of economic evaluation (e.g. CUA)		CEA	CEA	CEA
Other notable	Cluster	No	No	No

Author/study acronym		Halpern et al 2002 (363)	Hollis 2007 (247)	HPS (149, 364-368)
aspects of study design	randomised			
	Cross-over	No	No	No
	Intentionally unbalanced	yes	Yes	No
	Interventions target different diseases	No	No	No
	Other			
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Logistic regression	Unclear	At-the-margins
		Not stated	No mention of interaction term	No mention of interaction term
		Used chi squared tests to evaluate main effects and logistic regression to evaluate two-way interactions between main effects	"We used multiple logistic regression to compare the moderate and intensive interventions to brief intervention, the NRT offer to no NRT offer and an NRT offer with the moderate compared to intensive arms." In results, present OR for NRT vs. no NRT and for moderate and intensive vs. brief+NoNRT, but also refer to moderate*NRT and intensive*NRT interaction terms	Log-rank tests used for primary endpoints. No explicit comments on how interactions will be investigated, but does comment that secondary analyses will assess treatmetn effects of each therapy with/without the other (1999)
	Base case economic evaluation	At-the-margins	Inside-the-table (treating each cell separately)	Event-based CEA
				No interaction term included
			Compared each arm with brief counselling and no NRT, but used OLS to calculate mean total costs and SD for each arm	BMJ 2006: event-based analysis using regression equations predicting event rates, cost of events and life expectancy within a Markov model. Made no allowance for vitamin treatment
	Sensitivity analyses of economic evaluation	None		Other
				Lancet 2005: Effect of statins on incidence and cost of vascular events assumed to be proportional and drug costs assumed to be additive. Vitamin allocation ignored; study treated as 2-arm trial on statins

Author/study acronym	Halpern et al 2002 (363)	Hollis 2007 (247)	HPS (149, 364-368)
Presentation of economic results	Present costs and ICERs for $10 incentive vs. $5 incentive, based on costs and effects averaged over all other comparisons. Also presented ACERs (total cost/total responses) for $5 and $10.	Present ICERs for each arm vs. brief+NoNRT arm	Presented costs, effects and cost-effectiveness for statin vs. non-statin overall and in subgroups.
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)	Present 95% CI around ICERs based on bootstrapping	Present "ranges" for each ICERs "calculated using standard deviations and 12-month abstinence"	95% CI and extensive subgroup analyses but no CEACs
Statistical significance of differences of each comparison individually	$10 incentive significantly increased response rate vs. $5 incentive, but envelope size and mints had no significant effect.	NRT, intensive counselling and moderate counselling showed significant benefits	Vitamins had no significant effects on vascular events or mortality, whereas statins significantly reduced all vascular events and deaths
Significance of interaction effect	However, since either a large envelope OR a mint increases postal costs, but having both rather than just 1 has no effect on postage cost, there is authomatically an interaction for cost	Not significant for primary endpoint. Not reported for other endpoints	Not significant on multiplicative scale (χ-squared for "heterogeneity" between pts in the other allocation=0.03)
Results of comparing different methods for statistical analysis	N/A	N/A	N/A

Author/study acronym		NC STRIDES (79)	NIMH-MTA (242, 369-371)	PREVEND IT (372, 373)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2009	2005	2006
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2009	2005	2006
Year of publication for earliest paper presenting ***results*** of economic evaluation		2009	2005	2006
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Total sample size (across all arms)		825	579	864
Size of factorial matrix (e.g. 2x2)		2x2	2x2	2x2
Disease area		Healthy eating in general public and colorectal cancer survivors	ADHD	Prevention of cardiovascular disease; recruited patients with albuminuria
Sponsor		US public funding body	US public funding body	Manufacturer and other funding body
Country		US	US	Netherlands
Treatments (or treatment combinations) compared		Factor 1: tailored print communications (TPC) vs. control. Factor 2: brief telephone-based motivational interviewing (TMI) vs. control.	Factor 1: intensive medication management vs. not. Factor 2: multi-component behavioural treatment vs. not	Factor 1: Fosinopril vs. placebo. Factor 2: pravastatin vs. placebo
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Other	Pharmaceutical + other	Pharmaceutical
Power calculation allowed for interaction effect and/or powered to test for interaction?		No power calculation	Yes: powered based on inside-the-table comparisons	Not stated. Wording of sample size calculation implies that sample size was based on at-the-margins approach only.
Type of economic evaluation (e.g. CUA)		CEA	CEA	CEA
Other notable aspects of study design	Cluster randomised	No	no	No
	Cross-over	No	no	No
	Intentionally unbalanced	No	no	No
	Interventions	No	no	No

Author/study acronym		NC STRIDES (79)	NIMH-MTA (242, 369-371)	PREVEND IT (372, 373)
	target different diseases			
	Other	This study is also stratified by cancer status, since some pts were colorectal cancer survivors and others were matched controls.	In the combined treatment group, the two treatments were integrated by information sharing between teacher-consultant and pharmacotherapist that was used to guide overall decisions. As result, combination treatment was not the simple addition of the two unimodal treatments	RCT nested within an observational study. Economic evaluation is partly model-based and evaluates the cost-effectiveness of screening for albuminuria and treatment with fosinopril, rather than explicitly being limited to the comparisons in the factorial design.
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Linear regression/ANCOVA/ANOVA	Linear regression/ANCOVA/ANOVA	At-the-margins
		No mention of interaction term	Interaction term included regardless of significance	Not stated
		Not clearly stated whether both interventions were considered in the same model. Results section appears to be based on pair-wise comparisons between each of the 3 active treatments and control	"When omnibus RR [random-effects regression] analyses comparing all 4 groups were significant, 3 sets of pairwise comparisons were performed, each set addressing 1 of the principal study questions: (1) medication management vs. behavioral treatment (2 tailed); (2) combined treatment vs. medication management and combined treatment vs. behavioral treatment (1 tailed, assuming the superiority of combined treatment); and (3) community care vs. medication management, community care vs. behavioral treatment, and community care vs. combined treatment (1 tailed, assuming the superiority of MTA treatments)." Adjusted for multiple comparisons	
	Base case economic evaluation	Inside-the-table (treating each cell separately)	Inside-the-table (treating each cell separately)	At-the-margins
		Present pair-wise comparisons between arms. Methods used unclear		

Author/study acronym		NC STRIDES (79)	NIMH-MTA (242, 369-371)	PREVEND IT (372, 373)
	Sensitivity analyses of economic evaluation	Not stated	None	None
			No sensitivity analyses compared different ways of dealing with factorial design	
Presentation of economic results		Present pair-wise comparisons between arms. Methods used unclear	Present costs and outcomes separately for each group	Presented costs, effects and cost-effectiveness for screen and treat with fosinopril vs. no screening or treatment
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Unclear: no explicit mention of uncertainty, although present table with a column entitled "Cost-effective" with values between 0.822-0.992 that might be probability that treatment is cost-effective compared with a single comparator, although it is impossible to tell since these values are not mentioned in the paper.	Uncertainty dealt with using bootstrapping. Jensen et al show CE plane with separate clusters for each of the 4 groups. Foster et al show CEACs for multiple comparisons with the 4 strategies shown, based on bootstrapping and show NHB and 95% CI for each arm based on different outcome measures.	Presented pair-wise CEACs and CE planes for screen and treat with fosinopril vs. no screening or treatment
Statistical significance of differences of each comparison individually		Group receiving TPC+TMI produced significant improvements for fruit and vegetable consumption but not exercise; it is unclear whether this conclusion is based on comparison with baseline or between groups	Medication and combined groups showed significantly greater improvement than behavioural and community care groups. Combined treatment was not significantly better than medication alone	Fosinopril showed significant benefits, although pravastatin did not lower risk of CV mortality significantly.
Significance of interaction effect		Not stated.	Not stated	Not stated
Results of comparing different methods for statistical analysis		N/A	N/A	N/A

Author/study acronym		Richards (115, 116, 374)	Rodewald (375)	Saillour-Glenisson et al 2005 (253)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2006	1999	2005
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2006	1999	2005
Year of publication for earliest paper presenting ***results*** of economic evaluation		2006	1999	2005
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Total sample size (across all arms)		6133	3015	1306
Size of factorial matrix (e.g. 2x2)		2x2	2x2	2x2
Disease area		Breast cancer screening	Complete birth cohort	Thyroid tests
Sponsor		MRC	US public funding body	Other public funding body
Country		UK	US	France
Treatments (or treatment combinations) compared		Factor 1: letter from GP reminding about screening vs. no letter. Factor 2: flag in women's notes prompting discussion of screening, vs. no flag.	Factor 1: tracking with outreach to bring underimmunised children to primary care provider. Factor 2: prompting (primary care office policy change to identify and reduce missed immunisation opportunities	Factor 1: pocket card aide memoir [PAM] vs. no card. Factor 2: test request form giving information on when thyroid tests are necessary vs. no form [FPC]
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Logistical intervention targetted at health service	Logistical intervention targetted at health service	Logistical intervention targetted at health service
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated. Comment that the interaction was investigated with "only limited power"	No power calculation conducted	Unclear. Comments that sample size calculations took into consideration the comparision of the efficacy of FPC and PAM with respect to the control group, although interactions were not explicitly mentioned
Type of economic evaluation (e.g. CUA)		CEA and CUA	CEA	CEA
Other notable aspects of study design	Cluster randomised	Yes	No	Yes
	Cross-over	No	No	No

Author/study acronym		Richards (115, 116, 374)	Rodewald (375)	Saillour-Glenisson et al 2005 (253)
	Intentionally unbalanced	No	No	No
	Interventions target different diseases	No	No	No
	Other			FPC was cluster-randomised at the level of hospitals, while PAM was cluster-randomised at the level wards
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Logistic regression	Linear regression/ANCOVA/ANOVA	Logistic regression
		No interaction term included	Not stated	Interaction term excluded as it was not statistically significant
		"the interaction between the two interventions was also investigated, although with only limited power." Interaction terms included in regression model in a secondary analysis. Table 5 states that the OR for main effects are based on a model that did not contain an interaction term.	"Two-way ANOVA was used to test for effects of each of the interventions on the outcome measures, and for interactions between the interventions." no clear indication of whether interaction terms were included in the base case analysis	An interaction between FPC and PAM was evaluated. In the absence of an interaction, the proper effect of each intervention was studied using a "generalised linear mixed model" to account for cluster randomisation and stratification by pair of hospitals, although odds ratios were presented, implying that logistic regression was used. Results text implies that the efficacy of each intervention was based on simple effects (i.e. a comparison between cells A0 and 00), although Table II suggests that treatment effects were based on logistic regression with no interaction term
	Base case economic evaluation	Logistic regression	Unclear	Inside-the-table (treating each cell separately)
		Interaction term included regardless of significance	Not stated	
		Logistic regression model in WinBUGS, adjusting for clustering by practice and an interaction term. In the original publication (Brown 2006), the interaction term was assigned an	Appear to have taken the total cost per child of the outreach intervention and divided it by the "main effect for tracking/outreach in the ANOVA results". No estimation of uncertainty	ICERs were calculated relative to cell 00 (neither FPC nor PAM) on a inside-the-table basis.

Author/study acronym		Richards (115, 116, 374)	Rodewald (375)	Saillour-Glenisson et al 2005 (253)
		informative prior that it is unlikely (with 95% probability) to exceed the magnitude of the main effects. The logistic regression model estimated the probability of attending screening for each patient and the cost of screening and the cost of the intervention was added on (allowing for the uncertainty around the cost parameters).	around this figure. Unclear whether the main effect in question was from an analysis with or without interaction terms.	
	Sensitivity analyses of economic evaluation	Various	None	None
		1) Economic evaluation in primary clinical paper presented only point estimate for ICER (mean incremental cost/mean incremental effect) for flag or letter vs. control. (2) Welton et al estimated EVPPI prior to the start of the Richards study. (3) Welton et al estimated outcomes with an informative, evidence-based prior around treatment effect. (4) Welton et al estimated outcomes with no interaction.		
Presentation of economic results		Brown et al presented incremental cost and additional numbers of attendences for each active treatment arm vs. control group and presented incremental outcomes along the frontier, although the original paper gave only CER vs. control.	Cost-effectiveness only estimated for outreach/tracking vs. no outreach/tracking. Simply estimated point estimates for the ICERs by dividing costs by incremental benefits.	Present incremental cost relative to cell 00 and total and incremental outcomes; ICERs calculated relative to cell 00
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Presented incremental costs & effects vs. control on CE and CEAC frontier and EVPI allowing for multiple comparisons. These were calculated in WinBUGS. EVPPI estimated by Welton et al (both before and after the trial was conducted)	None	No presentation of uncertainty around economic results besides showing heterogeneity between types of hospitals

Author/study acronym	Richards (115, 116, 374)	Rodewald (375)	Saillour-Glenisson et al 2005 (253)
Statistical significance of differences of each comparison individually	Both flag and letter had a significant effect	Tracking/outreach had significant effect. Prompting "did not have a measurable effect"	FPC showed significantly superiority, but PAM had no significant effect
Significance of interaction effect	No significant interaction. "The findings ruled out any marked disadvantage from using both interventions simultaneously."	No significant interaction	No significant interaction for clinical outcome (p=0.21)
Results of comparing different methods for statistical analysis	ICER for letter vs. control was £27/attendance in Welton analysis that allowed for uncertainty and included interaction term, vs. £26/attendance in original clinical paper. By taking an incremental analysis, Brown et al concluded that the flag was inefficient (being dominated by letter), whereas the primary clinical paper concluded only that flag was less cost-effective than the letter based on CERs for each intervention vs. control. The difference was even more stark for the comparison between flag+letter and letter alone, which was £171/attendance in Welton et al and £41/attendance in the clinical paper. In the additional analyses presented by Welton et al (Table 4), the total NMB of the 4 interventions differs depending on whether conclusions are based on evidence before the Richards study, the evidence after Richards study with an evidence-based prior, or whether or not an interaction effect was included in an analysis after Richards study with a vague prior for main effect.	N/A	N/A

Author/study acronym	Salize/Twardella (240, 376)	SAPPHIRE (269, 281)	Thomas et al (234, 377)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation	2009	2009	2005
Year of publication for earliest paper presenting ***methods*** of economic evaluation	2009	2009	2005
Year of publication for earliest paper presenting ***results*** of economic evaluation	2009	2009	2005
Type of factorial design (e.g. partial/fractional)	full factorial	Partial factorial	Incomplete factorial
Total sample size (across all arms)	577	200	786
Size of factorial matrix (e.g. 2x2)	2x2	2x2	2x2x2
Disease area	Smoking cessation	Shoulder pain	Knee pain
Sponsor	Other public funding body	MRC	Other UK public funding body
Country	Germany	UK	UK
Treatments (or treatment combinations) compared	Factor 1: Training and incentive (TI) for GPs. Factor 2: Training and medication (TM: I.e. GPs were trained and had opportunity to prescribe NRT or buproprion at no cost to patient). Control group appear to have received no training.	Factor 1 (cluster randomised): GP training vs. no training. Factor 2 (individually randomised): lignocaine or cortisone	Factor 1 (all pts): exercise programme vs. no exercise programme. Factor 2 (all pts): Monthly telephone contact giving advice on pain management. Factor 3 (00 and AB groups only): placebo health food tablet vs. no placebo.
Type of comparisons	0 vs. a vs. b vs. ab	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated	Pharmacological + logistical intervention targetted at health service	Pharmaceutical + other	Physiotherapy and related techniques
Power calculation allowed for interaction effect and/or powered to test for interaction?	Unclear. Comment that "power calculations were based on the expected point prevalence of abstinence of 5% in usual care, 10% in interventions TI and TM, and 15% in the intervention TI+TM in a 12-month follow-up", which would imply	Not stated, although discussion of sample size considerations implies that separate power calculations were conducted for each factor	Not stated

Author/study acronym		Salize/Twardella (240, 376)	SAPPHIRE (269, 281)	Thomas et al (234, 377)
		no interaction on additive scale, but a large super-additive interaction on the logarithmic scale that was used to analyse results. Interactions not mentioned.		
Type of economic evaluation (e.g. CUA)		CEA	CUA	CEA
Other notable aspects of study design	Cluster randomised	yes	Cluster-randomised for 1 factor but not all	No
	Cross-over	no	No	No
	Intentionally unbalanced	no	No	yes
	Interventions target different diseases	no	No	No
	Other	Since it appears that no training was given to the control group and the TI+TM group appear to have only attended one training session, this study has confounded the interaction term and therefore cannot assess interactions and it is debatable whether it is truly factorial		
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Logistic regression	At-the-margins	Linear regression/ANCOVA/ANOVA
		No mention of interaction term	Interaction term excluded as it was not statistically significant	Interaction term excluded as it was not statistically significant
		Used logistic regression, adjusting for clustering. Present OR for TI vs. not TI and TM vs. not TM	"The first analysis was tested for an interaction between substance and training. An interaction term between substance and training was added to the model. This interaction was not significant, hence the analysis focused on two separable trials, training vs. no training and lignocaine vs. cortisone."	Authors decided to present main effects not simple main effects as interaction was significant only at 6 months.

Author/study acronym		Salize/Twardella (240, 376)	SAPPHIRE (269, 281)	Thomas et al (234, 377)
	Base case economic evaluation	Inside-the-table (treating each cell separately)	Linear regression/ANCOVA/ANOVA	At-the-margins
			No interaction term included	
			Used seemingly-unrelated regression to allow for clustering, but conducted analyses on the two factors separately, without controlling for the other factor or including interaction terms. Bayesian analsyis in MLwiN. "As it would not be expected that whether or not a GP is trained in delivering a shoulder injection would impact on the relative cost effectiveness of alternative drugs (which is supported by the absence of statistically significant interactions between treatment and training [10]), we compared treatment separately from GP training"	Presented cost-effectiveness only for with/without exercise intervention
	Sensitivity analyses of economic evaluation	None	None	None
Presentation of economic results		Presented ICERs for TM vs. control and TI+TM vs. control and TI+TM vs. TM. Excluded TI arm as the prevalence of abstinence did not differ significantly from that in control arm	Compared costs and effects between all trained vs. all not trained and for all lignocaine vs. all cortisone; results for individual cells not presented	Present detailed costs and cost-effectiveness for exercise vs. no exercise. Present total cost of telephone intervention, but no detailed information or ICERs
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Pair-wise CEACs for TM vs. control and TI+TM vs. control and TI+TM vs. TM (based on bootstrapping)	Presented pair-wise CEACs for training vs. no training and for lignocaine vs. cortisone, but not multiple comparison CEACs	Pair-wise CEACs calculated using bootstrapping
Statistical significance of differences of each comparison individually		TI had no significant effect, but TM significantly increased the odds of cessation	No significant difference between lidocaine and cortisone or between training and no training	Exercise showed significant improvement. Telephone counselling and placebo had no significant effect
Significance of interaction effect		Not stated	Not significant	Significant quantitative interaction was observed between exercise and

Author/study acronym	Salize/Twardella (240, 376)	SAPPHIRE (269, 281)	Thomas et al (234, 377)
			telephone contact at 6 months, but not at any other timepoint; this isolated significant interaction was interpreted as being due to chance
Results of comparing different methods for statistical analysis	N/A	N/A	N/A

Author/study acronym	Thompson et al 1986 (237)	TIMI 3B (378, 379)	Tjan-Heijnen et al (380, 381)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation	1986	2003	2003
Year of publication for earliest paper presenting ***methods*** of economic evaluation	1986	2003	2003
Year of publication for earliest paper presenting ***results*** of economic evaluation	1986	2003	2003
Type of factorial design (e.g. partial/fractional)	Full factorial	full factorial	full factorial
Total sample size (across all arms)	507	1473	163
Size of factorial matrix (e.g. 2x2)	2x2x2 + 2	2x2	2x2
Disease area	Colorectal cancer screening	Angina and myocardial infarction	Lung cancer
Sponsor	US public funding body	US public funding body	Manufacturer
Country	US	US and Canada	Europe
Treatments (or treatment combinations) compared	Factor 1: Physician talk about test vs. nurse talk vs. no talk; the nurse talk comparison was added later and not fully crossed with others, such that double combinations of nurse talk were not considered. Factor 2: Reminder postcard vs. no postcard; Factor 3: Telephone follow-up to those not returning test in 10 days vs. no telephone follow-up.	Factor 1: Tissue plasminogen activator (TPA) vs. placebo. Factor 2: early invasive strategy (early coronary arteriography then revascularisation when appropriate) vs. early conservative strategy (coronary arteriography then revascularisation after initial medical therapy)	Factor 1: standard triple chemotherapy (CDE) vs. intensified CDE + G-CSF. Factor 2: antibiotics (ciprofloxacin+roxithromcyin) vs. placebo.
Type of comparisons	0 vs. a vs. b vs. ab	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)
Type(s) of treatment evaluated	Logistical intervention targetted at health service	Pharmaceutical and surgery	Pharmaceutical
Power calculation allowed for interaction effect and/or powered to test for interaction?	No power calculation	Not stated, although wording implies that power calculation was based on at-the-margins analysis and the study was not powered for subgroup analyses unless there were large	Not powered to detect interaction, although unclear whether this is because antibiotics vs. placebo comparison stopped early on ethical grounds.

Author/study acronym		Thompson et al 1986 (237)	TIMI 3B (378, 379)	Tjan-Heijnen et al (380, 381)
			differences	
Type of economic evaluation (e.g. CUA)		CCA	CEA	CEA
Other notable aspects of study design	Cluster randomised	no	No	No
	Cross-over	no	No	No
	Intentionally unbalanced	no	No	No
	Interventions target different diseases	no	no	yes
	Other			
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Inside-the-table (treating each cell separately)	At-the-margins	At-the-margins
			Interaction term excluded as it was not statistically significant	
		Results are presented inside-the-table, although at-the-margins analysis is used to test overall effect of physician talk and adjusted binary regression is also conducted (which appears to have been evaluated with dummies for each of the 9 active treatment groups)	Two-stage testing procedure: tested for interaction using Breslow-Day test; "If treatment by strategy interaction was not detected for the initial therapy end point, results were pooled across the two strategies to provide a single TPA vs. placebo group comparison. Likewise, if a treatment by strategy interaction was not detected for the strategy primary end point, results were pooled across the two treatment assignments (TPA and placebo) to provide a single early invasive strategy vs. early conservative strategy comparison.	Present both at-the-margins analysis and logistic regression (with interaction term omitted as it was not significant)
	Base case economic evaluation	At-the-margins	Logistic regression	At-the-margins
		No interaction term included	Unclear	
		Compare crude costs of all 4 interventions; conclude that postcard is best as it is cheapest and gives comparable results to other arms.	Logistic regression model was used to explore how outcomes and treatment effects varied between risk groups; economic evaluation	

Author/study acronym		Thompson et al 1986 (237)	TIMI 3B (378, 379)	Tjan-Heijnen et al (380, 381)
		Conduct more detailed costing analysis to assess incremental cost of postcard vs. no postcard (including savings from early detection) and show postcard to be cost-saving (and therefore dominant).	focussed only on invasive vs. conservative comparison and is highly unlikely to have included interactions between treatment strategies	
	Sensitivity analyses of economic evaluation	None	None	
Presentation of economic results		see above	Presented only incremental outcomes and costs for invasive and conservative management (at-the-margins presentation)	Simply compare costs and effects for antibiotics vs. placebo for Germany and Netherlands, concluding that antibiotics dominate placebo
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		None	None	threshold and 3-way sensitivity analyses
Statistical significance of differences of each comparison individually		Postcard and phone call alone are significantly better than control; when analysed at-the-margins, physician talk significantly increases compliance, with no signficant difference between nurse and physician talk.	No significant difference in clinical endpoints for either factor, although TPA-treated patients had significantly more Mis	Antibiotic chemotherapy significantly reduced incidenceof febrile leucopaenia; results for standard vs. intensive chemo not yet reported
Significance of interaction effect		Not stated; authors conclude effect is additive and that all treatments have equal effect based on comparison of crude compliance data.	Not significant	Non-significant qualitative interaction for incidence of febrile leukopenia (p=0.130) inferred from table 2, not discussed in paper.
Results of comparing different methods for statistical analysis		N/A	None	N/A

Author/study acronym	UK BEAM (244, 382)	UKPDS (211, 263, 282-288, 383)
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation	2004	1998
Year of publication for earliest paper presenting ***methods*** of economic evaluation	2004	1998
Year of publication for earliest paper presenting ***results*** of economic evaluation	2004	1998
Type of factorial design (e.g. partial/fractional)	Full factorial	Partial factorial
Total sample size (across all arms)	1,334	4,209
Size of factorial matrix (e.g. 2x2)	2x3	4x3x2
Disease area	Low back pain	Diabetes and cardiovascular disease
Sponsor	MRC	Various public and private organisations
Country	UK	UK
Treatments (or treatment combinations) compared	Factor 1: spinal manipulation in NHS clinic, spinal manipulation in private clinic, or best care with GP (no manipulation). Factor 2: exercise programme, vs. best care.	**Factor 1 (UKPDS):** Diet vs. insulin vs. sulphonylurea (vs. metformin in obese patients only); patients randomised to diet were randomised to active treatments if diet failed; patients randomised to sulphonylurea were randomised to receive chlorpropamide rather than glibenclamide (centres started pre-1987) or glipizide (centres started 1987-8); additional randomisations for 2nd line or add-in therapy made later on. **Factor 2 (HDS, hypertensive pts only):** Tight BP control (pts randomised to captopril OR atenolol) or less tight BP control. **Factor 3 (UKPDS Acarbose Study)**: Acarbose vs. placebo (in 1946 patients)
Type of comparisons	0 vs. a vs. b vs. ab	Other
Type(s) of treatment evaluated	Physiotherapy and related techniques	Pharmaceutical
Power calculation allowed for interaction effect and/or powered to test for interaction?	Not stated	Not stated
Type of economic evaluation (e.g.	CUA	CEA and CUA

Author/study acronym		UK BEAM (244, 382)	UKPDS (211, 263, 282-288, 383)
CUA)			
Other notable aspects of study design	Cluster randomised	No	No
	Cross-over	no	No
	Intentionally unbalanced	yes	Yes
	Interventions target different diseases	No	No
	Other	1:1:1:1 randomisation between GP care, exercise, manipulation and manipulation+exercise, with 50% of the pts randomised to the latter 2 categories being allocated to private rather than NHS care	Additional randomisations within different cells of the factorial design
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Linear regression/ANCOVA/ANOVA	At-the-margins
		Interaction term included as reached statistical significance	No mention of interaction term
		"We analysed the data in steps. Firstly, we used multilevel modelling to make allowance for the innate clustering of participants by centre, exercise class, manipulator, and practice. Secondly, we tested the effect of exercise without manipulation by comparing participants allocated to best care with those allocated to best care plus exercise (table 2). Thirdly, we tested the effect of manipulation without exercise by comparing participants allocated to best care with those allocated to best care plus manipulation (table 3). If this was significant, we tested for differences between manipulation in NHS and private premises. Finally, if either exercise or manipulation gave significant results, we tested for interactions between exercise and manipulation—that is, whether the estimated improvement in participants allocated to best care, manipulation, and exercise (table 4) differed significantly from the sum of the estimated improvement due to manipulation (table 3) and that due to exercise (table 2)."	For Factors 1 & 2: Cox regression used in papers 33, 34, 38 and 39, with each comparison analysed separately. For Factor 3: mean outcomes across factor 1 groups (at-the-margins) and results disaggregated by factor 1 allocation (inside-the-table) are both presented.
	Base case economic evaluation	Other	At-the-margins
		Interaction term included as reached statistical significance	No mention of interaction term
		"Because the accompanying clinical paper found statistically significant interactions between manipulation and exercise,1 it	40, 41, 51 compared mean costs & QALYs and calculated ICERs separately for a single

Author/study acronym		UK BEAM (244, 382)	UKPDS (211, 263, 282-288, 383)
		compares four distinct treatments within the factorial design.19 Although costs show no interaction between treatments, this paper also compares these four treatments for three reasons. Firstly, as costs vary much more than clinical outcomes, this is prudent. Secondly, those people responsible for allocating resources need to choose between these four treatments. Finally, this epitomises the Bayesian statistical approach adopted in this paper." Private vs. NHS manipulation comparison is ignored; in base case, all pts were assumed to have NHS care. Adjusted for clustering in economic evaluation using bivariate multi-level Bayesian analysis in MLwiN	comparison, ignoring all others.
	Sensitivity analyses of economic evaluation		At-the-margins, including event-based CEA conducted at-the-margins
		Conducted a sensitivity analysis using private costs for manipulation and one in which 50% of pts were assumed to use private premises	Various sensitivity analyses presented, although these all appear to be conducted at-the-margins
Presentation of economic results		NHS vs. private care comparison is ignored. Costs and effects presented for the 4 remaining arms in a inside-the-table approach. CE frontier was identified and ICERs calculated vs. next most costly treatment on frontier	40, 41, 51 reported mean costs QALYs and ICER separately for a single comparison, ignoring all others. 72 reported mean costs, QALYs and ICERs for each of the 3 main comparisons at-the-margins
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Presented CEACs for multiple comparisons and those for manipulation alone vs. best care and for exercise vs. best care	CEACs for pair-wise comparisons
Statistical significance of differences of each comparison individually		Exercise significantly superior to no exercise. Manipulation significantly superior to no manipulation. No significant difference between NHS and private manipulation	Intensive glucose control significantly reduced diabetic events in non-overweight (33) and overweight (34) patients. Tight BP control significantly reduced diabetic events (38). Captopril and atenolol were equally effective (39). Acarbose significantly reduced HbA1c (44).
Significance of interaction effect		Significant negative interaction between manipulation and exercise at 3 months, but not at 12 months. No interaction observed between treatments was observed for costs.	Not stated
Results of comparing different methods for statistical analysis		N/A	N/A

Author/study acronym		**VIP (238)**	**Wesson et al 2008 (384)**
Year of publication for earliest paper presenting ***methods or results*** of economic evaluation		2005	2008
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2005	2008
Year of publication for earliest paper presenting ***results*** of economic evaluation		2005	2008
Type of factorial design (e.g. partial/fractional)		full factorial	Full factorial
Total sample size (across all arms)		391	45 clinics
Size of factorial matrix (e.g. 2x2)		2x2	2x2
Disease area		Falls in people with visual impairment	Family planning
Sponsor		Other public funding body	Other public funding body
Country		New Zealand	Kenya
Treatments (or treatment combinations) compared		Factor 1: home safety programme vs. no home safety programme. Factor 2: exercise programme vs. no exercise programme. Control group received two social home visits.	Factor 1: Educational outreach intervention ("detailing") for clinic providers vs. no detailing for clinic providers. Factor 2: Detailing for community based distribution agents (CBD) vs. no detailing for CBD
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Physiotherapy and related techniques	Logistical intervention targetted at health service
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated	No - assumed no interaction
Type of economic evaluation (e.g. CUA)		CEA	CEA
Other notable aspects of study design	Cluster randomised	No	Yes
	Cross-over	No	No
	Intentionally unbalanced	No	No
	Interventions target different diseases	No	No
	Other		
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Poisson regression	Linear regression/ANCOVA/ANOVA
		Interaction term included as reached statistical significance	Not stated
		Since interaction was statistically significant, authors "presented pre-planned combined group comparisons plus appropriate single group comparisons"	Used linear regression although primary outcome was the percentage or absolute number of IUDs implanted. Authors comment that "The interaction effect of detailing at a study site for both target groups and the main effects of detailing at study

Author/study acronym		VIP (238)	Wesson et al 2008 (384)
			sites were examined for either providers or CBD agents." No other mention of interactions in the paper
	Base case economic evaluation	At-the-margins	Inside-the-table (treating each cell separately)
		Only presented cost-effectiveness for home safety vs. no home safety programme since exercise was not found to be effective. Appear to have based this on the at-the-margins estimate of the efficacy of home safety despite a qualitative interaction (ICER=total cost/difference in total falls with/without home safety programme)	Divided the total cost of the CBD+detailing intervention (plus the cost of the IUDs implanted) by the number of additional IUDs implanted compared with *baseline*. Compared this ICER with that for depot injections.
	Sensitivity analyses of economic evaluation	None	None
Presentation of economic results		Presented only cost of home safety programme and cost/fall prevented. Results for exercise programme were not presented as it was ineffective	Results only presented for CBD+detailing vs. pre-study baseline.
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		None besides calculation of SD for cost and one-way sensitivity analyses	None
Statistical significance of differences of each comparison individually		Home safety programme significantly reduced falls. Exercise programme non-significantly increased falls.	Not stated - compared each intervention with baseline
Significance of interaction effect		Statistically significant qualitative interaction for primary endpoint	Not stated
Results of comparing different methods for statistical analysis		For clinical endpoints, there is a very large qualitative interaction that means that exercise programme increases fall rate in presence of home safety programme and reduces it when used alone; as result, at-the-margins analysis finds exercise to be non-significantly harmful, although it reduces falls when used alone.	N/A

Table G2. Protocols and terminated studies

Author/study acronym	**ACCORD (232, 385)**	**AED study (386)**	**CENEX (387, 388)**
Year of publication for earliest paper presenting ***methods*** of economic evaluation	2007	2009	2007
Protocol or completed study?	Protocol	Protocol	Protocol
Anticipated publication date of ec anal (for protocols only)	Not stated	Not stated	Not stated
Type of factorial design (e.g. partial/fractional)	Partial factorial	Full factorial	Full factorial
Projected total sample size (across all arms)	4733	1179	2800
Size of factorial matrix (e.g. 2x2)	double 2x2	3x3	2x2
Disease area	Diabetes	Alcohol abuse	Nutrition and exercise in older people
Sponsor	US public funding body	Department of Health	Wellcome Trust
Country	US and Canada	UK	Chile
Treatments (or treatment combinations) compared	Factor 1 (all patients): Intensive glycaemic control targetting HbA1c<6.0%, vs. standard glycaemic control targetting HbA1c<7.0-7.9%. Factor 2 differed between groups: For the 5518 pts with moderate dyslipidemia, Factor 2 comprised fenofibrate vs. placebo. For the remaining 4733 pts, Factor 2 comprised intensive blood pressure therapy targeting systolic BP<120 mmHg, vs. standard BP therapy targeting systolic BP<140 mmHg.	Factor 1: screening method (M-SASQ vs. FAST vs. SIPS-PAT questionnaires). Factor 2: intervention (patient information leaflet [PIL]; brief advice [BA]; or brief lifestyle counselling [BLC])	Factor 1: Nutritional supplement vs. no supplement. Factor 2: exercise vs. no exercise
Type of comparisons	0 vs. a vs. b vs. ab	A1 vs. A2 vs. B1 vs. B2 (≥2 dosing regimens of ≥2 interventions)	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated	Pharmaceutical	Other	Nutritional supplement/diet + exercise
Power calculation allowed for interaction effect and/or powered to test for interaction?	Interactions not mentioned in main protocol, although protocol does mention that power calculations on each treatment assumed that the other 2 interventions would achieve the effect sizes for which they were	Not stated: comments that: "The sample size calculation is designed to account primarily for intervention level outcomes. Powering the study in this way will also account statistically for appropriate outcomes	No - assumes no interactions

Author/study acronym		ACCORD (232, 385)	AED study (386)	CENEX (387, 388)
		powered	for screening approach and screening method."	
Type of economic evaluation (e.g. CUA)		CEA and CUA	CEA and CUA	CEA and CUA
Other notable aspects of study design	Cluster randomised	No	Yes	Yes
	Cross-over	No	No	No
	Intentionally unbalanced	No	No	Yes
	Interventions target different diseases	No	No	Yes
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Cox regression	Logistic regression	Not stated
		No mention of interaction term	Not stated	No interaction term included
			Comment that "Due to the nested factorial nature of the study, we will use multilevel modelling to explore potential interactions between each of the levels nested within the trial"	
	Base case economic evaluation	At-the-margins	Unclear	Not stated
		Describe primary hypotheses as "main effects" and describe 3 hypotheses	"The economic analysis will calculate the incremental cost-effectiveness of the control condition with the AHW condition under study"	
	Sensitivity analyses of economic evaluation	Inside-the-table (treating each cell separately)	Not stated	Not stated
		Define two secondary hypotheses comparing AB group with 00 group for each half of the factorial design		
Presentation of economic results		Not stated	Not stated	Not stated
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Plan to use bootstrapping and calculate 95% CI around ICERs	Plan to construct CEACs showing "the probability that the brief intervention is more cost-effective than usual care"	Not stated

Author/study acronym		CJS (389)	Coombes (390)	IMPAC3T (391)
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2009	2009	2007
Protocol or completed study?		Protocol	Protocol	Protocol
Anticipated publication date of ec anal (for protocols only)		30 month trial began May 2008	Not stated	
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Projected total sample size (across all arms)		480	132	400 patients; 20 practices
Size of factorial matrix (e.g. 2x2)		2x3	2x2	2x2
Disease area		Alcohol abuse	Lateral epicondylagia	Lower respiratory tract infections
Sponsor		Department of Health	Other public funding body	Other public funding body
Country		UK	Australia	Netherlands
Treatments (or treatment combinations) compared		Factor 1: screening method (FAST or M-SASQ questionnaires). Factor 2: intensity of brief intervention (client information leaflet [CIL]; brief advice [BA]; or brief lifestyle counselling [BLC]).	Factor 1: corticosteroid injection vs. placebo. Factor 2: Physiotherapy vs. no physiotherapy	Factor 1: communication skills training for GPs. Factor 2: C-reactive protein test at point of care
Type of comparisons		A1 vs. A2 vs. B1 vs. B2 (≥2 dosing regimens of ≥2 interventions)	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Other	Pharmaceutical and physiotherapy	Logistical intervention targetted at health service
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated. Comments that "The sample size calculation is designed to account primarily for intervention level outcomes. Powering the study in this way will also account statistically for appropriate outcomes for screening approach and screening method."	No - "The power calculation is based on previous findings that […] the combination of the two will be additive, but not synergistic in effect"	Not stated, although authors comment that interactions can affect power
Type of economic evaluation (e.g. CUA)		CEA and CUA	CUA and CBA	Not stated
Other notable aspects of study	Cluster randomised	Yes	No	Yes

Author/study acronym		CJS (389)	Coombes (390)	IMPAC3T (391)
design	Cross-over	No	No	No
	Intentionally unbalanced	No	No	No
	Interventions target different diseases	No	No	No
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Logistic regression	Linear regression/ANCOVA/ANOVA	Linear regression/ANCOVA/ANOVA
		Not stated	Not stated	Interaction term included regardless of significance
		Comment that "Due to the nested factorial nature of the study, we will use multi-level modelling to explore potential interactions between each of the levels nested within the trial (screening method and intervention)"	"The outcomes measured at 4, 8, 12, 26 and 52 weeks will be used to generate efficacy and recurrence indices that will be analysed using linear mixed and logistic regression models. […] Alpha will be set at 0.01 to compensate for the possible increase in type I error rates that may result from multiple testing."	Patients clustered within GPs within practices
	Base case economic evaluation	Unclear	Not stated	Unclear
		"The economic analysis will calculate the incremental cost-effectiveness of the control condition with the AHW condition under study"		Protocol not clear on methods for economic evaluation: Secondary outcomes will be compared using regression analysis. Regarding economic evaluation, simply state that "The statistical uncertainty regarding the cost-effectiveness of the different strategies will be assessed using non-parametric bootstrap resampling techniques and result in estimates of net monetary benefit values"
	Sensitivity analyses of economic evaluation	Not stated	Not stated	Not stated
Presentation of economic results		N/A	N/A	N/A
Presentation of uncertainty (e.g.		Plan to construct CEACs showing	N/A	N/A

Author/study acronym	CJS (389)	Coombes (390)	IMPAC3T (391)
pair-wise CEAC, EVPI)	"the probability that the brief intervention is more cost-effective than usual care"		

Author/study acronym		MOA (392)	OPTIMA (393)	OPTIMAL (394)
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2009	2003	2009
Protocol or completed study?		Protocol	Protocol	Protocol
Anticipated publication date of ec anal (for protocols only)		1-year data anticipated in June 2010; 2-year data in 2011	Clinical results were published in 2011	
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Projected total sample size (across all arms)		224	1700 planned	340
Size of factorial matrix (e.g. 2x2)		2x2	2x2	2x2
Disease area		Osteoarthritis	HIV	Prolapse and stress incontinence symptoms
Sponsor		Other public funding body	Other UK public funding body	US public funding body
Country		New Zealand	US, Canada and UK	US
Treatments (or treatment combinations) compared		Factor 1: multi-modal exercise vs. no exercise. Factor 2: individualised manual therapy vs. no manual therapy	Factor 1: mega-ART (antiretroviral therapy) vs. standard ART. Factor 2: 3-month antiretroviral drug-free period (ARDFP).	Factor 1: sacrospinous ligament fixation vs. uterosacral vaginal vault suspension (two different surgical interventions). Factor 2: perioperative behavioural and pelvic floor muscle training vs. usual perioperative care
Type of comparisons		0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab	A0 vs. AX vs. B0 vs. BX (Factor 1: interventions A and B; Factor 2: X vs. 0)
Type(s) of treatment evaluated		Physiotherapy and related techniques	Pharmaceutical	Surgery and physiotherapy
Power calculation allowed for interaction effect and/or powered to test for interaction?		46% power to detect interaction; no interaction is expected	Unclear. Power calculation section in Kyriakides 2003 mentions only mega vs. standard, not ARDFP	Power calculation assumes no interaction and was based on at-the-margins analysis
Type of economic evaluation (e.g. CUA)		CUA	CUA	CUA
Other notable aspects of study design	Cluster randomised	No	No	No
	Cross-over	No	No	No
	Intentionally unbalanced	No	No	No
	Interventions	No	No	Yes

Author/study acronym		MOA (392)	OPTIMA (393)	OPTIMAL (394)
	target different diseases			
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Linear regression/ANCOVA/ANOVA	At-the-margins	Linear regression/ANCOVA/ANOVA
		Interactions will be ignored unless statistically significant		Not stated
		Secondary analyses will assess interaction by including interaction terms in the models. The coefficients for interaction effects and their 95% CI will be reported.	"Primary comparisons will be the main effects: standard ART vs. mega-ART and ARDFP vs. no ARDFP". "If the primary analyses of main effects show a significant difference between the strategies, secondary analyses will compare the pair-wise treatment strategy comparisons [adjusting for multiple tests]." "Qualitative interactions between the main effects (ART and ARDFP) are not anticipated. If, however, such interactions are found, then no main effects will be reported. Quantitative interactions, which are more likely to be present, will be assessed prior to the analysis of the main effects; these main effects will nonetheless be reported."	Logistic regression will be used for surgical success and GLMs will be used for outcomes for periopreative interventions. Both models will include both treatment assignments. Although it is not explicitly stated whether interactions will be taken into account, the protocol states that they will test for interactions.
	Base case economic evaluation	Not stated	Unclear	Linear regression/ANCOVA/ANOVA
		Only very brief details of economic evaluation given in protocol	Only brief details given	"The incremental cost-effectiveness ratio (ICER) will be calculated as the differential costs between the perioperative BPMT group and the usual care group, divided by the differential QALYs between the two groups. General linear regression analysis will be conducted to estimate the differential mean costs and differential mean QALYs between the BPMT group and the usual care group, while adjusting for surgical treatment (SSLF or ULS), hysterectomy (concomitant or prior), surgeon, and

Author/study acronym		MOA (392)	OPTIMA (393)	OPTIMAL (394)
				other baseline measures identified as unbalanced between the two groups. The measures of costs and QALYs may be transformed before analysis if the variables demonstrate skewed distributions."
	Sensitivity analyses of economic evaluation	Not stated		Not stated
Presentation of economic results		Not stated		Cost-effectiveness of perioperative behavioural intervention vs. usual care will be assessed based on regression analysis. Cost-effectiveness of surgical interventions does not appear to be assessed
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Not stated	Protocol indicates that authors plan to use bootstrapping and calculate CEACs. Only brief details given, although the comment that CEACs plot "the probability of one strategy being more cost-effective than the other two" suggests that the authors plan to generate CEACs for multiple comparisons, but ignore one of the 4 arms.	Not stated. 95% CI around ICER will be based on bootstrapping

Author/study acronym		PHC study (395)	PORTSSS (396)	SIGNET (397)
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2009	2009	2007
Protocol or completed study?		Protocol	Protocol	Protocol
Anticipated publication date of ec anal (for protocols only)		30 month trial began April 2008	Not stated	Summer 2009
Type of factorial design (e.g. partial/fractional)		Full factorial	Full factorial	Full factorial
Projected total sample size (across all arms)		744	2576	500
Size of factorial matrix (e.g. 2x2)		3x2x2	2x2	2x2
Disease area		Alcohol abuse	Smoking cessation	Parenteral nutrition
Sponsor		Department of Health	Department of Health	MRC
Country		UK	UK	UK
Treatments (or treatment combinations) compared		Factor 1: intensity of brief intervention (patient information leaflet [PIL]; brief advice [BA]; and brief lifestyle counselling [BLC]). Factor 2: screening approach (targeted vs. universal screening). Factor 3: screening method (FAST or M-SASQ questionnaires)	Factor 1: Proactive vs. reactive support. Factor 2: the offer of NRT vs. no offer.	Factor 1: glutamine vs. no additional glutamine. Factor 2: selenium vs. no additional selenium
Type of comparisons		A1 vs. A2 vs. B1 vs. B2 (≥2 dosing regimens of ≥2 interventions)	0 vs. a vs. b vs. ab	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Other	Pharmaceutical + other	Nutritional supplement
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated. Comments that "The sample size calculation was designed to account primarily for intervention level outcomes. Powering the study in this way will also account statistically for appropriate outcomes for screening approach and screening method."	No - assumes no interactions based on previous evidence	Not stated - protocol comments that "evidence will be sought for an interaction between glutamine and selenium" but power calculation appears to be based on at-the-margins approach with no interaction.
Type of economic evaluation (e.g. CUA)		CEA and CUA	CEA and CUA	CUA
Other notable aspects of	Cluster randomised	Yes	No	No
	Cross-over	No	No	No
	Intentionally	No	No	No

Author/study acronym		PHC study (395)	PORTSSS (396)	SIGNET (397)
study design	unbalanced			
	Interventions target different diseases	No	No	No
Form of statistical analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	For primary clinical outcome	Logistic regression	At-the-margins	At-the-margins
		Not stated	Interactions will be ignored unless statistically significant	Not stated
		Comment that "we will use multi-level modelling to explore potential interactions between each of the three levels nested within the trial (screening approach, screening method and intervention)"		The first two "principal comparisons" comprised at-the-margins comparisons, while the third comprised "all randomised glutamine and selenium vs. all randomised glutamine alone vs. all randomised selenium alone vs. all randomised placebo". Authors stated that "Dichotomous endpoints will be analysed using statistical tests such as the Chi-squared test"
	Base case economic evaluation	Unclear	Not stated	Not stated
		"The economic analysis will calculate the incremental cost-effectiveness of the control condition with each of the screening and brief intervention conditions under study"	In addition to within trial analysis (cost/additional quitter), plan to use model to calculate lifetime cost/QALY	Outline of economic evaluation in the protocol simply says: “The type of economic evaluation performed will depend upon the findings
	Sensitivity analyses of economic evaluation	Not stated	Not stated	Not stated
Presentation of economic results		N/A	N/A	N/A
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		Plan to construct CEACs showing "the probability that the brief intervention is more cost-effective than usual care"	State that plan to conduct PSA, but give no other details	N/A

Author/study acronym		**STOOL (233)**	**ZAMSTAR (398)**
Year of publication for earliest paper presenting ***methods*** of economic evaluation		2008	2008
Protocol or completed study?		Terminated early, so no results; treated as protocol	Protocol
Anticipated publication date of ec anal (for protocols only)		N/A	Not stated
Type of factorial design (e.g. partial/fractional)		Incomplete factorial	Full factorial
Projected total sample size (across all arms)		1002 planned; 19 recruited; 7 completed trial taking intervention as allocated	24 communities covering approx 120,000 individuals
Size of factorial matrix (e.g. 2x2)		3x3	2x2
Disease area		Constipation	Tuberculosis and HIV
Sponsor		NIHR HTA	Other public funding body
Country		UK	Zambia and South Africa
Treatments (or treatment combinations) compared		Factor 1: bulk laxative, stimulant laxative or osmotic laxative. Factor 2: adding in either a bulk laxative, stimulant laxative or osmotic laxative. Cells of the design where patients would have added in the same type of drug as used first-line were omitted from the design.	Factor 1: enhanced case finding intervention. Factor 2: household intervention
Type of comparisons		Other	0 vs. a vs. b vs. ab
Type(s) of treatment evaluated		Pharmaceutical	Other
Power calculation allowed for interaction effect and/or powered to test for interaction?		Not stated	No: Powered based on no interaction on multiplicative scale and not powered to detect interactions
Type of economic evaluation (e.g. CUA)		Planned to conduct CMA if no "statistically significant evidence that one strategy is more effective than another", and otherwise CCA and CEA; CUA was also mentioned in Appendix 5	Not stated
Other notable aspects of study design	Cluster randomised	No	yes
	Cross-over	No	No
	Intentionally unbalanced	No	No
	Interventions target different diseases	No	No
	Other		Both interventions primarily target TB but also increase awareness of HIV
Form of statistical	For primary clinical	Poisson regression	Linear regression/ANCOVA/ANOVA

Author/study acronym		STOOL (233)	ZAMSTAR (398)
analysis to deal with factorial design (e.g. inside-the-table): including type of regression model if appropriate	outcome	No mention of interaction term	Not stated
			GLMs on log-incidence that allow for clustering. No clear indication of whether interaction terms will be included
	Base case economic evaluation	Not stated	Not stated
			Not stated: methods indicate that costs will be evaluated using linear regression and that cost-effectiveness will be assessed, but methods not stated
	Sensitivity analyses of economic evaluation	None	Not stated
Presentation of economic results		N/A	N/A
Presentation of uncertainty (e.g. pair-wise CEAC, EVPI)		N/A	Not stated

Appendix H: Data inputs, additional results and Stata code for the simulation analyses described in Chapter 5

Data inputs used in the simulations described in Chapter 5

Table H1. Data inputs for each study included in the simulation. The data inputs shown in this table were kept constant across all simulated datasets from the same trial.

Study	Constant: cost	dA: cost	dB: cost	Type of cost distribution	SD: cost	Units for health benefits	Constant: health	dA: health	dB: health	SD: health	Max health units	Correlation between costs and health
AFIST II (245)	£27,026	£6,842	-£4,301	Heteroskedastic	SDMR A: 1.398; SDMR B: 0.609	% pts with atrial fibrillation§	0.380§	0.020§	-0.100§	0.449	1	0.30*
ATEAM (246)	£54	£204	£100	Heteroskedastic	SDMR A: 2.617; SDMR B: 1.070	QALY	0.700	-0.010	0.040	0.087	1	-0.10*
Boyle (235)	£493	£26	£407	Ratio	SDMR: 0.249	Language score	1.590	2.910	0.860	7.365	50‡	0.05*
Hollis 2007 (247)	£67	£40	£126	Ratio	SDMR: 0.349	% smoking abstinence	0.117	0.021	0.054	0.360	1	0.15*
NIMH-MTA (242)	£975	£4	£5,158	Homoskedastic	SD: 917	Functioning Score§	-0.600§	-0.320§	-0.100§	0.898	15‡	0.30*
UKBEAM (244)	£346	£140	£195	Ratio	SDMR: 1.509	QALY	0.618	0.017	0.041	0.239	1	-0.10*
1-year IVAN results (255)	£9,663¥	-£3,159	-£7,864	Ratio	SDMR: 0.341	QALY	0.824	0.005	0.002	0.175	1	-0.03†
KAT mobile (Chapter 7)	£8,481	£2,619	£688	Homoskedastic	SD: 6,845	QALY	5.029	-0.297	-0.070	2.829	8.61	-0.28†
KAT metal (Chapter 7)	£8,085	-£303	-£252	Homoskedastic	SD: 2,826	QALY	5.569	-0.593	-0.523	2.476	8.61	-0.27†

Abbreviations: AFIST, Atrial Fibrillation Suppression Trial; ATEAM, Alexander Technique, Exercise and Massage; dA [dB], simple effect of treatment A [B] $(\mu_a - \mu_0)$; NIMH-MTA, National Institute of Mental Health Multimodal Treatment study of children with ADHD; SDMR, ratio of mean divided by SD; UKBEAM, United Kingdom Back pain Exercise And Manipulation.

* Hypothetical figure informed by the clinical area and any available scatter graphs.

† Estimated from primary data.

‡ Arbitrary limit chosen to be well above means for each group, since no maximum score was indicated in the trial reports.
§ AFIST and NIMH-MTA measured health benefits on scales in which high values indicate worse scores rather than better scores. The trials were simulated and mixed models were estimated using the actual scores (i.e. the proportion of patients with atrial fibrillation and negative functioning scores). However, to ensure that NMB was calculated correctly, model coefficients were transformed before NMB was calculated or criteria for assessing interactions were applied. For both trials, all interaction terms and treatment effects were subtracted from 0 to change the sign. Constant terms for NIMH-MTA were subtracted from zero, while those for AFIST were subtracted from 1. This transformation has no effect on AIC, BIC, SEs, degrees of freedom or p-values, so these outcomes were not adjusted.
¥ For IVAN, the interaction between drug and treatment regimen is so large that simply setting the interaction to 0 generates negative costs (Chapter 7). To avoid this arising within the simulated data, the constant term was increased by £2,000 in the two scenarios with zero interaction.

Table H2. Data inputs used for each simulated version of each study.

Scenario Name	c_Int	q_Int	Patients /arm	Best Tx*	Interaction:effect ratio†		
					Cost	QALYs	NMB*
AFIST	-£10,870	-14%	40	*ab*	-1.589	-7.000	-1.667
AFIST_DoubleInt	-£21,740	-28%	40	*ab*	-3.177	-14.000	-3.333
AFIST_HalfInt	-£5,435	-7%	40	*b*	-0.794	-3.500	-0.833
AFIST_DoubleSize	-£10,870	-14%	80	*ab*	-1.589	-7.000	-1.667
AFIST_NoInt	£0	0%	40	*b*	0.000	0.000	0.000
AFIST_NoInt_DoubleSize	£0	0%	80	*b*	0.000	0.000	0.000
ATEAM	-£91	-0.02	73	*b*	-0.909	-0.500	-0.442
ATEAM_DoubleInt	-£182	-0.04	73	*b*	-1.818	-1.000	-0.883
ATEAM_HalfInt	-£45	-0.01	73	*b*	-0.454	-0.250	-0.221
ATEAM_DoubleSize	-£91	-0.02	146	*b*	-0.909	-0.500	-0.442
ATEAM_NoInt	£0	0	73	*b*	0.000	0.000	0.000
ATEAM_NoInt_DoubleSize	£0	0	146	*b*	0.000	0.000	0.000
Boyle	£218	-2.04	31	*a*	8.385	-2.372	-2.676
Boyle_DoubleInt	£436	-4.08	31	*a*	16.769	-4.744	-5.352
Boyle_HalfInt	£109	-1.02	31	*a*	4.192	-1.186	-1.338
Boyle_DoubleSize	£218	-2.04	62	*a*	8.385	-2.372	-2.676
Boyle_NoInt	£0	0	31	*ab*	0.000	0.000	0.000
Boyle_NoInt_DoubleSize	£0	0	62	*ab*	0.000	0.000	0.000
Hollis	£9	1%	793	*ab*	0.225	0.429	0.554
Hollis_DoubleInt	£18	2%	793	*ab*	0.450	0.857	1.108
Hollis_HalfInt	£5	0%	793	*ab*	0.113	0.214	0.277
Hollis_DoubleSize	£9	1%	1586	*ab*	0.225	0.429	0.554
Hollis_NoInt	£0	0%	793	*ab*	0.000	0.000	0.000
Hollis_NoInt_DoubleSize	£0	0%	1586	*ab*	0.000	0.000	0.000
NIMH-MTA	£927	-0.16	35	*0*	231.750	-1.600	0.079
NIMH-MTA_DoubleInt	£1,854	-0.32	35	*0*	463.500	-3.200	0.158
NIMH-MTA_HalfInt	£464	-0.08	35	*0*	115.875	-0.800	0.040
NIMH-MTA_DoubleSize	£927	-0.16	70	*0*	231.750	-1.600	0.079
NIMH-MTA_NoInt	£0	0	35	*0*	0.000	0.000	0.000
NIMH-MTA_NoInt_DoubleSize	£0	0	70	*0*	0.000	0.000	0.000
UKBEAM	-£210	-0.025	322	*b*	-1.500	-1.471	-1.450
UKBEAM_DoubleInt	-£420	-0.05	322	*b*	-3.000	-2.941	-2.900
UKBEAM_HalfInt	-£105	-0.0125	322	*ab*	-0.750	-0.735	-0.725
UKBEAM_DoubleSize	-£210	-0.025	644	*b*	-1.500	-1.471	-1.450
UKBEAM_NoInt	£0	0	322	*ab*	0.000	0.000	0.000
UKBEAM_NoInt_DoubleSize	£0	0	644	*ab*	0.000	0.000	0.000
IVAN	£2,861	-0.015	153	*ab*	-0.906	-7.500	-0.970
IVAN_DoubleInt	£5,722	-0.03	153	*b*	-1.811	-15.000	-1.940
IVAN_HalfInt	£1,431	-0.0075	153	*ab*	-0.453	-3.750	-0.485
IVAN_DoubleSize	£2,861	-0.015	306	*ab*	-0.906	-7.500	-0.970
IVAN_NoInt	£0	0	153	*ab*	0.000	0.000	0.000
IVAN_NoInt_DoubleSize	£0	0	306	*ab*	0.000	0.000	0.000
KATmobile	-£2,720	0.897	48	*ab*	-3.953	-12.814	-9.895
KATmobile_DoubleInt	-£5,440	1.794	48	*ab*	-7.907	-25.629	-19.789
KATmobile_HalfInt	-£1,360	0.4485	48	*0*	-1.977	-6.407	-4.947
KATmobile_DoubleSize	-£2,720	0.897	96	*ab*	-3.953	-12.814	-9.895
KATmobile_NoInt	£0	0	48	*0*	0.000	0.000	0.000
KATmobile_NoInt_DoubleSize	£0	0	96	*0*	0.000	0.000	0.000
KATmetal	£506	1.065	36	*0*	-2.008	-2.036	-2.037
KATmetal_DoubleInt	£1,012	2.13	36	*ab*	-4.016	-4.073	-4.074
KATmetal_HalfInt	£253	0.5325	36	*0*	-1.004	-1.018	-1.019
KATmetal_DoubleSize	£506	1.065	72	*0*	-2.008	-2.036	-2.037
KATmetal_NoInt	£0	0	36	*0*	0.000	0.000	0.000
KATmetal_NoInt_DoubleSize	£0	0	72	*0*	0.000	0.000	0.000

Abbreviations: AFIST, Atrial Fibrillation Suppression Trial; ATEAM, Alexander Technique, Exercise and Massage; c_Int, interaction for cost; NIMH-MTA, National Institute of Mental Health Multimodal Treatment study of children with ADHD; q_Int, interaction for health benefits; Tx, treatment-combination; UKBEAM, United Kingdom Back pain Exercise And Manipulation.

* NMB was calculated at a £20,000/QALY ceiling ratio; a £5,000 (or $5,000) per unit ceiling ratio was used for all other measures of health benefit.

† The interaction:effect ratio (IER_{AB}) was calculated as the ratio of the interaction (I_{AB}) divided by the main effect of A (δ_A): i.e. as $IER_{AB} = I_{AB} / \delta_A = (\mu_0 - \mu_a - \mu_b + \mu_{ab}) / (\mu_a - \mu_0)$, where the factor defined as A is chosen such that $|\mu_a - \mu_0| < |\mu_b - \mu_0|$ in the case of non-mixed interactions and such that δ_A has the opposite sign to I_{AB} in the case of mixed interactions.

Stata code used in the simulations described in Chapter 5

```
import excel "Recalculating results to assess interactions", sheet("Matrix for Stata") firstrow
sum NoPts
local I=r(N)
mkmat Study c_cons c_A c_B c_Int c_DistType c_sd c_SDOverMean c_sdOverMeanB0 c_sdOverMeanB1 Outcome q_cons q_A q_B q_Int q_sd q_limit rho
NoPts, matrix(Data)
clear

local N=300 //Number of datasets to generate with each combination of interaction terms across all sample sizes
forvalues n=1(1)`N' {
forvalues i=1(1)`I' {
display "Dataset `i', Run `n'"
quietly {
clear

local w=1
local variable "Study c_cons c_A c_B c_Int c_DistType c_sd c_SDOverMean c_sdOverMeanB0 c_sdOverMeanB1 Outcome q_cons q_A q_B q_Int q_sd
q_limit rho NoPts"
foreach v of local variable {
        local `v' = Data[`i',`w']
        local w=`w'+1
}

//GENERATING A DATASET
local NoPtsTotal = `NoPts'*4
set obs     `NoPtsTotal' //Total number of patients in trial

generate ptid=_n
generate treatment=ceil(ptid/`NoPts')
generate a=0
replace  a=1 if treatment==2 | treatment==4
generate b=0
replace  b=1 if treatment==3 | treatment==4
generate interaction = a*b
generate mean_cost = `c_cons' + a*`c_A' + b*`c_B' + interaction*`c_Int'
if (`c_DistType'==1) generate sd_cost   = `c_sd' //Homoskedastic
if (`c_DistType'==2) generate sd_cost   = `c_SDOverMean'*mean_cost //Ratio.
if (`c_DistType'==3) generate sd_cost   = `c_sdOverMeanB0'*mean_cost if b==0 //Heteroskedastic ratio
```

```
if (`c_DistType'==3) replace  sd_cost   = `c_sdOverMeanB1'*mean_cost if b==1 //Heteroskedastic ratio

generate cost  = rgamma((mean_cost/sd_cost)^2,(sd_cost^2)/mean_cost)
generate rand_qalys = rnormal()
generate rand_cost_standard=(cost-mean_cost)/sd_cost
generate mean_qalys = `q_cons' + a*`q_A' + b*`q_B' + interaction*`q_Int'
generate qalys      = `q_cons' + a*`q_A' + b*`q_B' + interaction*`q_Int' + `q_sd'*(rand_cost_standard*`rho' + rand_qalys*(1-`rho'^2)^0.5)
replace qalys = `q_limit' if qalys>`q_limit' //Imposing a ceiling effect on QALYs

sum cost qalys if treatment==1 //Means should equal c_cons                    SD should equal c_sd
sum cost qalys if treatment==2 //Means should equal c_cons + c_A              SD should equal c_sd
sum cost qalys if treatment==3 //Means should equal c_cons + c_B              SD should equal c_sd
sum cost qalys if treatment==4 //Means should equal c_cons + c_A + c_B + c_Int; SD should equal c_sd
correl cost qalys //Correlation should equal rho
regress cost  a b interaction //coefficients should equal parameters set above; SE for constant should equal SD/sqrt(n*4), SE for A & B should be
SD/sqrt(n*2), SE for interaction should be SD/sqrt(n)
regress qalys a b interaction //coefficients should equal parameters set above

rename qalys outcome1
rename cost outcome2
reshape long outcome, i(ptid) j(cost)
replace cost=cost-1 //Variable called cost has been generated in the reshape process, but is currently equal to 1 if outcome=health, and we want it to equal 1
if outcome=cost
tab cost
sum outcome if cost==1
sum outcome if cost==0

generate A_cost = cost* a
generate B_cost = (cost)* b
generate Interaction_cost = (cost)* interaction
generate A_QALY = (1-cost)* a
generate B_QALY = (1-cost)* b
generate Interaction_QALY = (1-cost)* interaction
generate QALYconstant=1-cost

generate cost_noint = 0
generate qaly_noint = 0
```

```
xtmixed outcome cost A_cost B_cost cost_noint       QALYconstant A_QALY B_QALY qaly_noint      , noconstant || ptid: , noconstant residuals(unstructured,
t(cost))
        matrix coefficients_aux =  e(b)
        matrix coleq coefficients_aux =""
        matrix colnames coefficients_aux = C_cons_cN_qN   C_A_cN_qN   C_B_cN_qN   C_Int_cN_qN   Q_cons_cN_qN   Q_A_cN_qN   Q_B_cN_qN
Q_Int_cN_qN   C_SD_cN_qN   Q_SD_cN_qN   CQ_SD_cN_qN
        scalar C_Int_test_cN_qN = coefficients_aux[1, 4]
        scalar Q_Int_test_cN_qN = coefficients_aux[1, 8]
        matrix V=e(V)
        scalar C_Int_SE_cN_qN=sqrt(V[4,4])
        scalar Q_Int_SE_cN_qN=sqrt(V[8,8])
        scalar C_A_SE_cN_qN=sqrt(V[2,2])
        scalar Q_A_SE_cN_qN=sqrt(V[6,6])
        scalar C_B_SE_cN_qN=sqrt(V[3,3])
        scalar Q_B_SE_cN_qN=sqrt(V[7,7])
        matrix Int         = C_Int_test_cN_qN,  C_Int_SE_cN_qN, C_A_SE_cN_qN, C_B_SE_cN_qN, Q_Int_test_cN_qN,  Q_Int_SE_cN_qN,
Q_A_SE_cN_qN, Q_B_SE_cN_qN
        matrix colnames Int = C_Int_test_cN_qN   C_Int_SE_cN_qN  C_A_SE_cN_qN  C_B_SE_cN_qN  Q_Int_test_cN_qN   Q_Int_SE_cN_qN
Q_A_SE_cN_qN  Q_B_SE_cN_qN
        matrix LL = e(df_m), e(ll), e(r2_a)
        matrix colnames LL = dfm_cN_qN LL_cN_qN R2Adj_cN_qN
        matrix coefficients_all =  coefficients_aux,  Int, LL

xtmixed outcome cost A_cost B_cost Interaction_cost QALYconstant A_QALY B_QALY qaly_noint      , noconstant || ptid: , noconstant
residuals(unstructured, t(cost))
        matrix coefficients_aux =  e(b)
        matrix coleq coefficients_aux =""
        matrix colnames coefficients_aux = C_cons_cI_qN   C_A_cI_qN   C_B_cI_qN   C_Int_cI_qN   Q_cons_cI_qN   Q_A_cI_qN   Q_B_cI_qN
Q_Int_cI_qN   C_SD_cI_qN   Q_SD_cI_qN   CQ_SD_cI_qN
        scalar C_Int_test_cI_qN = coefficients_aux[1, 4]
        scalar Q_Int_test_cI_qN = coefficients_aux[1, 8]
        matrix V=e(V)
        scalar C_Int_SE_cI_qN=sqrt(V[4,4])
        scalar Q_Int_SE_cI_qN=sqrt(V[8,8])
        scalar C_A_SE_cI_qN=sqrt(V[2,2])
        scalar Q_A_SE_cI_qN=sqrt(V[6,6])
        scalar C_B_SE_cI_qN=sqrt(V[3,3])
        scalar Q_B_SE_cI_qN=sqrt(V[7,7])
```

```
        matrix Int          = C_Int_test_cI_qN,  C_Int_SE_cI_qN, C_A_SE_cI_qN, C_B_SE_cI_qN, Q_Int_test_cI_qN,  Q_Int_SE_cI_qN, Q_A_SE_cI_qN,
Q_B_SE_cI_qN
        matrix colnames Int = C_Int_test_cI_qN   C_Int_SE_cI_qN  C_A_SE_cI_qN  C_B_SE_cI_qN  Q_Int_test_cI_qN   Q_Int_SE_cI_qN  Q_A_SE_cI_qN
Q_B_SE_cI_qN
        matrix LL = e(df_m), e(ll), e(r2_a)
        matrix colnames LL = dfm_cI_qN LL_cI_qN R2Adj_cI_qN
        matrix coefficients_all =  coefficients_all , coefficients_aux, Int, LL

xtmixed outcome cost A_cost B_cost cost_noint       QALYconstant A_QALY B_QALY Interaction_QALY, noconstant || ptid: , noconstant
residuals(unstructured, t(cost))
        matrix coefficients_aux =  e(b)
        matrix coleq coefficients_aux =""
        matrix colnames coefficients_aux = C_cons_cN_qI   C_A_cN_qI   C_B_cN_qI   C_Int_cN_qI   Q_cons_cN_qI   Q_A_cN_qI   Q_B_cN_qI
Q_Int_cN_qI   C_SD_cN_qI   Q_SD_cN_qI   CQ_SD_cN_qI
        scalar C_Int_test_cN_qI = coefficients_aux[1, 4]
        scalar Q_Int_test_cN_qI = coefficients_aux[1, 8]
        matrix V=e(V)
        scalar C_Int_SE_cN_qI=sqrt(V[4,4])
        scalar Q_Int_SE_cN_qI=sqrt(V[8,8])
        scalar C_A_SE_cN_qI=sqrt(V[2,2])
        scalar Q_A_SE_cN_qI=sqrt(V[6,6])
        scalar C_B_SE_cN_qI=sqrt(V[3,3])
        scalar Q_B_SE_cN_qI=sqrt(V[7,7])
        matrix Int          = C_Int_test_cN_qI,  C_Int_SE_cN_qI,  C_A_SE_cN_qI,  C_B_SE_cN_qI, Q_Int_test_cN_qI,  Q_Int_SE_cN_qI, Q_A_SE_cN_qI,
Q_B_SE_cN_qI
        matrix colnames Int = C_Int_test_cN_qI   C_Int_SE_cN_qI   C_A_SE_cN_qI   C_B_SE_cN_qI  Q_Int_test_cN_qI   Q_Int_SE_cN_qI  Q_A_SE_cN_qI
Q_B_SE_cN_qI
        matrix LL = e(df_m), e(ll), e(r2_a)
        matrix colnames LL = dfm_cN_qI LL_cN_qI R2Adj_cN_qI
        matrix coefficients_all =  coefficients_all , coefficients_aux, Int, LL

xtmixed outcome cost A_cost B_cost Interaction_cost QALYconstant A_QALY B_QALY Interaction_QALY, noconstant || ptid: , noconstant
residuals(unstructured, t(cost))
        matrix coefficients_aux =  e(b)
        matrix coleq coefficients_aux =""
        matrix colnames coefficients_aux = C_cons_cI_qI   C_A_cI_qI   C_B_cI_qI   C_Int_cI_qI   Q_cons_cI_qI   Q_A_cI_qI   Q_B_cI_qI   Q_Int_cI_qI
C_SD_cI_qI   Q_SD_cI_qI   CQ_SD_cI_qI
        scalar C_Int_test_cI_qI = coefficients_aux[1, 4]
        scalar Q_Int_test_cI_qI = coefficients_aux[1, 8]
```

```
		matrix V=e(V)
		scalar C_Int_SE_cl_ql=sqrt(V[4,4])
		scalar Q_Int_SE_cl_ql=sqrt(V[8,8])
		scalar C_A_SE_cl_ql=sqrt(V[2,2])
		scalar Q_A_SE_cl_ql=sqrt(V[6,6])
		scalar C_B_SE_cl_ql=sqrt(V[3,3])
		scalar Q_B_SE_cl_ql=sqrt(V[7,7])
		matrix Int	= C_Int_test_cl_ql,  C_Int_SE_cl_ql,  C_A_SE_cl_ql, C_B_SE_cl_ql, Q_Int_test_cl_ql,  Q_Int_SE_cl_ql, Q_A_SE_cl_ql,
Q_B_SE_cl_ql
		matrix colnames Int = C_Int_test_cl_ql   C_Int_SE_cl_ql   C_A_SE_cl_ql  C_B_SE_cl_ql  Q_Int_test_cl_ql   Q_Int_SE_cl_ql Q_A_SE_cl_ql
Q_B_SE_cl_ql
		matrix LL = e(df_m), e(ll), e(r2_a)
		matrix colnames LL = dfm_cl_ql LL_cl_ql R2Adj_cl_ql
		matrix coefficients_all =  coefficients_all , coefficients_aux, Int, LL

clear
svmat coefficients_all, names(col)
generate Dataset = `i'
generate run = `n'
local input = "NoPts rho c_cons c_A c_B c_Int c_sd q_cons q_A q_B q_Int q_sd"
foreach v of local input {
		generate `v'=``v''
}
if (`i'==53 & `n'==1) save "Allcoefficients.dta", replace
if (`i'!=53 | `n'!=1) append using "Allcoefficients.dta"
if (`i'!=53 | `n'!=1) save "Allcoefficients.dta", replace
} //quietly
} //for i
} //for n



tab Dataset computer

order rho , first
order c_sd , first
order c_cons , first
order c_B , first
order c_A , first
```

```
order q_sd , first
order q_cons , first
order q_B , first
order q_A , first
order q_Int , first
order c_Int , first
order NoPts , first
order run, first
order Dataset, first
sort Dataset NoPts run

//Adjustment for outcomes in AFIST II and MTA where higher outcome numbers are bad. AFIST was analysed as % pts with AF & want to report as % pts without AF. MTA was analysed as negative numbers, but want to analyse as positive ones
sum q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI q_A  q_B  q_Int Q_Int_cN_qN Q_A_cN_qN  Q_B_cN_qN  Q_Int_test_cN_qN Q_Int_cN_qI Q_A_cN_qI  Q_B_cN_qI  Q_Int_test_cN_qI Q_Int_cI_qN Q_A_cI_qN  Q_B_cI_qN  Q_Int_test_cI_qN Q_Int_cI_qI Q_A_cI_qI  Q_B_cI_qI Q_Int_test_cI_qI if Dataset >=19 & Dataset<=24
sum q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI q_A  q_B  q_Int Q_Int_cN_qN Q_A_cN_qN  Q_B_cN_qN  Q_Int_test_cN_qN Q_Int_cN_qI Q_A_cN_qI  Q_B_cN_qI  Q_Int_test_cN_qI Q_Int_cI_qN Q_A_cI_qN  Q_B_cI_qN  Q_Int_test_cI_qN Q_Int_cI_qI Q_A_cI_qI  Q_B_cI_qI Q_Int_test_cI_qI if Dataset >=25 & Dataset<=30
sum q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI q_A  q_B  q_Int Q_Int_cN_qN Q_A_cN_qN  Q_B_cN_qN  Q_Int_test_cN_qN Q_Int_cN_qI Q_A_cN_qI  Q_B_cN_qI  Q_Int_test_cN_qI Q_Int_cI_qN Q_A_cI_qN  Q_B_cI_qN  Q_Int_test_cI_qN Q_Int_cI_qI Q_A_cI_qI  Q_B_cI_qI Q_Int_test_cI_qI if Dataset <19  | Dataset>30
local outcome="q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI"
foreach o of local outcome {
        replace `o'=1-`o' if Dataset >=19 & Dataset<=24 //AFIST: Constants are set to be 100% - % with AF
        replace `o'=0-`o' if Dataset >=25 & Dataset<=30 //MTA:   Constants are set to be  - X
}
local outcome="q_A  q_B  q_Int"
foreach o of local outcome {
        replace `o'=0-`o' if Dataset >=19 & Dataset<=30 //AFIST & MTA: Sign is changed on interaction terms & all treatment effects
}
local outcome="Q_Int Q_A  Q_B  Q_Int_test"
foreach o of local outcome {
        local model="_cN_qN   _cI_qN   _cN_qI   _cI_qI"
        foreach m of local model {
                replace `o'`m'=0-`o'`m' if Dataset >=19 & Dataset<=30 //AFIST & MTA: Sign is changed on interaction terms & all treatment effects
        }
}
```

```
sum q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI q_A  q_B  q_Int Q_Int_cN_qN Q_A_cN_qN  Q_B_cN_qN  Q_Int_test_cN_qN
Q_Int_cN_qI Q_A_cN_qI  Q_B_cN_qI  Q_Int_test_cN_qI Q_Int_cI_qN Q_A_cI_qN  Q_B_cI_qN  Q_Int_test_cI_qN Q_Int_cI_qI Q_A_cI_qI  Q_B_cI_qI
Q_Int_test_cI_qI if Dataset >=19 & Dataset<=24
sum q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI q_A  q_B  q_Int Q_Int_cN_qN Q_A_cN_qN  Q_B_cN_qN  Q_Int_test_cN_qN
Q_Int_cN_qI Q_A_cN_qI  Q_B_cN_qI  Q_Int_test_cN_qI Q_Int_cI_qN Q_A_cI_qN  Q_B_cI_qN  Q_Int_test_cI_qN Q_Int_cI_qI Q_A_cI_qI  Q_B_cI_qI
Q_Int_test_cI_qI if Dataset >=25 & Dataset<=30
sum q_cons Q_cons_cN_qN  Q_cons_cN_qI   Q_cons_cI_qN   Q_cons_cI_qI q_A  q_B  q_Int Q_Int_cN_qN Q_A_cN_qN  Q_B_cN_qN  Q_Int_test_cN_qN
Q_Int_cN_qI Q_A_cN_qI  Q_B_cN_qI  Q_Int_test_cN_qI Q_Int_cI_qN Q_A_cI_qN  Q_B_cI_qN  Q_Int_test_cI_qN Q_Int_cI_qI Q_A_cI_qI  Q_B_cI_qI
Q_Int_test_cI_qI if Dataset <19  | Dataset>30


//CALCULATING CRITERIA FOR INTERACTIONS
local model="_cN_qN   _cI_qN   _cN_qI   _cI_qI"
foreach m of local model {
        generate aic`m' = 2*dfm`m'-2*LL`m'
        generate bic`m' = -2*LL`m' + dfm`m'*ln(NoPts*4)
        local outcome="C Q"
        foreach o of local outcome {
                generate `o'_pvalueInt`m' = ttail(NoPts*4-dfm`m', abs(`o'_Int`m'/`o'_Int_SE`m'))
                generate `o'_pvalueA`m'   = ttail(NoPts*4-dfm`m', abs(`o'_A`m'  /`o'_A_SE`m'))
                generate `o'_pvalueB`m'   = ttail(NoPts*4-dfm`m', abs(`o'_B`m'  /`o'_B_SE`m'))
                egen minTxeffect = rowmin(`o'_A`m' `o'_B`m')
                egen maxTxeffect = rowmax(`o'_A`m' `o'_B`m')
                replace minTxeffect=maxTxeffect if abs(minTxeffect)>abs(maxTxeffect)
                generate `o'_IntTyp`m'    = 0  if `o'_Int`m'==0
                replace  `o'_IntTyp`m'    = -1 if (`o'_Int`m'<0 & minTxeffect>=0)                                  | (`o'_Int`m'>0 & minTxeffect<=0) //Sub-additive
                replace  `o'_IntTyp`m'    = -2 if (`o'_Int`m'<0 & minTxeffect>=0 & abs(`o'_Int`m')>abs(minTxeffect)) | (`o'_Int`m'>0 & minTxeffect<=0 &
abs(`o'_Int`m')>abs(minTxeffect)) //Qualitative
                replace  `o'_IntTyp`m'    = 1  if (`o'_Int`m'>0 & minTxeffect> 0)                                  | (`o'_Int`m'<0 & minTxeffect< 0) //Super-additive
                replace  `o'_IntTyp`m'    = 2  if (`o'_Int`m'>0 & minTxeffect> 0 & abs(`o'_Int`m')>abs(minTxeffect)) | (`o'_Int`m'<0 & minTxeffect< 0 &
abs(`o'_Int`m')>abs(minTxeffect)) //Super-additive & larger than main effect
                replace  `o'_IntTyp`m'    = -3 if `o'_A`m'/`o'_B`m'<0 & `o'_Int`m'!=0 //Mixed interaction (could at this stage be qualitative or not)
                replace  `o'_IntTyp`m'    = -4 if `o'_A`m'/`o'_B`m'<0 & `o'_Int`m'!=0 & ((`o'_Int`m'/`o'_A`m'<0 & abs(`o'_Int`m')>abs(`o'_A`m')) |
(`o'_Int`m'/`o'_B`m'<0 & abs(`o'_Int`m')>abs(`o'_B`m'))) //Mixed qualitative interaction, where minTxeffect is the one with the opposing sign
                generate `o'_IntRatio`m'  = `o'_Int`m'/minTxeffect if `o'_IntTyp`m'>-3
                replace  `o'_IntRatio`m'  = `o'_Int`m'/`o'_A`m'    if `o'_IntTyp`m'<=-3 & `o'_Int`m'/`o'_A`m'<0
                replace  `o'_IntRatio`m'  = `o'_Int`m'/`o'_B`m'    if `o'_IntTyp`m'<=-3 & `o'_Int`m'/`o'_B`m'<0
                drop minTxeffect maxTxeffect
```

```
		label define `o'_IntTyp`m' -4 "Mixed Qualitative" -3 "Mixed Quant" -2 "Qualitative" -1 "SubAdd Quant" 0 "Zero" 1 "SuperAdd" 2
"LargeSuperAdd"
		label values `o'_IntTyp`m' `o'_IntTyp`m'
		generate `o'_Bias`m' = 0.5*`o'_Int`m'
	}
	forvalues R=0(5000)40000 {
		generate NB_00`m'_`R' = (Q_cons`m'                          )*`R'-(C_cons`m')
		generate NB_A0`m'_`R' = (Q_cons`m' + Q_A`m'                 )*`R'-(C_cons`m' + C_A`m')
		generate NB_0B`m'_`R' = (Q_cons`m'          + Q_B`m'        )*`R'-(C_cons`m'          + C_B`m')
		generate NB_AB`m'_`R' = (Q_cons`m' + Q_A`m' + Q_B`m' + Q_Int`m')*`R'-(C_cons`m' + C_A`m' + C_B`m' + C_Int`m')
		generate NB_Int`m'_`R'= NB_00`m'_`R' + NB_AB`m'_`R' - NB_A0`m'_`R' - NB_0B`m'_`R'
		generate txA= NB_A0`m'_`R'-NB_00`m'_`R'
		generate txB= NB_0B`m'_`R'-NB_00`m'_`R'
		egen minTxeffect = rowmin(txA txB)
		egen maxTxeffect = rowmax(txA txB)
		replace minTxeffect=maxTxeffect if abs(minTxeffect)>abs(maxTxeffect)
		generate NB_IntTyp`m'_`R' = 0  if NB_Int`m'_`R'==0
		replace  NB_IntTyp`m'_`R' = -1 if (NB_Int`m'_`R'<0 & minTxeffect>=0)                                  | (NB_Int`m'_`R'>0 & minTxeffect<0)
		replace  NB_IntTyp`m'_`R' = -2 if (NB_Int`m'_`R'<0 & minTxeffect>=0 & abs(NB_Int`m'_`R')>abs(minTxeffect)) | (NB_Int`m'_`R'>0 &
minTxeffect<0  & abs(NB_Int`m'_`R')>abs(minTxeffect))
		replace  NB_IntTyp`m'_`R' = 1  if (NB_Int`m'_`R'>0 & minTxeffect> 0)                                  | (NB_Int`m'_`R'<0 & minTxeffect< 0)
		replace  NB_IntTyp`m'_`R' = 2  if (NB_Int`m'_`R'>0 & minTxeffect> 0 & abs(NB_Int`m'_`R')>abs(minTxeffect)) | (NB_Int`m'_`R'<0 &
minTxeffect< 0 & abs(NB_Int`m'_`R')>abs(minTxeffect))
		replace  NB_IntTyp`m'_`R' = -3 if txA/txB<0 & NB_Int`m'_`R'!=0 //Mixed interaction (could at this stage be qualitative or not)
		replace  NB_IntTyp`m'_`R' = -4 if txA/txB<0 & NB_Int`m'_`R'!=0 & ((NB_Int`m'_`R'/txA<0 & abs(NB_Int`m'_`R')>abs(txA)) |
(NB_Int`m'_`R'/txB<0 & abs(NB_Int`m'_`R')>abs(txB))) //Mixed qualitative interaction, where minTxeffect is the one with the opposing sign
		generate NB_IntRatio`m'_`R' = NB_Int`m'_`R'/minTxeffect if NB_IntTyp`m'_`R'>-3
		replace  NB_IntRatio`m'_`R' = NB_Int`m'_`R'/txA                    if NB_IntTyp`m'_`R'<=-3 & NB_Int`m'_`R'/txA<0
		replace  NB_IntRatio`m'_`R' = NB_Int`m'_`R'/txB                    if NB_IntTyp`m'_`R'<=-3 & NB_Int`m'_`R'/txB<0
		drop minTxeffect maxTxeffect txA txB
		label define NB_IntTyp`m'_`R' -4 "Mixed Qualitative" -3 "Mixed Quant" -2 "Qualitative" -1 "SubAdd Quant" 0 "Zero" 1 "SuperAdd" 2
"LargeSuperAdd"
		label values NB_IntTyp`m'_`R' NB_IntTyp`m'_`R'
		egen MaxNBCurrent`m'_`R' = rowmax(NB_00`m'_`R' NB_A0`m'_`R' NB_0B`m'_`R' NB_AB`m'_`R')
		generate BestTx`m'_`R' = 1 if NB_00`m'_`R'==MaxNBCurrent`m'_`R'
		replace  BestTx`m'_`R' = 2 if NB_A0`m'_`R'==MaxNBCurrent`m'_`R'
		replace  BestTx`m'_`R' = 3 if NB_0B`m'_`R'==MaxNBCurrent`m'_`R'
		replace  BestTx`m'_`R' = 4 if NB_AB`m'_`R'==MaxNBCurrent`m'_`R'
```

```
        }
}
generate BestAnalysis_alwaysinclude= "cI_qI"
generate BestAnalysis_alwaysexclude= "cN_qN"
generate BestAnalysis_p0_05       = "cN_qN"
replace  BestAnalysis_p0_05       = "cN_qI" if C_pvalueInt_cI_qI>=0.05 & Q_pvalueInt_cI_qI< 0.05
replace  BestAnalysis_p0_05       = "cI_qN" if C_pvalueInt_cI_qI< 0.05 & Q_pvalueInt_cI_qI>=0.05
replace  BestAnalysis_p0_05       = "cI_qI" if C_pvalueInt_cI_qI< 0.05 & Q_pvalueInt_cI_qI< 0.05
generate BestAnalysis_p0_10       = "cN_qN"
replace  BestAnalysis_p0_10       = "cN_qI" if C_pvalueInt_cI_qI>=0.10 & Q_pvalueInt_cI_qI< 0.10
replace  BestAnalysis_p0_10       = "cI_qN" if C_pvalueInt_cI_qI< 0.10 & Q_pvalueInt_cI_qI>=0.10
replace  BestAnalysis_p0_10       = "cI_qI" if C_pvalueInt_cI_qI< 0.10 & Q_pvalueInt_cI_qI< 0.10
generate BestAnalysis_p0_25       = "cN_qN"
replace  BestAnalysis_p0_25       = "cN_qI" if C_pvalueInt_cI_qI>=0.25 & Q_pvalueInt_cI_qI< 0.25
replace  BestAnalysis_p0_25       = "cI_qN" if C_pvalueInt_cI_qI< 0.25 & Q_pvalueInt_cI_qI>=0.25
replace  BestAnalysis_p0_25       = "cI_qI" if C_pvalueInt_cI_qI< 0.25 & Q_pvalueInt_cI_qI< 0.25
local ic = "aic bic"
foreach i of local ic {
        generate BestAnalysis_`i'        = "cN_qN"
        replace  BestAnalysis_`i'        = "cN_qI" if `i'_cN_qI<`i'_cN_qN & `i'_cN_qI<`i'_cI_qN & `i'_cN_qI<`i'_cI_qI
        replace  BestAnalysis_`i'        = "cI_qN" if `i'_cI_qN<`i'_cN_qN & `i'_cI_qN<`i'_cN_qI & `i'_cI_qN<`i'_cI_qI
        replace  BestAnalysis_`i'        = "cI_qI" if `i'_cI_qI<`i'_cN_qN & `i'_cI_qI<`i'_cI_qN & `i'_cI_qI<`i'_cN_qI
        generate BestAnalysis_`i'3       = "cN_qN"
        replace  BestAnalysis_`i'3       = "cN_qI" if `i'_cN_qI+3<`i'_cN_qN & `i'_cN_qI+3<`i'_cI_qN & `i'_cN_qI+3<`i'_cI_qI
        replace  BestAnalysis_`i'3       = "cI_qN" if `i'_cI_qN+3<`i'_cN_qN & `i'_cI_qN+3<`i'_cN_qI & `i'_cI_qN+3<`i'_cI_qI
        replace  BestAnalysis_`i'3       = "cI_qI" if `i'_cI_qI+3<`i'_cN_qN & `i'_cI_qI+3<`i'_cI_qN & `i'_cI_qI+3<`i'_cN_qI
}
generate BestAnalysis_R2        = "cN_qN"
replace  BestAnalysis_R2        = "cN_qI" if R2Adj_cN_qI>R2Adj_cN_qN & R2Adj_cN_qI>R2Adj_cI_qN & R2Adj_cN_qI>R2Adj_cI_qI
replace  BestAnalysis_R2        = "cI_qN" if R2Adj_cI_qN>R2Adj_cN_qN & R2Adj_cI_qN>R2Adj_cN_qI & R2Adj_cI_qN>R2Adj_cI_qI
replace  BestAnalysis_R2        = "cI_qI" if R2Adj_cI_qI>R2Adj_cN_qN & R2Adj_cI_qI>R2Adj_cI_qN & R2Adj_cI_qI>R2Adj_cN_qI

generate BestAnalysis_qualitative = "cN_qN"
replace  BestAnalysis_qualitative = "cN_qI" if  C_IntTyp_cI_qI!=-2 & C_IntTyp_cI_qI!=-4  & (Q_IntTyp_cI_qI==-2 | Q_IntTyp_cI_qI==-4)
replace  BestAnalysis_qualitative = "cI_qN" if (C_IntTyp_cI_qI==-2 | C_IntTyp_cI_qI==-4) &  Q_IntTyp_cI_qI!=-2 & Q_IntTyp_cI_qI!=-4
replace  BestAnalysis_qualitative = "cI_qI" if (C_IntTyp_cI_qI==-2 | C_IntTyp_cI_qI==-4) & (Q_IntTyp_cI_qI==-2 | Q_IntTyp_cI_qI==-4)

generate BestAnalysis_qual_large = "cN_qN"
```

```
replace  BestAnalysis_qual_large = "cN_qI" if (C_IntTyp_cI_qI!=-2 & C_IntTyp_cI_qI!=2 & C_IntTyp_cI_qI!=-4) & (Q_IntTyp_cI_qI==-2 | Q_IntTyp_cI_qI==2 |
Q_IntTyp_cI_qI==-4)
replace  BestAnalysis_qual_large = "cI_qN" if (C_IntTyp_cI_qI==-2 | C_IntTyp_cI_qI==2 | C_IntTyp_cI_qI==-4) & (Q_IntTyp_cI_qI!=-2 & Q_IntTyp_cI_qI!=2 &
Q_IntTyp_cI_qI!=-4)
replace  BestAnalysis_qual_large = "cI_qI" if (C_IntTyp_cI_qI==-2 | C_IntTyp_cI_qI==2 | C_IntTyp_cI_qI==-4) & (Q_IntTyp_cI_qI==-2 | Q_IntTyp_cI_qI==2 |
Q_IntTyp_cI_qI==-4)

generate BestAnalysis_qualitativeNB = "cN_qN"
replace  BestAnalysis_qualitativeNB = "cN_qI" if (Q_IntTyp_cI_qI==-2 | NB_IntTyp_cN_qI_5000==-2 | NB_IntTyp_cN_qI_10000==-2 |
NB_IntTyp_cN_qI_15000==-2 | NB_IntTyp_cN_qI_20000==-2 | NB_IntTyp_cN_qI_25000==-2 | NB_IntTyp_cN_qI_30000==-2 | NB_IntTyp_cN_qI_35000==-2
| NB_IntTyp_cN_qI_40000==-2 | Q_IntTyp_cI_qI==-4 | NB_IntTyp_cN_qI_5000==-4 | NB_IntTyp_cN_qI_10000==-4 | NB_IntTyp_cN_qI_15000==-4 |
NB_IntTyp_cN_qI_20000==-4 | NB_IntTyp_cN_qI_25000==-4 | NB_IntTyp_cN_qI_30000==-4 | NB_IntTyp_cN_qI_35000==-4 | NB_IntTyp_cN_qI_40000==-
4) &  C_IntTyp_cI_qI!=-2 & NB_IntTyp_cI_qN_5000!=-2 & NB_IntTyp_cI_qN_10000!=-2 & NB_IntTyp_cI_qN_15000!=-2 & NB_IntTyp_cI_qN_20000!=-2 &
NB_IntTyp_cI_qN_25000!=-2 & NB_IntTyp_cI_qN_30000!=-2 & NB_IntTyp_cI_qN_35000!=-2 & NB_IntTyp_cI_qN_40000!=-2 & C_IntTyp_cI_qI!=-4 &
NB_IntTyp_cI_qN_5000!=-4 & NB_IntTyp_cI_qN_10000!=-4 & NB_IntTyp_cI_qN_15000!=-4 & NB_IntTyp_cI_qN_20000!=-4 & NB_IntTyp_cI_qN_25000!=-4
& NB_IntTyp_cI_qN_30000!=-4 & NB_IntTyp_cI_qN_35000!=-4 & NB_IntTyp_cI_qN_40000!=-4
replace  BestAnalysis_qualitativeNB = "cI_qN" if (Q_IntTyp_cI_qI!=-2 & NB_IntTyp_cN_qI_5000!=-2 & NB_IntTyp_cN_qI_10000!=-2 &
NB_IntTyp_cN_qI_15000!=-2 & NB_IntTyp_cN_qI_20000!=-2 & NB_IntTyp_cN_qI_25000!=-2 & NB_IntTyp_cN_qI_30000!=-2 & NB_IntTyp_cN_qI_35000!=-
2 & NB_IntTyp_cN_qI_40000!=-2 & Q_IntTyp_cI_qI!=-4 & NB_IntTyp_cN_qI_5000!=-4 & NB_IntTyp_cN_qI_10000!=-4 & NB_IntTyp_cN_qI_15000!=-4 &
NB_IntTyp_cN_qI_20000!=-4 & NB_IntTyp_cN_qI_25000!=-4 & NB_IntTyp_cN_qI_30000!=-4 & NB_IntTyp_cN_qI_35000!=-4 & NB_IntTyp_cN_qI_40000!=-
4) & (C_IntTyp_cI_qI==-2 | NB_IntTyp_cI_qN_5000==-2 | NB_IntTyp_cI_qN_10000==-2 | NB_IntTyp_cI_qN_15000==-2 | NB_IntTyp_cI_qN_20000==-2 |
NB_IntTyp_cI_qN_25000==-2 | NB_IntTyp_cI_qN_30000==-2 | NB_IntTyp_cI_qN_35000==-2 | NB_IntTyp_cI_qN_40000==-2 | C_IntTyp_cI_qI==-4 |
NB_IntTyp_cI_qN_5000==-4 | NB_IntTyp_cI_qN_10000==-4 | NB_IntTyp_cI_qN_15000==-4 | NB_IntTyp_cI_qN_20000==-4 | NB_IntTyp_cI_qN_25000==-4 |
NB_IntTyp_cI_qN_30000==-4 | NB_IntTyp_cI_qN_35000==-4 | NB_IntTyp_cI_qN_40000==-4)
replace  BestAnalysis_qualitativeNB = "cI_qI" if (Q_IntTyp_cI_qI==-2 | NB_IntTyp_cN_qI_5000==-2 | NB_IntTyp_cN_qI_10000==-2 |
NB_IntTyp_cN_qI_15000==-2 | NB_IntTyp_cN_qI_20000==-2 | NB_IntTyp_cN_qI_25000==-2 | NB_IntTyp_cN_qI_30000==-2 | NB_IntTyp_cN_qI_35000==-2
| NB_IntTyp_cN_qI_40000==-2 | Q_IntTyp_cI_qI==-4 | NB_IntTyp_cN_qI_5000==-4 | NB_IntTyp_cN_qI_10000==-4 | NB_IntTyp_cN_qI_15000==-4 |
NB_IntTyp_cN_qI_20000==-4 | NB_IntTyp_cN_qI_25000==-4 | NB_IntTyp_cN_qI_30000==-4 | NB_IntTyp_cN_qI_35000==-4 | NB_IntTyp_cN_qI_40000==-
4) & (C_IntTyp_cI_qI==-2 | NB_IntTyp_cI_qN_5000==-2 | NB_IntTyp_cI_qN_10000==-2 | NB_IntTyp_cI_qN_15000==-2 | NB_IntTyp_cI_qN_20000==-2 |
NB_IntTyp_cI_qN_25000==-2 | NB_IntTyp_cI_qN_30000==-2 | NB_IntTyp_cI_qN_35000==-2 | NB_IntTyp_cI_qN_40000==-2 | C_IntTyp_cI_qI==-4 |
NB_IntTyp_cI_qN_5000==-4 | NB_IntTyp_cI_qN_10000==-4 | NB_IntTyp_cI_qN_15000==-4 | NB_IntTyp_cI_qN_20000==-4 | NB_IntTyp_cI_qN_25000==-4 |
NB_IntTyp_cI_qN_30000==-4 | NB_IntTyp_cI_qN_35000==-4 | NB_IntTyp_cI_qN_40000==-4)

generate BestAnalysis_qual_largeNB = "cN_qN"
replace  BestAnalysis_qual_largeNB = "cN_qI" if (Q_IntTyp_cI_qI==-2 | NB_IntTyp_cN_qI_5000==-2 | NB_IntTyp_cN_qI_10000==-2 |
NB_IntTyp_cN_qI_15000==-2 | NB_IntTyp_cN_qI_20000==-2 | NB_IntTyp_cN_qI_25000==-2 | NB_IntTyp_cN_qI_30000==-2 | NB_IntTyp_cN_qI_35000==-2
| NB_IntTyp_cN_qI_40000==-2 | Q_IntTyp_cI_qI==-4 | NB_IntTyp_cN_qI_5000==-4 | NB_IntTyp_cN_qI_10000==-4 | NB_IntTyp_cN_qI_15000==-4 |
NB_IntTyp_cN_qI_20000==-4 | NB_IntTyp_cN_qI_25000==-4 | NB_IntTyp_cN_qI_30000==-4 | NB_IntTyp_cN_qI_35000==-4 | NB_IntTyp_cN_qI_40000==-4
| Q_IntTyp_cI_qI==2 | NB_IntTyp_cN_qI_5000==2 | NB_IntTyp_cN_qI_10000==2 | NB_IntTyp_cN_qI_15000==2 | NB_IntTyp_cN_qI_20000==2 |
```

```
NB_IntTyp_cN_qI_25000==2 | NB_IntTyp_cN_qI_30000==2 | NB_IntTyp_cN_qI_35000==2 | NB_IntTyp_cN_qI_40000==2) &  C_IntTyp_cI_qI!=-2 &
NB_IntTyp_cI_qN_5000!=-2 & NB_IntTyp_cI_qN_10000!=-2 & NB_IntTyp_cI_qN_15000!=-2 & NB_IntTyp_cI_qN_20000!=-2 & NB_IntTyp_cI_qN_25000!=-2
& NB_IntTyp_cI_qN_30000!=-2 & NB_IntTyp_cI_qN_35000!=-2 & NB_IntTyp_cI_qN_40000!=-2 & C_IntTyp_cI_qI!=-4 & NB_IntTyp_cI_qN_5000!=-4 &
NB_IntTyp_cI_qN_10000!=-4 & NB_IntTyp_cI_qN_15000!=-4 & NB_IntTyp_cI_qN_20000!=-4 & NB_IntTyp_cI_qN_25000!=-4 & NB_IntTyp_cI_qN_30000!=-
4 & NB_IntTyp_cI_qN_35000!=-4 & NB_IntTyp_cI_qN_40000!=-4 & C_IntTyp_cI_qI!=2 & NB_IntTyp_cI_qN_5000!=2 & NB_IntTyp_cI_qN_10000!=2 &
NB_IntTyp_cI_qN_15000!=2 & NB_IntTyp_cI_qN_20000!=2 & NB_IntTyp_cI_qN_25000!=2 & NB_IntTyp_cI_qN_30000!=2 & NB_IntTyp_cI_qN_35000!=2 &
NB_IntTyp_cI_qN_40000!=2
replace  BestAnalysis_qual_largeNB = "cI_qN" if (Q_IntTyp_cI_qI!=-2 & NB_IntTyp_cN_qI_5000!=-2 & NB_IntTyp_cN_qI_10000!=-2 &
NB_IntTyp_cN_qI_15000!=-2 & NB_IntTyp_cN_qI_20000!=-2 & NB_IntTyp_cN_qI_25000!=-2 & NB_IntTyp_cN_qI_30000!=-2 & NB_IntTyp_cN_qI_35000!=-
2 & NB_IntTyp_cN_qI_40000!=-2 & Q_IntTyp_cI_qI!=-4 & NB_IntTyp_cN_qI_5000!=-4 & NB_IntTyp_cN_qI_10000!=-4 & NB_IntTyp_cN_qI_15000!=-4 &
NB_IntTyp_cN_qI_20000!=-4 & NB_IntTyp_cN_qI_25000!=-4 & NB_IntTyp_cN_qI_30000!=-4 & NB_IntTyp_cN_qI_35000!=-4 & NB_IntTyp_cN_qI_40000!=-
4 & Q_IntTyp_cI_qI!=2 & NB_IntTyp_cN_qI_5000!=2 & NB_IntTyp_cN_qI_10000!=2 & NB_IntTyp_cN_qI_15000!=2 & NB_IntTyp_cN_qI_20000!=2 &
NB_IntTyp_cN_qI_25000!=2 & NB_IntTyp_cN_qI_30000!=2 & NB_IntTyp_cN_qI_35000!=2 & NB_IntTyp_cN_qI_40000!=2) & (C_IntTyp_cI_qI==-2 |
NB_IntTyp_cI_qN_5000==-2 | NB_IntTyp_cI_qN_10000==-2 | NB_IntTyp_cI_qN_15000==-2 | NB_IntTyp_cI_qN_20000==-2 | NB_IntTyp_cI_qN_25000==-2 |
NB_IntTyp_cI_qN_30000==-2 | NB_IntTyp_cI_qN_35000==-2 | NB_IntTyp_cI_qN_40000==-2 | C_IntTyp_cI_qI==-4 | NB_IntTyp_cI_qN_5000==-4 |
NB_IntTyp_cI_qN_10000==-4 | NB_IntTyp_cI_qN_15000==-4 | NB_IntTyp_cI_qN_20000==-4 | NB_IntTyp_cI_qN_25000==-4 | NB_IntTyp_cI_qN_30000==-4
| NB_IntTyp_cI_qN_35000==-4 | NB_IntTyp_cI_qN_40000==-4 | C_IntTyp_cI_qI==2 | NB_IntTyp_cI_qN_5000==2 | NB_IntTyp_cI_qN_10000==2 |
NB_IntTyp_cI_qN_15000==2 | NB_IntTyp_cI_qN_20000==2 | NB_IntTyp_cI_qN_25000==2 | NB_IntTyp_cI_qN_30000==2 | NB_IntTyp_cI_qN_35000==2 |
NB_IntTyp_cI_qN_40000==2)
replace  BestAnalysis_qual_largeNB = "cI_qI" if (Q_IntTyp_cI_qI==-2 | NB_IntTyp_cN_qI_5000==-2 | NB_IntTyp_cN_qI_10000==-2 |
NB_IntTyp_cN_qI_15000==-2 | NB_IntTyp_cN_qI_20000==-2 | NB_IntTyp_cN_qI_25000==-2 | NB_IntTyp_cN_qI_30000==-2 | NB_IntTyp_cN_qI_35000==-2
| NB_IntTyp_cN_qI_40000==-2 | Q_IntTyp_cI_qI==-4 | NB_IntTyp_cN_qI_5000==-4 | NB_IntTyp_cN_qI_10000==-4 | NB_IntTyp_cN_qI_15000==-4 |
NB_IntTyp_cN_qI_20000==-4 | NB_IntTyp_cN_qI_25000==-4 | NB_IntTyp_cN_qI_30000==-4 | NB_IntTyp_cN_qI_35000==-4 | NB_IntTyp_cN_qI_40000==-4
| Q_IntTyp_cI_qI==2 | NB_IntTyp_cN_qI_5000==2 | NB_IntTyp_cN_qI_10000==2 | NB_IntTyp_cN_qI_15000==2 | NB_IntTyp_cN_qI_20000==2 |
NB_IntTyp_cN_qI_25000==2 | NB_IntTyp_cN_qI_30000==2 | NB_IntTyp_cN_qI_35000==2 | NB_IntTyp_cN_qI_40000==2) & (C_IntTyp_cI_qI==-2 |
NB_IntTyp_cI_qN_5000==-2 | NB_IntTyp_cI_qN_10000==-2 | NB_IntTyp_cI_qN_15000==-2 | NB_IntTyp_cI_qN_20000==-2 | NB_IntTyp_cI_qN_25000==-2 |
NB_IntTyp_cI_qN_30000==-2 | NB_IntTyp_cI_qN_35000==-2 | NB_IntTyp_cI_qN_40000==-2 | C_IntTyp_cI_qI==-4 | NB_IntTyp_cI_qN_5000==-4 |
NB_IntTyp_cI_qN_10000==-4 | NB_IntTyp_cI_qN_15000==-4 | NB_IntTyp_cI_qN_20000==-4 | NB_IntTyp_cI_qN_25000==-4 | NB_IntTyp_cI_qN_30000==-4
| NB_IntTyp_cI_qN_35000==-4 | NB_IntTyp_cI_qN_40000==-4 | C_IntTyp_cI_qI==2 | NB_IntTyp_cI_qN_5000==2 | NB_IntTyp_cI_qN_10000==2 |
NB_IntTyp_cI_qN_15000==2 | NB_IntTyp_cI_qN_20000==2 | NB_IntTyp_cI_qN_25000==2 | NB_IntTyp_cI_qN_30000==2 | NB_IntTyp_cI_qN_35000==2 |
NB_IntTyp_cI_qN_40000==2)

generate BestAnalysis_qual_largeSig = "cN_qN"
replace  BestAnalysis_qual_largeSig = "cN_qI" if (C_IntTyp_cI_qI!=-2 & C_IntTyp_cI_qI!=2 & C_IntTyp_cI_qI!=-4 & C_pvalueInt_cI_qI>=0.05) &
(Q_IntTyp_cI_qI==-2 | Q_IntTyp_cI_qI==2 | Q_IntTyp_cI_qI==-4 | Q_pvalueInt_cI_qI< 0.05)
replace  BestAnalysis_qual_largeSig = "cI_qN" if (C_IntTyp_cI_qI==-2 | C_IntTyp_cI_qI==2 | C_IntTyp_cI_qI==-4 | C_pvalueInt_cI_qI< 0.05) &
(Q_IntTyp_cI_qI!=-2 & Q_IntTyp_cI_qI!=2 & Q_IntTyp_cI_qI!=-4 & Q_pvalueInt_cI_qI>=0.05)
replace  BestAnalysis_qual_largeSig = "cI_qI" if (C_IntTyp_cI_qI==-2 | C_IntTyp_cI_qI==2 | C_IntTyp_cI_qI==-4 | C_pvalueInt_cI_qI< 0.05) &
(Q_IntTyp_cI_qI==-2 | Q_IntTyp_cI_qI==2 | Q_IntTyp_cI_qI==-4 | Q_pvalueInt_cI_qI< 0.05)
```

```
local size = "250 500 1000"
foreach s of local size {
	generate BestAnalysis_`s' = "cN_qN"
	replace  BestAnalysis_`s' = "cN_qI" if abs(Q_Int_cI_qI)>=`s'/1000 & abs(C_Int_cI_qI)< `s'
	replace  BestAnalysis_`s' = "cI_qN" if abs(Q_Int_cI_qI)< `s'/1000 & abs(C_Int_cI_qI)>=`s'
	replace  BestAnalysis_`s' = "cI_qI" if abs(Q_Int_cI_qI)>=`s'/1000 & abs(C_Int_cI_qI)>=`s'
}

forvalues R=0(5000)40000 {
	generate Correct_NB_00_`R' = (q_cons                )*`R'-(c_cons)
	generate Correct_NB_A0_`R' = (q_cons + q_A          )*`R'-(c_cons + c_A)
	generate Correct_NB_0B_`R' = (q_cons       + q_B    )*`R'-(c_cons       + c_B)
	generate Correct_NB_AB_`R' = (q_cons + q_A + q_B + q_Int)*`R'-(c_cons + c_A + c_B + c_Int)
	generate Correct_NB_Int_`R'= Correct_NB_00_`R' + Correct_NB_AB_`R' - Correct_NB_A0_`R' - Correct_NB_0B_`R'
	egen Correct_MaxNBCurrent_`R' = rowmax(Correct_NB_00_`R' Correct_NB_A0_`R' Correct_NB_0B_`R' Correct_NB_AB_`R')
	generate Correct_BestTx_`R' = 1 if Correct_NB_00_`R'==Correct_MaxNBCurrent_`R'
	replace  Correct_BestTx_`R' = 2 if Correct_NB_A0_`R'==Correct_MaxNBCurrent_`R'
	replace  Correct_BestTx_`R' = 3 if Correct_NB_0B_`R'==Correct_MaxNBCurrent_`R'
	replace  Correct_BestTx_`R' = 4 if Correct_NB_AB_`R'==Correct_MaxNBCurrent_`R'
}

local criteria = "alwaysinclude alwaysexclude p0_05 p0_10 p0_25 aic bic aic3 bic3 R2 qualitative qual_large qualitativeNB qual_largeNB qual_largeSig 250
500 1000"
forvalues R=5000(15000)20000 {
	foreach c of local criteria {
		generate BestTx_`c'_`R' = BestTx_cN_qN_`R' if BestAnalysis_`c'=="cN_qN"
		replace  BestTx_`c'_`R' = BestTx_cI_qN_`R' if BestAnalysis_`c'=="cI_qN"
		replace  BestTx_`c'_`R' = BestTx_cN_qI_`R' if BestAnalysis_`c'=="cN_qI"
		replace  BestTx_`c'_`R' = BestTx_cI_qI_`R' if BestAnalysis_`c'=="cI_qI"

		generate RightBestTx`c'_`R' = 0
		replace  RightBestTx`c'_`R' = 1 if BestTx_`c'_`R'==Correct_BestTx_`R'
		generate netloss`c'_`R' = Correct_MaxNBCurrent_`R'-Correct_NB_00_`R' if BestTx_`c'_`R'==1
		replace  netloss`c'_`R' = Correct_MaxNBCurrent_`R'-Correct_NB_A0_`R' if BestTx_`c'_`R'==2
		replace  netloss`c'_`R' = Correct_MaxNBCurrent_`R'-Correct_NB_0B_`R' if BestTx_`c'_`R'==3
		replace  netloss`c'_`R' = Correct_MaxNBCurrent_`R'-Correct_NB_AB_`R' if BestTx_`c'_`R'==4

		generate EffectA_`c'_`R' = NB_A0_cN_qN_`R'-NB_00_cN_qN_`R' if BestAnalysis_`c'=="cN_qN"
```

```
                replace  EffectA_`c'_`R' = NB_A0_cI_qN_`R'-NB_00_cI_qN_`R' if BestAnalysis_`c'=="cI_qN"
                replace  EffectA_`c'_`R' = NB_A0_cN_qI_`R'-NB_00_cN_qI_`R' if BestAnalysis_`c'=="cN_qI"
                replace  EffectA_`c'_`R' = NB_A0_cI_qI_`R'-NB_00_cI_qI_`R' if BestAnalysis_`c'=="cI_qI"

                generate EffectB_`c'_`R' = NB_0B_cN_qN_`R'-NB_00_cN_qN_`R' if BestAnalysis_`c'=="cN_qN"
                replace  EffectB_`c'_`R' = NB_0B_cI_qN_`R'-NB_00_cI_qN_`R' if BestAnalysis_`c'=="cI_qN"
                replace  EffectB_`c'_`R' = NB_0B_cN_qI_`R'-NB_00_cN_qI_`R' if BestAnalysis_`c'=="cN_qI"
                replace  EffectB_`c'_`R' = NB_0B_cI_qI_`R'-NB_00_cI_qI_`R' if BestAnalysis_`c'=="cI_qI"
        }
}

foreach c of local criteria {
        generate RightBestTx`c'_Rc      = RightBestTx`c'_20000 if (Dataset>=7 & Dataset<=18)  | Dataset>=37 //ATEAM & UKBEAM, KAT, IVAN use QALYs
        generate netloss`c'_Rc      = netloss`c'_20000          if (Dataset>=7 & Dataset<=18)  | Dataset>=37 //ATEAM & UKBEAM, KAT, IVAN use QALYs
        replace  RightBestTx`c'_Rc      = RightBestTx`c'_5000  if (Dataset>=19 & Dataset<=36) | Dataset<=6 //Boyle, AFIST, MTA, Hollis use other
outcomes
        replace  netloss`c'_Rc      = netloss`c'_5000           if (Dataset>=19 & Dataset<=36) | Dataset<=6 //Boyle, AFIST, MTA, Hollis use other
outcomes
}
generate Correct_BestTx_Rc              = Correct_BestTx_20000      if (Dataset>=7 & Dataset<=18)  | Dataset>=37 //ATEAM & UKBEAM, KAT, IVAN use
QALYs
replace Correct_BestTx_Rc               = Correct_BestTx_5000       if (Dataset>=19 & Dataset<=36) | Dataset<=6 //Boyle, AFIST, MTA, Hollis use other
outcomes
sort Dataset run

//Identifying true int type & ratio
generate N_Int= Correct_NB_Int_20000                                if (Dataset>=7 & Dataset<=18)  | Dataset>=37 //ATEAM & UKBEAM, KAT, IVAN use
QALYs
replace  N_Int= Correct_NB_Int_5000                                 if (Dataset>=19 & Dataset<=36) | Dataset<=6 //Boyle, AFIST, MTA, Hollis use other
outcomes
generate N_A = Correct_NB_A0_20000-Correct_NB_00_20000  if (Dataset>=7 & Dataset<=18)  | Dataset>=37 //ATEAM & UKBEAM, KAT, IVAN use QALYs
generate N_B = Correct_NB_0B_20000-Correct_NB_00_20000  if (Dataset>=7 & Dataset<=18)  | Dataset>=37 //ATEAM & UKBEAM, KAT, IVAN use QALYs
replace  N_A = Correct_NB_A0_5000 -Correct_NB_00_5000   if (Dataset>=19 & Dataset<=36) | Dataset<=6 //Boyle, AFIST, MTA, Hollis use other outcomes
replace  N_B = Correct_NB_0B_5000 -Correct_NB_00_5000   if (Dataset>=19 & Dataset<=36) | Dataset<=6 //Boyle, AFIST, MTA, Hollis use other outcomes
local outcome "q c N"
foreach o of local outcome {
                egen minTxeffect = rowmin(`o'_A `o'_B)
                egen maxTxeffect = rowmax(`o'_A `o'_B)
                replace minTxeffect=maxTxeffect if abs(minTxeffect)>abs(maxTxeffect)
```

```
		generate `o'_IntTyp    = 0  if `o'_Int==0
		replace  `o'_IntTyp    = -1 if (`o'_Int<0 & minTxeffect>=0)                              | (`o'_Int>0 & minTxeffect<=0) //Sub-additive
		replace  `o'_IntTyp    = -2 if (`o'_Int<0 & minTxeffect>=0 & abs(`o'_Int)>abs(minTxeffect)) | (`o'_Int>0 & minTxeffect<=0 &
abs(`o'_Int)>abs(minTxeffect)) //Qualitative
		replace  `o'_IntTyp    = 1  if (`o'_Int>0 & minTxeffect> 0)                              | (`o'_Int<0 & minTxeffect< 0) //Super-additive
		replace  `o'_IntTyp    = 2  if (`o'_Int>0 & minTxeffect> 0 & abs(`o'_Int)>abs(minTxeffect)) | (`o'_Int<0 & minTxeffect< 0 &
abs(`o'_Int)>abs(minTxeffect)) //Super-additive & larger than main effect
		replace  `o'_IntTyp    = -3 if `o'_A/`o'_B<0 & `o'_Int!=0 //Mixed interaction (could at this stage be qualitative or not)
		replace  `o'_IntTyp    = -4 if `o'_A/`o'_B<0 & `o'_Int!=0 & ((`o'_Int/`o'_A<0 & abs(`o'_Int)>abs(`o'_A)) | (`o'_Int/`o'_B<0 &
abs(`o'_Int)>abs(`o'_B))) //Mixed qualitative interaction, where minTxeffect is the one with the opposing sign
		generate `o'_IntRatio  = `o'_Int/minTxeffect if `o'_IntTyp>-3
		replace  `o'_IntRatio  = `o'_Int/`o'_A    if `o'_IntTyp<=-3 & `o'_Int/`o'_A<0
		replace  `o'_IntRatio  = `o'_Int/`o'_B    if `o'_IntTyp<=-3 & `o'_Int/`o'_B<0
		drop minTxeffect maxTxeffect
		label define `o'_IntTyp -4 "Mixed Qualitative" -3 "Mixed Quant" -2 "Qualitative" -1 "SubAdd Quant" 0 "Zero" 1 "SuperAdd" 2 "LargeSuperAdd"
		label values `o'_IntTyp `o'_IntTyp
}

//Coverage and power
local model="_cN_qN   _cI_qN   _cN_qI   _cI_qI"
foreach m of local model {
	local treatment "A B"
	foreach t of local treatment {
		generate coverage_C`m'_`t' = 0
		replace  coverage_C`m'_`t' = 1 if c_`t' < C_`t'`m' + 1.959963985*C_`t'_SE`m' & c_`t' > C_`t'`m' - 1.959963985*C_`t'_SE`m' //Coverage is 1 for
this run if true value is within the 95% CI: i.e. above the lower limit and below the upper limit
		generate coverage_Q`m'_`t' = 0
		replace  coverage_Q`m'_`t' = 1 if q_`t' < Q_`t'`m' + 1.959963985*Q_`t'_SE`m' & q_`t' > Q_`t'`m' - 1.959963985*Q_`t'_SE`m'
	}
}

local criteria = "alwaysinclude alwaysexclude p0_05 p0_10 p0_25 aic bic aic3 bic3 R2 qualitative qual_large qualitativeNB qual_largeNB qual_largeSig 250
500 1000"
foreach c of local criteria {
	local treatment "A B"
	foreach t of local treatment {
		local outcome "C Q"
		foreach o of local outcome {
			generate coverage_`o'_`t'_`c' = coverage_`o'_cN_qN_`t' if BestAnalysis_`c'=="cN_qN"
```

```
                        replace  coverage_`o'_`t'_`c' = coverage_`o'_cN_qI_`t' if BestAnalysis_`c'=="cN_qI"
                        replace  coverage_`o'_`t'_`c' = coverage_`o'_cI_qN_`t' if BestAnalysis_`c'=="cI_qN"
                        replace  coverage_`o'_`t'_`c' = coverage_`o'_cI_qI_`t' if BestAnalysis_`c'=="cI_qI"

                        generate power_`o'_`t'_`c'    = 0
                        replace  power_`o'_`t'_`c'    = 1 if `o'_pvalue`t'_cN_qN<0.05 & BestAnalysis_`c'=="cN_qN" //Power is 1 for this run if p<0.05 for main
effect of A
                        replace  power_`o'_`t'_`c'    = 1 if `o'_pvalue`t'_cN_qI<0.05 & BestAnalysis_`c'=="cN_qI"
                        replace  power_`o'_`t'_`c'    = 1 if `o'_pvalue`t'_cI_qN<0.05 & BestAnalysis_`c'=="cI_qN"
                        replace  power_`o'_`t'_`c'    = 1 if `o'_pvalue`t'_cI_qI<0.05 & BestAnalysis_`c'=="cI_qI"
                }
        }
}
```

Additional results from Chapter 5

Table H3. Variation in the best performing criterion between trials and scenarios

Study Name	Interaction type for NMB*	Strategies with lowest opportunity cost†	Opportunity cost			
			Min	Always include	Never include	Include if >simple effect
AFIST	MQual	Always,	£857	£857	£2,936	£919
AFIST_DoubleInt	MQual	Always, p<0.05, p<0.10, p<0.25, AIC, Qual, QualBig, QualNMB, QualBigNMB, QualBigSig, ≥250, ≥500, ≥1000,	£0	£0	£2,106	£0
AFIST_HalfInt	Mixed	Never,	£593	£1,453	£593	£1,457
AFIST_DoubleSize	MQual	Always,	£140	£140	£2,900	£171
AFIST_NoInt	Zero	Never,	£1,777	£1,946	£1,777	£1,946
AFIST_NoInt_DoubleSize	Zero	AIC,	£510	£766	£514	£750
ATEAM	Mixed	Never, BIC, ≥500, ≥1000,	£9	£26	£9	£26
ATEAM_DoubleInt	Mixed	Qual, QualBig, QualNMB, QualBigNMB, QualBigSig,	£2	£5	£19	£2
ATEAM_HalfInt	Mixed	Never, ≥500, ≥1000,	£6	£25	£6	£21
ATEAM_DoubleSize	Mixed	ALL	£0	£0	£0	£0
ATEAM_NoInt	Zero	Never, ≥250, ≥500, ≥1000,	£17	£44	£17	£44
ATEAM_NoInt_DoubleSize	Zero	Never, ≥250, ≥500, ≥1000,	£0	£8	£0	£11
Boyle	Qual	QualNMB,	£2,403	£2,466	£3,427	£2,466
Boyle_DoubleInt	Qual	Always, QualBig, QualBigNMB, QualBigSig, ≥250,	£1,796	£1,796	£5,802	£1,796
Boyle_HalfInt	Qual	Never,	£1,082	£1,455	£1,082	£1,455

Study Name	Interaction type for NMB*	Strategies with lowest opportunity cost†	Opportunity cost			
			Min	Always include	Never include	Include if >simple effect
Boyle_DoubleSize	Qual	QualNMB,	£1,319	£1,501	£3,007	£1,523
Boyle_NoInt	Zero	Never,	£1,397	£2,031	£1,397	£2,044
Boyle_NoInt_DoubleSize	Zero	Never,	£840	£1,388	£840	£1,362
Hollis	Super	QualBigSig,	£13	£20	£14	£22
Hollis_DoubleInt	LargeSuper	Qual, QualBig,	£3	£8	£4	£15
Hollis_HalfInt	Super	Never, ≥250, ≥500, ≥1000,	£9	£19	£9	£20
Hollis_DoubleSize	Super	QualBig,	£3	£7	£4	£8
Hollis_NoInt	Zero	QualBig,	£13	£21	£14	£20
Hollis_NoInt_DoubleSize	Zero	Never, ≥250, ≥500, ≥1000,	£4	£13	£4	£11
MTA	MQual	≥250,	£138	£144	£372	£149
MTA_DoubleInt	MQual	Always, QualBig, QualBigNMB, QualBigSig,	£138	£138	£942	£138
MTA_HalfInt	Mixed	p<0.10, AIC,	£101	£117	£144	£128
MTA_DoubleSize	MQual	Always, QualBigNMB, QualBigSig,	£43	£43	£170	£43
MTA_NoInt	Zero	Never, ≥1000,	£53	£133	£53	£117
MTA_NoInt_DoubleSize	Zero	Never, ≥1000,	£0	£48	£0	£48
UKBEAM	Qual	QualNMB, QualBigNMB,	£51	£52	£56	£51
UKBEAM_DoubleInt	Qual	QualNMB, QualBigNMB,	£102	£105	£170	£102
UKBEAM_HalfInt	Sub	Never, ≥250, ≥500, ≥1000,	£22	£50	£22	£50
UKBEAM_DoubleSize	Qual	QualNMB, QualBigNMB,	£43	£44	£63	£43
UKBEAM_NoInt	Zero	Qual,	£51	£73	£52	£73
UKBEAM_NoInt_DoubleSize	Zero	Never, ≥250, ≥500, ≥1000,	£35	£52	£35	£52
IVAN	SubAdd	Never, Qual, QualBig,	£0	£28	£0	£20
IVAN_DoubleInt	Qual	Always, p<0.05, p<0.10, p<0.25, AIC, BIC, Qual, QualBig, QualNMB, QualBigNMB, QualBigSig, ≥250, ≥500, ≥1000,	£0	£0	£2,073	£0
IVAN_HalfInt	SubAdd	ALL	£0	£0	£0	£0
IVAN_DoubleSize	SubAdd	Never, Qual, QualBig,	£0	£32	£0	£18
IVAN_NoInt	Zero	ALL	£0	£0	£0	£0
IVAN_NoInt_DoubleSize	Zero	ALL	£0	£0	£0	£0
KATmobile	Qual	Always,	£3,922	£3,922	£7,255	£4,050
KATmobile_DoubleInt	Qual	Always, p<0.25, Qual, QualBig, QualNMB, QualBigNMB, QualBigSig, ≥250, ≥500,	£416	£416	£4,628	£416
KATmobile_HalfInt	Qual	Always,	£1,273	£1,273	£2,072	£1,432
KATmobile_DoubleSize	Qual	Always,	£2,401	£2,401	£5,928	£2,586

Study Name	Interaction type for NMB*	Strategies with lowest opportunity cost†	Opportunity cost			
			Min	Always include	Never include	Include if >simple effect
KATmobile_NoInt	Zero	≥1000,	£2,033	£2,411	£2,168	£2,475
KATmobile_NoInt_DoubleSize	Zero	Never,	£1,397	£1,672	£1,397	£1,664
KATmetal	Qual	QualNMB,	£2,423	£2,449	£5,969	£2,457
KATmetal_DoubleInt	Qual	Always, Qual, QualBig, QualNMB, QualBigNMB, QualBigSig, ≥250,	£1,763	£1,763	£8,805	£1,763
KATmetal_HalfInt	Qual	Always,	£4,020	£4,020	£4,438	£4,054
KATmetal_DoubleSize	Qual	Always, Qual, QualBig, QualBigSig,	£1,310	£1,310	£5,423	£1,311
KATmetal_NoInt	Zero	Never,	£2,467	£3,521	£2,467	£3,525
KATmetal_NoInt_DoubleSize	Zero	Never,	£558	£1,837	£558	£1,735

* Abbreviations for interaction types: LargeSuper, super-additive interaction larger than the smaller of the two simple effects; Mixed, mixed non-qualitative interaction; MQual, mixed qualitative interaction; Qual, non-mixed qualitative interaction; SubAdd, sub-additive interaction; Super, super-additive interaction smaller than either the simple effect; Zero, interaction equals zero.

† Abbreviations for analyses: Always, always include all interactions [criterion 1]; Never, never include any interactions [criterion 2]; $p<0.05$ [or 0.10, 0.25], including interactions where $p<0.05$ [criterion 3], $p<0.10$ [criterion 4]; or $p<0.25$ [criterion 5]; AIC, include interactions decreasing AIC [criterion 6]; BIC, include interactions decreasing BIC [criterion 7]; Qual, include qualitative interactions in costs or benefits [criterion 8]; QualBig, include interactions for costs or benefits if >simple effect [criterion 9]; QualBigSig, include interactions for costs or benefits if $p<0.05$ or >simple effect [criterion 10]; QualNMB, include qualitative interactions for costs, benefits or NMB [criterion 11]; QualBigNMB, include interactions for costs, benefits or NMB if >simple effect [criterion 12]; ≥250, include |interactions| ≥0.25 or ≥£250 [criterion 13]; ≥500, include |interactions| ≥0.5 or ≥£500 [criterion 14]; ≥1000, include |interactions| ≥1 or ≥£250 [criterion 15].

Appendix I: Additional results and methods for the IVAN economic evaluation

Additional results from Chapter 6

Table I1. Goodness of fit for models with and without interactions

	Mean total AIC (% of bootstrap replicates in which this analysis has lower AIC than base case)			
	1: Base case	**2: No interactions (drug cost analysed on natural scale)**	**3: No interactions (drug cost analysed on logarithmic scale)**	**4: Interactions in all quarters**
Drug cost	66,688	67,641 (0%)	66,675 (100%)	66,690 (35%)
Administration & monitoring costs	52,519	52,519 (100%)	52,519 (100%)	52,529 (32%)
Medications & medical service cost	66,540	66,532 (79%)	66,532 (79%)	66,539 (76%)
QALYs	-17,422	-17,432 (72%)	-17,432 (72%)	-17,419 (26%)

Table I2. Results of Scenario 2: At-the-margins ignoring all interactions

	Total QALYs (SE)	**Mean (95% CI) drug cost†**	**Mean (95% CI) administration & monitoring cost**	**Mean (95% CI) medications/ medical service cost†**	**Total cost (SE)**	**Total NMB (SE)†**
Discontinuous Avastin	1.584 (1.543, 1.624)	-£1,031 (-£1,212, -£850)	£1,824 (£1,708, £1,940)	£566 (£262, £871)	£1,360 (£986, £1,734)	£30,312 (£29,432, £31,192)
Continuous Avastin	1.604 (1.566, 1.643)	£2,770 (£2,584, £2,956)	£1,951 (£1,859, £2,043)	£544 (£267, £821)	£5,265 (£4,915, £5,615)	£26,817 (£25,960, £27,674)
Discontinuous Lucentis	1.584 (1.539, 1.629)	£10,812 (£10,313, £11,311)	£1,840 (£1,727, £1,953)	£394 (£210, £578)	£13,046 (£12,481, £13,610)	£18,630 (£17,634, £19,626)
Continuous Lucentis	1.607 (1.567, 1.646)	£14,649 (£14,372, £14,925)	£1,969 (£1,883, £2,055)	£372 (£214, £530)	£16,989 (£16,642, £17,337)	£15,144 (£14,381, £15,907)
Difference: Lucentis vs. Avastin	0.001 (-0.038, 0.041)	£11,861 (£11,499, £12,222)*	£17 (-£108, £141)	-£172 (-£452, £108)	£11,705 (£11,218, £12,192)*	-£11,677 (-£12,574, -£10,781)*
Difference: continuous vs. discontinuous	0.022 (-0.016, 0.060)	£3,819 (£3,464, £4,173)*	£128 (£19, £237)*	-£22 (-£287, £243)	£3,924 (£3,458, £4,391)*	-£3,491 (-£4,364, -£2,618)*

* Significantly different from zero ($p<0.05$)

† Analysis allowed for interactions

‡ NMB equals QALYs multiplied by ceiling ratio minus costs; the table shows NMB at a £20,000/QALY ceiling ratio.

Manuscript submitted to BMJ Open describing the IVAN economic evaluation

Cost-effectiveness of ranibizumab and bevacizumab for age-related macular degeneration: 2-year findings from the IVAN randomised trial

Authors: Helen A Dakin *Senior Researcher*,[1] Sarah Wordsworth *University Research Lecturer*,[1] Chris A Rogers *Reader in Medical Statistics*,[2] Giselle Abangma *Health Research Analyst,* [3] James Raftery *Professor of Health Technology Assessment*,[4] Simon P Harding *Professor of Clinical Ophthalmology*,[5] Andrew J Lotery *Professor of Ophthalmology,*[6] Susan M Downes *Consultant Ophthalmologist*,[7] Usha Chakravarthy *Clinical Professor*,[8†] Barnaby C Reeves *Professor of Health Services Research*,[2†] on behalf of the IVAN Study Investigators*

* The IVAN study investigators are listed online (available at http://aaojournal.org) in Appendix 1 of reference [1].

† These authors contributed equally to the manuscript and the trial.

[1] Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford, OX3 7LF, UK

[2] Clinical Trials and Evaluation Unit, School of Clinical Sciences, University of Bristol, Bristol, BS2 8HW, UK.

[3] Swiss Re Services Ltd, Research and Development. London EC3A 8EP, UK

[4] Wessex Institute for Health Research and Development, University of Southampton, SO16 7NS, UK

[5] Department of Eye and Vision Science, Institute of Ageing and Chronic Disease, University of Liverpool, Daulby Street, Liverpool, L69 3GAL, UK.

[6] Academic Unit of Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, SO16 6YD, UK

[7] Oxford Eye Hospital, Oxford University Hospitals NHS Trust, Oxford, OX3 9DU, UK.

[8] Centre for Experimental Medicine, The Queen's University of Belfast, Belfast, BT12 6BA, UK.

Funding and role of study sponsors

The IVAN trial is funded by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme (project number 07/36/01) and a full report will be published in *Health Technology Assessment*. Visit the HTA programme website (www.hta.ac.uk) for further project information. The views and opinions expressed are those of the authors and do not necessarily reflect those of the HTA programme, NIHR, the UK National Health Service or the Department of Health. The funder had no involvement in: the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. The authors had full access to the data (including statistical reports and tables), can take responsibility for the integrity of the data and the accuracy of the data analysis and are responsible for submitting this paper.

Competing interest statement

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author). All authors had financial support from the NIHR for the submitted work. HD, SW, GA and JR have no financial or non-financial interests relevant to the submitted work. UC, SH, SD and AL are principal investigators of trials sponsored by Novartis, the manufacturers of ranibizumab. UC has attended and been remunerated for attendance at advisory boards for Novartis, Bayer, Neovista, Oraya, Allergan, and Bausch and Lomb; and her employing institution has received payments from Novartis, Bayer, Neovista, Oraya, Alcon, and Pfizer. SH has attended and been remunerated for attendance at advisory boards for and received travel support from Novartis and Allergan. CR has received an honorarium from Novartis for a lecture. SD's and AL's employing institutions have received payments from Novartis. SD and AL have received honoraria from Novartis for lectures. AL has attended and been remunerated for attendance at advisory boards for Novartis and Bayer; likewise SD for Bayer and Ely Lilly. BR has received a fee for teaching from Janssen-Cilag.

Copyright/licence

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

Authorship statement

SW, CR, JR, SH, AL, SD, BR and UC conceived, designed and conducted the IVAN trial. HD, SW, GA and JR conceived and designed the economic evaluation, with extensive input from CR, SH and BR. CR supervised collation/cleaning of trial data, while HD, SW and GA collected additional resource use data from centres. HD conducted the economic analysis with help from GA under the supervision of SW. HD drafted the manuscript. All authors edited the manuscript for important intellectual content and approved the final version.

Acknowledgements

The IVAN Study Investigators are listed in Appendix 1 of reference.[1]

We would also like to thank the IVAN participants and the clinicians, nurses and clinic managers who completed resource use questionnaires.

Transparency declaration

The manuscript's guarantor (HD) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethics statement

Participants provided written informed consent. A UK National Health Service Research Ethics Committee approved the trial (07/NIR03/37).

Key words

Neovascular age-related macular degeneration (nAMD); vascular endothelial growth factor (VEGF) inhibitors; trial-based economic evaluation; cost-utility analysis; cost-minimisation analysis; cost-effectiveness.

Abstract

Objective: To assess the incremental cost and cost-effectiveness of continuous and discontinuous regimens of bevacizumab (Avastin®) and ranibizumab (Lucentis®) for neovascular age-related macular degeneration (nAMD) from a UK National Health Service perspective.

Design: A within-trial cost-utility analysis with two-year time horizon, based on a multi-centre factorial, non-inferiority randomised controlled trial.

Setting: 23 hospital ophthalmology clinics.

Participants: 610 patients aged ≥50 years with untreated nAMD in the study eye.

Interventions: 0.5 mg ranibizumab or 1.25 mg bevacizumab given continuous (monthly) or discontinuous (as-needed) for two years.

Main Outcome Measures: Quality-adjusted life-years (QALYs).

Results: Total two-year costs ranged from £3,002/patient ($4,700; 95% CI: £2,601 to £3,403) for discontinuous bevacizumab to £18,590/patient ($29,106; 95% CI: £18,258 to £18,922) for continuous ranibizumab. Ranibizumab was significantly more costly than bevacizumab for both continuous treatment (+£14,989/patient [$23,468; 95% CI: £14,522 to £15,456; p<0.001) and discontinuous treatment (+£8,498 [$13,305]; 95% CI: £7,700 to £9,295; p<0.001), with negligible difference in QALYs. Continuous ranibizumab would only be cost-effective compared with continuous bevacizumab if the NHS were willing to pay £3.5 million ($5.5 million) per additional QALY gained. Patients receiving continuous bevacizumab accrued higher total costs (+£599 [$938] 95% CI: £91 to £1,107; p=0.021) than those receiving discontinuous bevacizumab, but also accrued non-significantly more QALYs (+0.020; 95% CI: -0.032 to 0.071; p=0.452). Continuous bevacizumab therefore cost £30,220 ($47,316) per QALY gained compared with discontinuous bevacizumab. However, bootstrapping demonstrated that if the NHS is willing to pay £20,000/QALY gained, there is a 37% chance that continuous bevacizumab is cost-effective compared with discontinuous bevacizumab.

Conclusions: Ranibizumab is not cost-effective compared with bevacizumab, being substantially more costly and producing little/no QALY gain. Discontinuous bevacizumab is likely to be the most cost-effective of the four treatment strategies evaluated in IVAN, although there is a 37% chance that continuous bevacizumab is cost-effective.

Trial registration: ISRCTN92166560

Article summary: Strengths and limitations of this study

- We conducted a trial-based economic evaluation based on high-quality data on costs and quality of life prospectively collected within a randomised trial.
- This demonstrated that bevacizumab would achieve substantial cost-savings over ranibizumab with negligible differences in quality of life. In England, switching patients to bevacizumab could save at least £102 ($160) million per year. However, bevacizumab is not currently licensed for neovascular age-related macular degeneration (nAMD).
- Our study is the first to evaluate the cost-effectiveness of alternative treatment regimens for nAMD.
- Of the strategies for the treatment of nAMD evaluated in this trial, we found discontinuous (as needed) bevacizumab to be the least costly and most cost-effective. However, there was substantial uncertainty around this finding and sensitivity analyses suggested that the cost-effectiveness of using continuous (monthly) treatment rather than discontinuous treatment may vary between centres.

Introduction

Neovascular age-related macular degeneration (nAMD) is a common disorder of the ageing eye and if left untreated leads to severe central visual impairment. The current standard of care is treatment with biologicals that bind to or inhibit vascular endothelial growth factor (VEGF). Biologicals need to be injected into the vitreous cavity of the eye at 4-8 week intervals. However, the first treatment convincingly shown to be effective in preventing vision loss (ranibizumab, Lucentis®[2 3]) is expensive (£742/dose[4]). Another anti-VEGF biological, bevacizumab (Avastin®), is licensed to treat bowel cancer and has been used to treat nAMD, using smaller doses that cost much less than ranibizumab. Small non-randomised studies on bevacizumab have reported outcomes that were as good as those achievable with ranibizumab.[5] Comparative effectiveness randomised controlled trials (RCTs) of ranibizumab versus bevacizumab were therefore needed to provide unbiased estimates of relative efficacy and safety. The UK Inhibition of VEGF in Age-related choroidal Neovascularisation (IVAN) trial[1 6] and the US Comparison of Age-related macular degeneration Treatments Trials (CATT)[7 8] were amongst the first such trials to report findings.

Two-year IVAN results demonstrated that ranibizumab and bevacizumab produced similar improvements in visual function, with no significant difference in arteriothrombotic events or hospital admissions for heart failure, which have previously been linked with anti-VEGF therapy.[6] IVAN also compared discontinuous (as-needed) treatment against continuous monthly injections. Continuous and discontinuous treatment produced similar improvements in visual function, although mortality was significantly lower with continuous treatment (p=0.05).

Given the rising demands for healthcare and limited budgets, it is important to assess cost-effectiveness as well as the clinical effectiveness and safety of medical interventions. Evidence on incremental cost and cost-effectiveness is of particular importance in nAMD, owing to the potential savings and health implications of either reducing treatment frequency or substituting a much cheaper alternative (bevacizumab) for a more expensive drug (ranibizumab). Although ranibizumab costs many times more than bevacizumab, it is important to consider all relevant costs and assess cost-effectiveness to determine whether the more expensive therapy has added

health benefits that justify the additional costs or lead to savings that offset the price difference.

A recent systematic review[9] identified nine economic evaluations of ranibizumab and three of bevacizumab. Six further studies evaluating ranibizumab[10-15] and one evaluating bevacizumab[11] have since been published. Most studies found ranibizumab to be cost-effective versus other treatments, such as pegaptanib. Four studies concluded that bevacizumab was likely to be cost-effective compared with ranibizumab, but relied on observational data[11 15] or assumptions about relative efficacy.[16 17] We are unaware of any other RCT that has estimated the cost-effectiveness of bevacizumab or any study that evaluated the cost-effectiveness of continuous versus discontinuous treatment.

A key objective of the IVAN trial was to assess the incremental cost and incremental cost-effectiveness of continuous and discontinuous regimens of bevacizumab and ranibizumab in nAMD from the perspective of the UK National Health Service (NHS). The results of these analyses are reported here.

Methods

The study was based on the two-year results from the IVAN trial (ISRCTN92166560), which provided high-quality data on resource use and outcomes and comprises the only UK trial directly comparing these interventions. Trial design and methods have been described previously[1 6]; in brief this was a factorial, multi-centre non-inferiority trial in which 610 patients not previously treated for nAMD in their study eyes were randomised to either bevacizumab (0.5 mg/dose) or ranibizumab (1.25 mg/dose) and to either discontinuous treatment or continuous monthly injections for two years. Discontinuous treatment comprised an initial course of three monthly injections, followed by further courses of three injections given monthly if pre-specified clinical and optical coherence tomography (OCT) re-treatment criteria were met. The economic evaluation took a two-year time horizon to estimate within-trial cost-effectiveness. Following UK guidelines,[18] we took the perspective of the UK NHS, which excludes costs incurred by patients and their families or employers.

Detailed methods and additional results will be published as a monograph in *Health Technology Assessment*.

Since IVAN was factorial, it was important to consider the likelihood of interactions: that is, whether the differences in costs and/or quality of life between bevacizumab and ranibizumab differ between treatment regimens. Although no interactions were anticipated for visual acuity,[1] large interactions between drug and treatment regimen were expected for costs and cost-effectiveness, since reducing the number of injections would have a proportionately greater effect on drug costs for ranibizumab than for less expensive bevacizumab. Interactions for quality of life or costs are also possible if the number of injections required for discontinuous treatment differed between drugs. We therefore estimate the mean costs and mean quality-adjusted life-years (QALYs) for each of the four treatment combinations and interpret the results based on four pair-wise comparisons:

- Continuous ranibizumab versus discontinuous ranibizumab;
- Continuous bevacizumab versus discontinuous bevacizumab;
- Continuous ranibizumab versus continuous bevacizumab;
- Discontinuous ranibizumab versus discontinuous bevacizumab.

We report two forms of economic evaluation. Comparisons between drugs were based on cost-minimisation analysis, which compares costs between treatments that are assumed to have identical health effects.[19] Cost-minimisation analysis is appropriate only if the difference in cost is so large that no plausible difference in efficacy could cause the more costly treatment to be cost-effective.[20 21] This approach is justified for the comparisons between drugs because the large difference in drug costs was inevitably going to be the main influence on the incremental cost-effectiveness of ranibizumab versus bevacizumab. We therefore pre-specified that ranibizumab and bevacizumab would be compared using cost-minimisation analysis unless ranibizumab-treated patients accrued ≥0.05 more QALYs than those receiving bevacizumab. By contrast, we used cost-utility analysis, in which health outcomes are measured in QALYs, to compare continuous and discontinuous treatment, where incremental costs are smaller.

Measurement and valuation of resource use

Our base case analysis also focused on resource use associated with the study eye or associated with adverse events (AEs) or serious adverse events (SAEs) that were "*expected*": i.e. previously linked to anti-VEGF treatment (Appendix). Concomitant medications, hospitalisations and ambulatory consultations that were neither associated with the study eye nor attributable to expected AEs or expected SAEs were excluded to avoid including episodes of high healthcare resource use unrelated to treatment (e.g. renal failure or cancer), which might otherwise have swamped the main effect of treatment on costs.[22]

After enrolment, participants were monitored for disease activity on a monthly basis with visual acuity assessments, colour fundus imaging and OCT. Fundus fluorescein angiography (FFA) was undertaken at specified visits and when OCT was insufficient to reach a decision on disease activity. A pre-specified algorithm was used to determine the need for re-treatment. Patients allocated to discontinuous treatment began a new course of three monthly injections whenever they met re-treatment criteria. However, costing analyses excluded protocol-driven resource use; in particular, we assumed that patients would not require colour fundus photography, OCT or FFA unless this would affect treatment decisions. As such, patients on discontinuous treatment were assumed not to require these investigations at the second or third visit in a course of three injections, when treatment was mandated (Figure 1). Similarly, patients on continuous treatment were assumed to require monitoring consultations only once every three months, on the grounds that ophthalmologists would want information about disease progression periodically, irrespective of whether treatment decisions are required.

Micro-costing was used to estimate the cost of injection and monitoring consultations since the available national average costs[23 24] are not nAMD-specific and do not differentiate between consultations for monitoring and intravitreal drug delivery. Staff at 13 of the 23 IVAN centres completed questionnaires on overheads, the cost of setting up clinic facilities and equipment and/or the staff required to run injection and monitoring clinics.

The drug acquisition cost for ranibizumab was the NHS list price (£742.17[4]) and that for bevacizumab was the price typically charged by the not-for-profit NHS provider used in the trial (£49/prefilled syringe). All concomitant medications, contacts with medical professionals and hospitalisations were recorded at each monthly clinic visit. Concomitant medications applied to the study eye or indicated for any expected SAE/AE were valued using list prices.[4] Costs of other medications, including those applied to the fellow eye, were excluded from the analysis. Unit costs for consultations with general practitioners, district or general practice nurses and hospital staff outside IVAN clinics were obtained from routine sources.[24 25] These costs were applied to ambulatory consultations that were either related to the eye or that occurred within 30 days of an expected SAE/AE. Hospital stays linked to expected SAEs were valued using the mean cost per bed-day for associated HRGs.[24]

Resource use data and unit costs were combined to estimate quarterly costs of: bevacizumab/ranibizumab; drug administration and monitoring consultations; and hospitalisations, ambulatory consultations and medication changes for expected SAEs/AEs. Value added tax (VAT) was excluded from the economic evaluation and included in budget impact estimates, following guidelines.[18] Costs are reported in 2011 pounds sterling, accompanied by equivalents in US dollars (exchange rate: $1.57/pound).

Measurement and valuation of health benefits

The three-level EQ-5D questionnaire[26] was administered at baseline, 3, 12 and 24 months, and (if the patient was willing and able to do so) at study exit, after any SAE and after a drop in visual acuity in the study eye of ≥15 letters on the Early Treatment Diabetic Retinopathy Study vision chart between two consecutive visits (referred to subsequently as a 'reduction in visual acuity'). The Health Utilities Index questionnaire version 3 (HUI3) was administered at the same timepoints and used in sensitivity analyses; EQ-5D comprised the utility primary measure following UK guidelines.[18] Patients self-completed large-print EQ-5D questionnaires, with assistance from study nurses where necessary; responses were valued using the UK time-trade-off tariff to give "utilities".[26]

Missing utility data were imputed using multiple imputation,[27] which avoids bias and enables analysis of the whole sample. Multiple imputation was conducted using the ice command[28] (version 1.9.4) in Stata Version 12 (StataCorp, College Station, TX).

QALYs for each participant were calculated as the area under the curve. We assumed that utility changed linearly between consecutive EQ-5D measurements in the absence of SAEs. Since linear changes are unlikely for patients with SAEs, we assumed that SAEs and reductions in visual acuity caused a sudden drop in utility on the day of onset, followed by a linear rise as the patient recovered; the rate of this linear rise was estimated using mixed models (Appendix).

Statistical methods

Linear regression models were used to estimate the effect of drug and treatment regimen on QALYs, drug costs, administration/monitoring costs and medication/medical service use in each three-month period or "quarter" (Appendix). Interactions between drug and treatment regimen were included in the models for quarters 2-8 if they were either statistically significant or were larger than the main effect for drug or for treatment regimen. The analysis of QALYs, drug costs and medication/medical service use in quarters 2-8 therefore took account of interactions, while drug and treatment regimen were assumed to have additive effects on administration/monitoring costs. A variant on Kaplan-Meier sample averaging[29 30] was used to account for patients withdrawing early from the trial and exclude differences in mortality unrelated to treatment; regression predictions of quarterly costs and QALYs were weighted by the proportion of patients alive at the start of each quarter. Costs accrued in Year 2 were discounted at 3.5% to allow for time preference (i.e. the tendency to prefer benefits sooner and costs later).[18] Uncertainty around quarterly costs and QALYs was quantified by estimating models separately for 130 nonparametric bootstrap draws on each of 100 datasets generated in multiple imputation to capture the uncertainty around imputed utilities. The appendix gives further details of the statistical methodology.

Presentation of results and uncertainty

Incremental cost-effectiveness ratios (ICERs) were calculated by dividing the difference in cost between two study arms by the difference in QALYs. Results were interpreted assuming that the UK NHS would be willing to pay £20,000 to gain one QALY (a £20,000/QALY "ceiling ratio").[31] We also present net benefits for each of the four treatment arms: net benefit equals total QALYs multiplied by the ceiling ratio, minus total costs. Uncertainty around ICERs is presented as cost-effectiveness acceptability curves, which plot the probability of each of the four treatment groups having the highest net benefits (i.e. being most cost-effective) at a range of ceiling ratios.

Sensitivity analyses evaluated the impact of changing the costs (e.g. halving the cost of ranibizumab), methods (e.g. taking a one-year time horizon) and assumptions (e.g. including the costs of all SAEs, not just 'expected' SAEs).

Results

QALYs and quality of life

The number of QALYs accrued over the two-year trial period did not differ significantly between bevacizumab and ranibizumab, or between continuous and discontinuous treatment ($p \geq 0.381$; Table 1). Patients randomised to continuous treatment accrued non-significantly more QALYs than those randomised to discontinuous treatment (mean difference: 0.020 [95% CI: -0.032, 0.071] for bevacizumab, p=0.452 and 0.026 [95% CI: -0.032, 0.085] for ranibizumab, p=0.381), while differences between ranibizumab and bevacizumab were negligible.

Resource use and costs

Patients receiving continuous treatment received a mean of 22 injections, while those on discontinuous treatment received 13. Consequently, drug costs differed substantially between continuous and discontinuous treatment (Table 1; $p<0.001$), as well as between ranibizumab and bevacizumab ($p<0.001$). Since reducing treatment frequency produces larger savings for ranibizumab than for bevacizumab, there were significant interactions between drug and treatment regimen for drug cost ($p<0.001$).

Administration of bevacizumab or ranibizumab cost £61 ($96; standard deviation, SD: £14) per injection, while each consultation for monitoring cost £72 ($113; SD: £41), plus £39 ($61; SD: £16) for each FFA. Administering intravitreal injections and monitoring disease progression/remission cost between £1,825 and £1,970 per patient over the two-year trial period (Table 1). Discontinuous treatment reduced the number of injections required, but increased the number of monitoring consultations needed to assess disease status against retreatment criteria, since we assumed that OCT would only be done when it would inform treatment decisions. Since continuous treatment requires, on average, nine more injections ($p<0.001$), but avoids only six monitoring visits ($p<0.001$), drug administration and monitoring costs were higher with continuous treatment than discontinuous treatment (mean difference: £130 per patient ($204); 95% CI: £20, £239; $p=0.021$), with no significant difference between bevacizumab and ranibizumab ($p=0.80$).

The cost of medication changes, hospitalisations and ambulatory consultations associated with expected SAEs and expected AEs was relatively small (mean: £469 [$735] per patient), but varied substantially between patients (95 percentile range: £0, £1,401). There was no significant difference in such costs between drugs or treatment regimens ($p\geq0.163$).

Taking account of the drug cost, drug administration/monitoring and medication/medical service use, the mean total cost per patient over the two-year trial ranged from £18,590 ($29,119) for continuous ranibizumab to £3,002 ($4,702) for discontinuous bevacizumab (Table 1). Drug cost accounted for 80-88% of the total cost for patients randomised to ranibizumab and 21-30% of the cost for patients randomised to bevacizumab. Drug administration and monitoring accounted for 54-61% of the costs accrued by patients randomised to bevacizumab and 10-15% of costs for those randomised to ranibizumab.

Base case comparison between ranibizumab and bevacizumab

Since the difference in mean QALYs between ranibizumab and bevacizumab was less than the pre-specified non-inferiority margin (0.05 QALYs), cost-minimisation analysis was used to compare the two drugs on the basis of cost alone. Overall,

continuous ranibizumab cost £14,989 more per patient ($23,476 [95% CI: £14,522, £15,456], Table 1) than continuous bevacizumab over the two-year trial period (p<0.001). Discontinuous ranibizumab cost £8,498 more per patient ($13,308 [95% CI: £7,700, £9,295], p<0.001) compared with discontinuous bevacizumab. Bootstrapping analyses estimated the probability that switching from ranibizumab to bevacizumab would save money and found that this exceeds 99.9%.

Base case comparison between continuous and discontinuous treatment

Overall, using continuous rather than discontinuous treatment increased costs by £7,090 ($11,102 [95% CI: £6,337, £7,844], p<0.001) for ranibizumab and £599 ($938 [95% CI: £91, £1,107], p=0.021) for bevacizumab.

However, patients randomised to continuous bevacizumab also accrued non-significantly more QALYs than those randomised to discontinuous bevacizumab (Table 1; p=0.452). In line with best practice,[19] we took account of the non-significant differences in QALYs and allowed for the joint distribution of costs and QALYs, since assuming no difference in health outcomes can introduce bias and give misleading conclusions.[20 21] Dividing the difference in cost by the difference in QALYs suggests that continuous bevacizumab costs £30,220 ($47,316) per additional QALY gained compared with discontinuous bevacizumab. This ICER is somewhat higher than the £20,000 ($31,000) per QALY "ceiling ratio" below which the NHS generally considers treatments to be cost-effective.[31] However, the imprecision around QALY differences means that there is substantial uncertainty around this ICER. Bootstrapping demonstrated that there is a 37% chance that continuous bevacizumab is cost-effective compared with discontinuous bevacizumab at a £20,000/QALY ceiling ratio, which increases to 50% at £30,000/QALY.

Continuous ranibizumab cost £270,217 ($423,074) per QALY gained compared with discontinuous ranibizumab. Due to the substantial savings possible by giving ranibizumab less frequently, we can be >99.99% confident that continuous ranibizumab is poor value for money compared with discontinuous ranibizumab at a £20,000/QALY ceiling ratio.

Base case four-way comparison

It is also informative to consider the four trial treatment groups as four mutually-exclusive alternative strategies for managing nAMD. Framing the decision in this way demonstrates that discontinuous bevacizumab is the most cost-effective treatment strategy evaluated in IVAN, generating higher net benefits than the other three treatment strategies (Table 1), where net benefit equals QALYs multiplied by ceiling ratio (in this case £20,000/QALY) minus costs. Continuous ranibizumab would only be cost-effective compared with continuous bevacizumab if the NHS were willing to pay £3.5 million ($5.5 million) per additional QALY gained. Discontinuous ranibizumab is not cost-effective at any ceiling ratio, as it is more costly and less effective than continuous or discontinuous bevacizumab.

However, there remains substantial uncertainty around incremental QALY gains. This is illustrated by the cost-effectiveness acceptability curves plotting the probability of each treatment being the most cost-effective of the four strategies at different ceiling ratios (Figure 2). This demonstrates that, although we can be 98% confident that discontinuous bevacizumab is less costly than continuous bevacizumab, our confidence in the conclusion that discontinuous bevacizumab has highest net benefits decreases rapidly as the value we place on the small, non-significant QALY gains increases. At a £20,000/QALY ceiling ratio, there is a 63% probability that discontinuous bevacizumab is the most cost-effective strategy considered in IVAN and a 37% probability that continuous bevacizumab is the most cost-effective. By contrast, the probability of either continuous or discontinuous ranibizumab being the most cost-effective strategy for managing nAMD is <1% unless the NHS were willing to pay more than £100,000/QALY gained.

Sensitivity analyses

Sensitivity analyses demonstrated that the conclusions are very robust to changes in the assumptions and methods used to measure costs and utilities and conduct the analysis (Figure 3). Notably, no sensitivity analysis changed the conclusion that ranibizumab is not cost-effective compared with bevacizumab. However, three sensitivity analyses changed the conclusion that continuous bevacizumab is not cost-effective compared with discontinuous bevacizumab: assuming that FFA is only conducted at baseline, not at any subsequent monitoring consultation; measuring

quality of life using HUI3 rather than EQ-5D; and using unadjusted Kaplan-Meier estimates of the probability of surviving at any point in time to account for censoring, rather than excluding differences in deaths that were unrelated to study medication (see Appendix).

Threshold analyses demonstrated that the price of ranibizumab would need to be reduced to £63.46 per dose (a 91% price reduction) in order for continuous ranibizumab to be cost-effective compared with continuous bevacizumab at a £20,000/QALY ceiling ratio.

Discussion

This study demonstrates that in a UK setting, we can be >99% confident that ranibizumab represents very poor value for money compared with bevacizumab at the £20,000 ($31,000) per QALY ceiling ratio used within NHS decision-making.[31] Continuous ranibizumab would only be cost-effective compared with continuous bevacizumab if the NHS were willing to pay >£3.5 ($5.5) million/QALY gained. Furthermore, our analysis also shows that giving discontinuous bevacizumab, rather than discontinuous ranibizumab could save the UK NHS £8,498 ($13,341) per patient treated, with little or no impact on the health gains from treatment. If the 17,295 eyes requiring anti-VEGF therapy each year in England[32] were switched from discontinuous ranibizumab to discontinuous bevacizumab, the NHS could save at least £102 ($160) million per year (including 20% VAT) based on the treatment regimens evaluated in IVAN. It remains controversial as to whether a drug (bevacizumab) that has not been approved and licensed for nAMD by regulatory agencies should be used when a licensed drug (ranibizumab) is available. In the UK, clinicians may prescribe unlicensed medications within approved research projects, when no suitable medicine is licensed, or when the licensed alternative is unavailable,[33] although prescribing on cost grounds is not mentioned. By contrast, in the US, ophthalmologists use bevacizumab freely.[34] National guidance (rather than local hospital/clinician policies) is therefore needed in the UK to direct the choice between bevacizumab and ranibizumab. CATT and IVAN provide robust data to guide regulators in this decision, demonstrating that ranibizumab and bevacizumab

have comparable effects on vision and similar safety profiles,[6 7] but that ranibizumab costs £3.5 million per QALY compared with bevacizumab.

The base case analysis found that continuous bevacizumab cost £30,220 ($47,445) per QALY gained compared with discontinuous bevacizumab, suggesting that discontinuous bevacizumab is the most cost-effective strategy evaluated in IVAN if the NHS is willing to pay up to £20,000/QALY gained. However, there remains substantial uncertainty around this conclusion and there is a 37% chance that continuous bevacizumab is cost-effective. The finding of non-significantly higher QALYs with continuous treatment contradicts our prior hypothesis that avoiding monthly injections might improve quality of life, although the observed difference could be due to chance. Nonetheless, discontinuous bevacizumab would remain the most cost-effective strategy even if there were no difference in quality of life between treatment regimens. Other considerations may affect the choice of anti-VEGF delivery model. In particular, since discontinuous treatment requires regular clinical review and access to retinal imaging, it may be more practical to provide treatment every month, with monitoring restricted to specified points in time (e.g. six or 12 months after initiation of therapy). Indeed the label for the newest anti-VEGF (aflibercept) incorporates a limited clinical monitoring regime.[35] The discontinuous treatment regimen evaluated in the IVAN trial was chosen partly to minimise the possibility of disadvantage to participants in these groups and partly to minimise the number of retreatment decisions required. Neither monthly treatment nor treatments given as blocks of three are used widely in routine practice, although following publication of IVAN,[1 6] there appears to be increased interest in the "IVAN regimen". The cost-effectiveness of monthly versus intermittent treatment will therefore vary between treatment centres depending on local costs and clinical practice.

Unlike previous studies, our analysis is based on high-quality data from an RCT, which was powered to exclude any clinically-meaningful difference in visual acuity and with prospective measurements of costs and quality of life. It therefore provides unequivocally-unbiased estimates of incremental costs and QALYs. Nevertheless, our analysis confirms the findings of previous economic evaluations, namely that ranibizumab is not cost-effective compared with bevacizumab.[11 17] We are also (to our knowledge) the first to evaluate the cost-effectiveness of alternative treatment

regimens. In addition to following best practice for trial-based economic evaluation, this study includes several novel aspects, such as measuring quality of life after SAEs, excluding chance differences in deaths unrelated to treatment and allowing for the factorial design by including only large or statistically significant interactions.

The study also estimates the cost of consultations to administer ranibizumab/bevacizumab and monitor outcomes, which could be used in other economic evaluations. Micro-costing shows the main drivers of consultation costs and highlighted substantial variation in costs between centres; this variation means that the cost-effectiveness of continuous versus discontinuous bevacizumab (but not ranibizumab versus bevacizumab) will vary between centres. It is important to note that the costs were calculated to assess incremental cost-effectiveness in IVAN and should not be used to set the prices at which hospitals are reimbursed. In particular, they are bottom-up estimates that exclude unpaid overtime and VAT and make assumptions about overheads and proportion of staff-time spent on patient contacts. In most settings it is likely that the costs to healthcare commissioners will be higher and subject to local negotiations with care providers.

The base case analysis focused on mortality attributable to study medication and the costs associated with "expected" SAEs/AEs and excluded other costs. This reduced the risk that chance differences in resource use not associated with study medication could distort our conclusions. However, it also means that the unanticipated increase in the incidence of other SAEs (e.g. gastrointestinal events) with bevacizumab[16] (which comprised the only difference in SAEs between drugs) is not taken into account in the costing analysis. However, sensitivity analyses including the cost of all SAEs/AEs gave the same conclusions.

Further research is needed to assess the extent to which the cost-effectiveness findings generalise to other countries with different relative prices and management of nAMD and SAEs/AEs. For example, the incidence of SAEs was substantially lower in IVAN than CATT,[6 7] although sensitivity analyses suggested that this did not change the conclusions. The costs of the two drugs may vary between centres within the UK as hospitals may use different bevacizumab suppliers or have different discounts on ranibizumab. Nevertheless, because we collected very detailed

information on resource use, policy makers in other countries can review these data against their own to examine their similarity and, hence, the applicability of our findings to their setting. Future work combining data from IVAN with that from other trials, such as CATT,[7] may help reduce uncertainty and evaluate the extent to which the results can be generalised. However we believe that our primary finding of ranibizumab representing very poor value for money compared with bevacizumab does apply throughout the world.

References

1. The IVAN Study Investigators, Chakravarthy U, Harding SP, et al. Ranibizumab versus Bevacizumab to Treat Neovascular Age-related Macular Degeneration: One-Year Findings from the IVAN Randomized Trial. Ophthalmology 2012;119(7):1399-411.
2. Brown DM, Michels M, Kaiser PK, et al. Ranibizumab versus verteporfin photodynamic therapy for neovascular age-related macular degeneration: Two-year results of the ANCHOR study. Ophthalmology 2009;116(1):57-65.e5.
3. Rosenfeld PJ, Brown DM, Heier JS, et al. Ranibizumab for neovascular age-related macular degeneration. The New England journal of medicine 2006;355(14):1419-31.
4. British Medical Association. British National Formulary 62 (http://bnf.org/bnf/index.htm). London: Pharmaceutical Press, September 2011.
5. Mitchell P. A systematic review of the efficacy and safety outcomes of anti-VEGF agents used for treating neovascular age-related macular degeneration: comparison of ranibizumab and bevacizumab. Current Medical Research & Opinion 2011;27(7):1465-75.
6. Chakravarthy U, Harding SP, Rogers CA, et al. Alternative treatments to inhibit VEGF in age-related choroidal neovascularisation: 2-year findings of the IVAN randomised controlled trial. Lancet 2013;382(9900):1258-67.
7. Martin DF, Maguire MG, Fine SL, et al. Ranibizumab and Bevacizumab for Treatment of Neovascular Age-related Macular Degeneration: Two-Year Results. Ophthalmology 2012;119(7):1388-98.

8. Martin DF, Maguire MG, Ying GS, et al. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. N Engl J Med 2011;364(20):1897-908.
9. Mitchell P, Annemans L, White R, et al. Cost effectiveness of treatments for wet age-related macular degeneration. PharmacoEconomics 2011;29(2):107-31.
10. Yanagi Y, Aihara Y, Fukuda T, et al. [Cost-effectiveness of ranibizumab, photodynamic therapy and pegaptanib sodium in the treatment of neovascular age-related macular degeneration in Japanese]. Nihon Ganka Gakkai Zasshi 2011;115(9):825-31.
11. Patel JJ, Mendes MA, Bounthavong M, et al. Cost-utility analysis of bevacizumab versus ranibizumab in neovascular age-related macular degeneration using a Markov model. J Eval Clin Pract 2012;18(2):247-55.
12. Gower EW, Cassard SD, Bass EB, et al. A cost-effectiveness analysis of three treatments for age-related macular degeneration. Retina 2010;30(2):212-21.
13. Hernandez-Pastor LJ, Ortega A, Garcia-Layana A, et al. Cost-effectiveness of ranibizumab compared with pegaptanib in neovascular age-related macular degeneration. Graefes Archive for Clinical & Experimental Ophthalmology 2010;248(4):467-76.
14. Neubauer AS, Holz FG, Sauer S, et al. Cost-effectiveness of ranibizumab for the treatment of neovascular age-related macular degeneration in Germany: Model analysis from the perspective of Germany's statutory health insurance system. Clinical Therapeutics 2010;32(7):1343-56.
15. Smiddy WE. Economic implications of current age-related macular degeneration treatments. Ophthalmology 2009;116(3):481-7.
16. Fletcher EC, Lade RJ, Adewoyin T, et al. Computerized model of cost-utility analysis for treatment of age-related macular degeneration. Ophthalmology 2008;115(12):2192-8.
17. Raftery J, Clegg A, Jones J, et al. Ranibizumab (Lucentis) versus bevacizumab (Avastin): modelling cost effectiveness. Br J Ophthalmol 2007;91(9):1244-6.
18. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. Secondary Guide to the methods of technology appraisal 2013 2013. http://www.nice.org.uk/media/D45/1E/GuideToMethodsTechnologyAppraisal2013.pdf.

19. Drummond MF, Sculpher MJ, Torrance GW, et al. Methods for the economic evaluation of health care programmes. 3rd ed. New York: Oxford University Press, 2005.
20. Dakin H, Wordsworth S. Cost-minimisation analysis versus cost-effectiveness analysis, revisited. Health Econ 2013;22(1):22-34.
21. Briggs AH, O'Brien BJ. The death of cost-minimization analysis? Health Econ 2001;10(2):179-84.
22. Jonsson B, Weinstein MC. Economic evaluation alongside multinational clinical trials. Study considerations for GUSTO IIb. International journal of technology assessment in health care 1997;13(1):49-58.
23. Department of Health. PbR operational guidance and tariffs: 2011-12 tariff information spreadsheet. Secondary PbR operational guidance and tariffs: 2011-12 tariff information spreadsheet 2011. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_124356.
24. Department of Health. National Schedule of Reference Costs 2010-11 for NHS Trusts and PCTs combined. Secondary National Schedule of Reference Costs 2010-11 for NHS Trusts and PCTs combined November 2011. http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_131148.xls.
25. Curtis L. Unit Costs of Health and Social Care 2011. Pages 91, 129-150, 177-213. Secondary Unit Costs of Health and Social Care 2011. Pages 91, 129-150, 177-213 2011. http://www.pssru.ac.uk/archive/pdf/uc/uc2011/uc2011.pdf.
26. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35(11):1095-108.
27. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30(4):377-99.
28. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. Stata Journal 2007;7:445-64.
29. Hallstrom AP, Sullivan SD. On estimating costs for economic evaluation in failure time studies. Med Care 1998;36(3):433-6.
30. Gray A, Clarke P, Wolstenholme J, et al. Applied Methods of Cost-Effectiveness Analysis in Health Care Oxford: Oxford University Press, 2011.

31. National Institute for Health and Clinical Excellence. Social value judgements: Principles for the development of NICE guidance. Second Edition. Secondary Social value judgements: Principles for the development of NICE guidance. Second Edition 2008. http://www.nice.org.uk/media/C18/30/SVJ2PUBLICATION2008.pdf.

32. National Institute for Health and Clinical Excellence. Ranibizumab and pegaptanib for the treatment of age-related macular degeneration TA155: Costing template and report. Secondary Ranibizumab and pegaptanib for the treatment of age-related macular degeneration TA155: Costing template and report August 2008. (http://www.nice.org.uk/nicemedia/live/12057/41735/41735.xls).

33. General Medical Council. Good practice in prescribing and managing medicines and devices. Secondary Good practice in prescribing and managing medicines and devices 2013. http://www.gmc-uk.org/static/documents/content/Prescribing_guidance.pdf.

34. Curtis LH, Hammill BG, Schulman KA, et al. Risks of mortality, myocardial infarction, bleeding, and stroke associated with therapies for age-related macular degeneration.[Erratum appears in Arch Ophthalmol. 2010 Dec;128(12):1623]. Archives of Ophthalmology 2010;128(10):1273-9.

35. Bayer Plc. Summary of product characteristics for Eylea 40mg/ml solution for injection in a vial Secondary Summary of product characteristics for Eylea 40mg/ml solution for injection in a vial 2013. http://www.medicines.org.uk/emc/medicine/27224/SPC/Eylea+40mg+ml+solution+for+injection+in+a+vial/.

Table 1: Results of the economic evaluation

	Total QALYs (95% CI)‡	Mean (95% CI) drug cost‡	Mean (95% CI) administration & monitoring cost	Mean (95% CI) medication/medical service cost‡	Total cost (95% CI)‡	Total net benefits (95% CI)†‡
Discontinuous bevacizumab	1.584 (1.538, 1.630)	£651 (£605, £698)	£1,825 (£1,708, £1,941)	£526 (£144, £908)	£3,002 (£2,601, £3,403)	£28,683 (£27,707, £29,658)
Continuous bevacizumab	1.604 (1.563, 1.645)	£1,065 (£1,048, £1,081)	£1,952 (£1,860, £2,043)	£585 (£250, £919)	£3,601 (£3,259, £3,943)	£28,480 (£27,548, £29,412)
Discontinuous ranibizumab	1.582 (1.530, 1.634)	£9,229 (£8,584, £9,875)	£1,838 (£1,724, £1,952)	£432 (£253, £611)	£11,500 (£10,798, £12,202)	£20,142 (£18,963, £21,321)
Continuous ranibizumab	1.608 (1.565, 1.651)	£16,286 (£16,011, £16,562)	£1,970 (£1,883, £2,057)	£334 (£215, £452)	£18,590 (£18,258, £18,922)	£13,576 (£12,769, £14,383)
Difference: ranibizumab vs. bevacizumab	Continuous: 0.004 (-0.046, 0.054)	Continuous: £15,222 (£14,948, £15,495)*	£16 (-£109, £141)	Continuous: -£251 (-£604, £102)	Continuous: £14,989 (£14,522, £15,456)*	Continuous: -£14,904 (-£15,995, -£13,813)*
	Discontinuous: -0.002 (-0.064, 0.060)	Discontinuous: £8,578 (£7,932, £9,225)*		Discontinuous: -£94 (-£514, £326)	Discontinuous: £8,498 (£7,700, £9,295)*	Discontinuous: -£8,541 (-£9,939, -£7,144)*
Difference: continuous vs. discontinuous	Ranibizumab: 0.026 (-0.032, 0.085)	Ranibizumab: £7,057 (£6,364, £7,750)*	£130 (£20, £239)*	Ranibizumab: -£98 (-£310, £113)	Ranibizumab: £7,090 (£6,337, £7,844)*	Ranibizumab: -£6,566 (-£7,861, -£5,271)*
	Bevacizumab: 0.020 (-0.032, 0.071)	Bevacizumab: £413 (£365, £462)*		Bevacizumab: £59 (-£438, £556)	Bevacizumab: £599 (£91, £1,107)*	Bevacizumab: -£203 (-£1,372, £967)
Interaction	0.006 (-0.071, 0.084)	£6,643 (£5,949, £7,338)*	£5 (-£31, £42)	-£157 (-£696, £381)	£6,491 (£5,604, £7,379)*	-£6,363 (-£8,088, -£4,638)*

* Significantly different from zero ($p<0.05$).
† Net benefits equal QALYs multiplied by ceiling ratio minus costs; the net benefits shown in this table were calculated at a £20,000/QALY ceiling ratio.
‡ Analysis allowed for interactions.

Figure 1: Schematic illustrating the assumptions made about the frequency of injection and monitoring consultations within the costing analysis. The consultations required by patients on discontinuous treatment will depend on when they met treatment failure criteria; Patient 2 met the re-treatment criteria at visits 0, 7 and 11.

	Visit	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Patient 1: Continuous treatment	Injection	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓
	Monitoring consult	✓			✓			X			✓		
	FFA	✓			?			X			?		
Patient 2: Discontinuous treatment	Injection	✓	✓	✓					✓	✓	✓		✓
	Monitoring consult	✓			✓	✓	X	✓	✓			✓	✓
	FFA	✓			?	?	X	?	?			?	?

✓ Relevant consultation cost was applied.
? The cost of fundus fluorescein angiography (FFA) was applied if clinically indicated: for discontinuous patients, this was applied whenever the patient had FFA in the trial; for continuous patients, the proportion of patients having FFA was based on estimated use in routine clinical practice.
X No consultation cost was applied as the participant missed the visit.

Figure 2: Cost-effectiveness acceptability curve showing the probability that each treatment is the most cost-effective strategy evaluated in IVAN at a range of ceiling ratios. For example, at a ceiling ratio of £20,000/QALY gained (shown by the vertical dashed line), there is a 63% probability that discontinuous bevacizumab is best and a 37% probability that continuous bevacizumab is best, while the probability that either ranibizumab treatment regimen is best is approximately 0% (total = 100%).

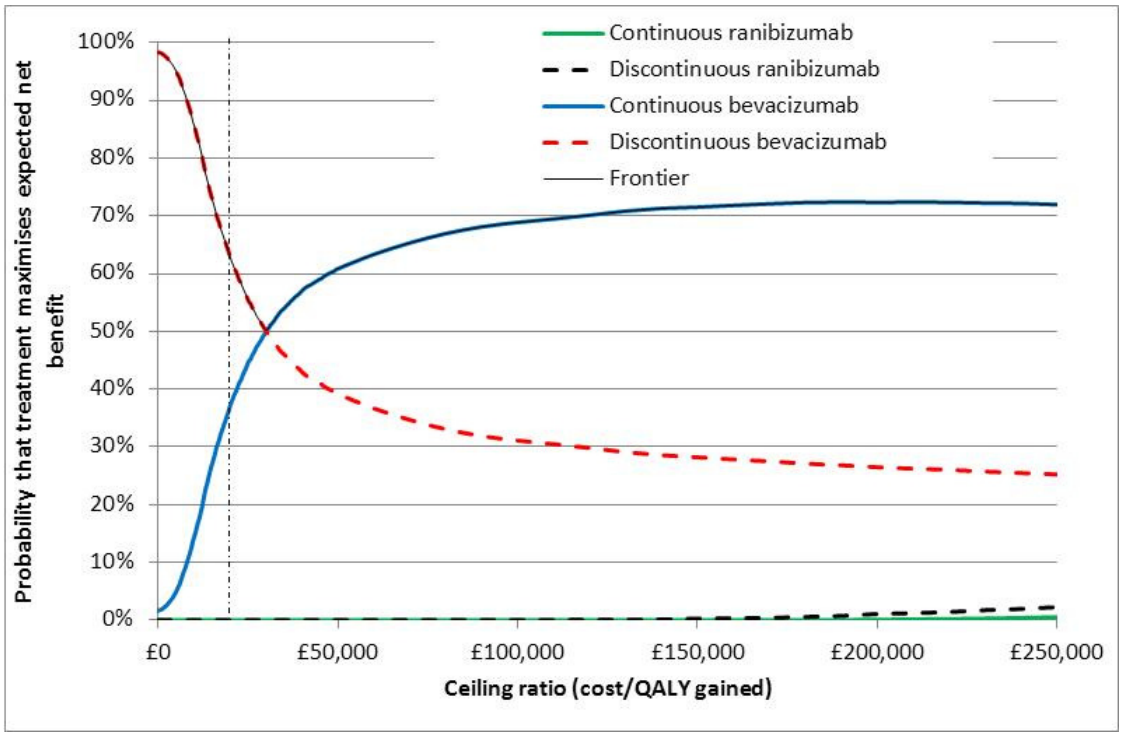
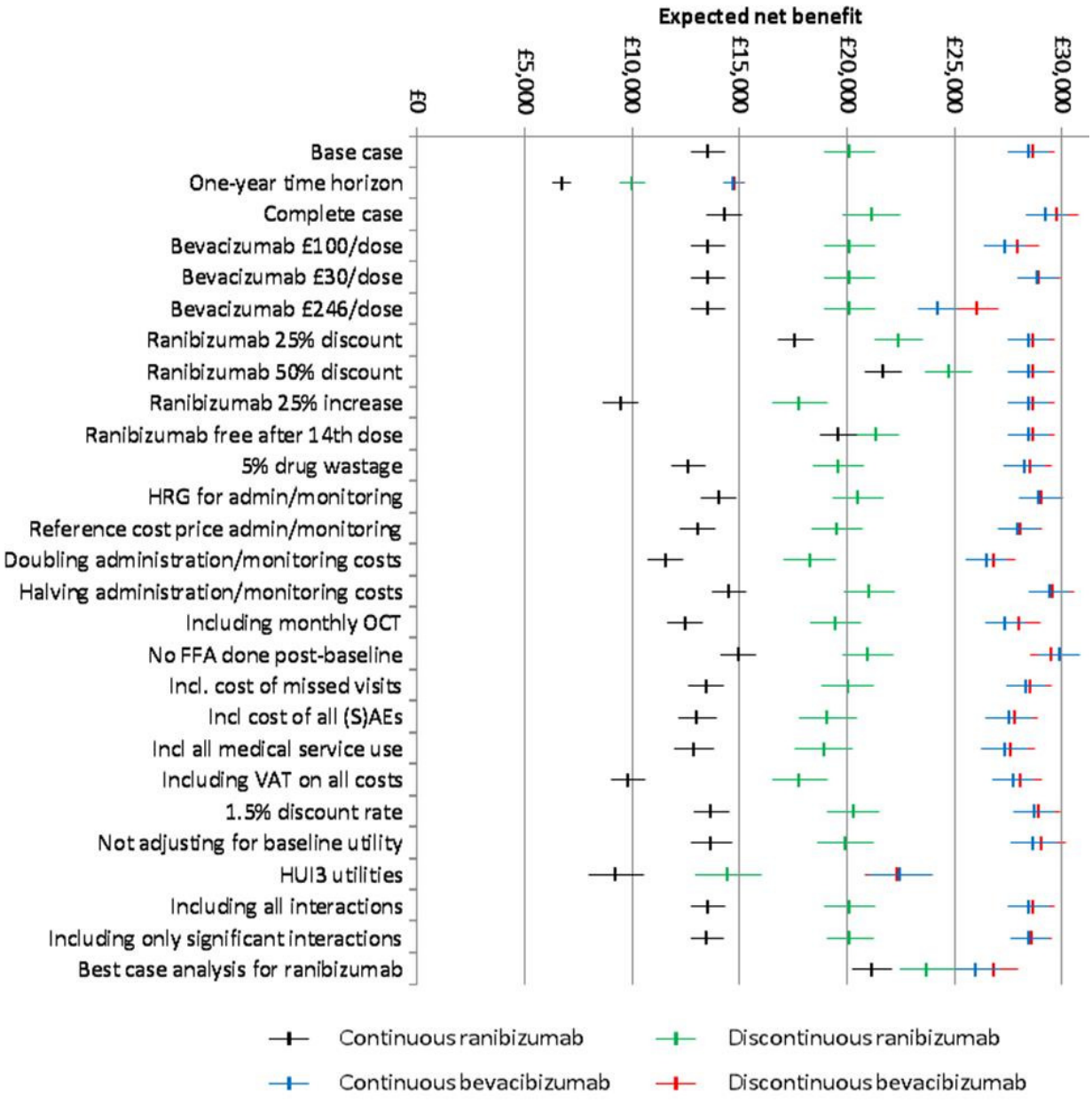


Figure 3: Effect of sensitivity analyses on total net benefits for each of the four treatment arms, assuming a £20,000/QALY ceiling ratio. Treatments that are more cost-effective have higher net benefits; the treatment furthest to the right is therefore most cost-effective, while the treatment furthest to the left is the least cost-effective. Error bars represent 95% confidence intervals.



Appendix: Additional details on the statistical analysis

Definition of expected AEs or SAEs

The base case analysis focused on resource use associated with the study eye or associated with adverse events (AEs) or serious adverse events (SAEs) that were "*expected*": i.e. previously linked to anti-VEGF treatment. The list of AEs and SAEs continued to be "expected" was based on the IVAN trial protocol.[1]

The following were considered to be expected SAEs within the economic evaluation: angina pectoris; arthralgia; cardiac arrest; cardiac failure; cardiovascular disorder; cataract traumatic; cerebrovascular accident; coronary artery bypass; deep vein thrombosis; endophthalmitis; haemorrhage; intraocular pressure increased; left ventricular failure; myocardial infarction; nausea; pulmonary embolism; retinal detachment; retinal pigment epithelial tear; retinal vein occlusion; transient ischaemic attack; upper respiratory tract infection; urinary tract infection; and uveitis.

The following AEs were considered to be expected: angina pectoris; arthralgia; bronchitis; cardiac disorder; cataract; cataract cortical; cataract nuclear; cataract operation; cataract traumatic; conjunctival haemorrhage; cough; eye inflammation; eye irritation; eye pain; haemorrhage; hallucination, visual; headache; hypertension; influenza; intraocular pressure increased; lacrimation increased; nasopharyngitis; nausea; pulmonary embolism; retinal detachment; retinal pigment epithelial tear; retinal vein occlusion; sinusitis; transient ischaemic attack; upper respiratory tract infection; urinary tract infection; uveitis; visual impairment; vitreous detachment; and vitreous floaters.

Measurement and valuation of health benefits

Mixed models were used to estimate the rate at which patients' EQ-5D utility improves after SAEs or reductions in visual acuity. For patients who experienced an SAE that reduced EQ-5D utility, models assumed that EQ-5D utility fell on the day of the SAE and rose linearly afterwards. Similar profiles have previously been used to model recovery from acute hepatitis[2] and chronic obstructive pulmonary disease exacerbations.[3] We focused on linear

recovery profiles to simplify subsequent QALY calculations and as models with quadratic recovery curves did not fit the data as well as those with linear profiles.

Mixed models were estimated on all post-baseline utility measurements using the xtmixed command in Stata. A basic model was defined and a pre-specified series of variations on this model were evaluated and included in the base case analysis if they reduced Akaike's information criterion (AIC). The final model divided SAEs into four categories:

- Ocular (including reductions in visual acuity, increased intraocular pressure and all SAEs in the "eye disorders" MedDRA category)
- Cardiovascular (including all SAEs classed as "cardiac disorders", plus cerebrovascular accident, coronary artery bypass, deep vein thrombosis, haemorrhage, pulmonary embolism and transient ischaemic attack)
- Cancer (comprising all events in the "Neoplasms benign, malignant and unspecified" MedDRA category)
- Other (all events not falling into one of the previous four categories)

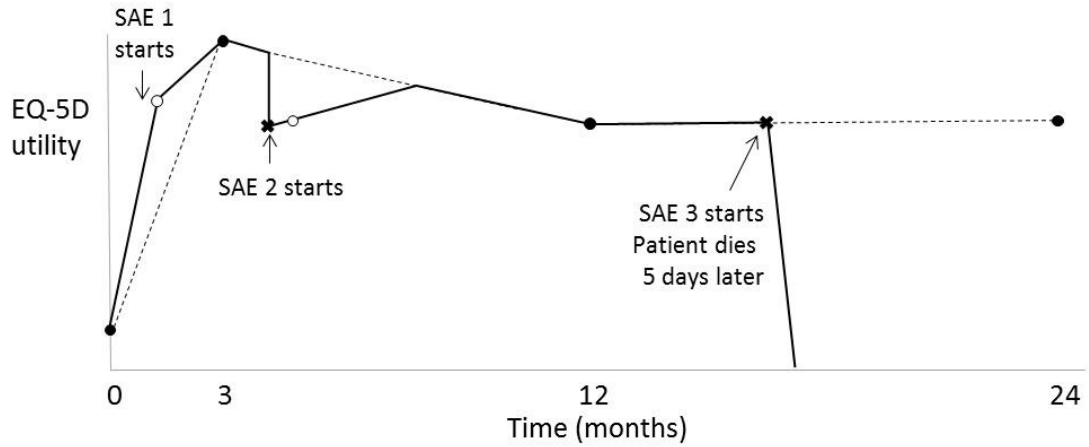
The model assumed that each type of SAE that patients had experienced reduced the EQ-5D utility of patient i at time j by $\beta_{Event,i}$, but that EQ-5D utility rose by a certain amount ($\beta_{EventRecovery}$) with each day that passed after each type of SAE. EQ-5D utility was also assumed to be a function of time since randomisation, treatment and baseline EQ-5D utility (centred by subtracting the mean baseline EQ-5D utility across all patients):

$$\begin{aligned}EQ\text{-}5D_{ij} = {} & Constant_i + \beta_{BL}.(BLEQ5D_{ij}\text{-}MeanBLEQ5D) + \beta_{Time,j}.Time_{ij} \\ & + \beta_{Bevacizumab}.Bevacizumab_i + \beta_{Discontinuous}.Discontinuous_i \\ & + \beta_{Interact}.Bevacizumab_i.Discontinuous_i \\ & + \beta_{CVD,i}.CVD_{ij} + \beta_{CVDRecovery}.TimeSinceCVD_{ij} \\ & + \beta_{Ocular,i}.Ocular_{ij} + \beta_{OcularRecovery}.TimeSinceOcular_{ij} \\ & + \beta_{Cancer,i}.Cancer_{ij} + \beta_{CancerRecovery}.TimeSinceCancer_{ij} \\ & + \beta_{Other,i}.Other_{ij} + \beta_{OtherRecovery}.TimeSinceOther_{ij}\end{aligned}$$

The slopes estimated in the mixed models (e.g. $\beta_{CVDRecovery}$) were used alongside the observed EQ-5D measurements for each patient to estimate EQ-5D utility on the day the SAE started and identify the point at which EQ-5D utility returned to the level that would be expected from the EQ-5D utility measurements that were not taken after SAEs (Figure A). However, some post-SAE measurements were higher than would have been expected from

the other measurements for that patient (e.g. Figure A); in these cases, we assumed that EQ-5D utility changed linearly between the routine measurements (Figure A). For patients dying 1-7 days after the latest SAE, EQ-5D utility was assumed to fall linearly to 0 between the date the SAE started and the date of death. Further details will be reported in *Health Technology Assessment*.

Figure A Illustration of the utility profile around SAEs. EQ-5D utility measurements after SAEs are shown in white circles, while scheduled measurements are shown in black circles. The EQ-5D utility measurement after this patient's first set of SAEs is higher than would be expected from the baseline and three-month measurements; we therefore assumed that EQ-5D utility rose linearly from baseline to the post-SAE measurement and from this onto the 3-month measurement. EQ-5D utility is lower after their second set of SAEs; here, we use the slope coefficients from the mixed model that show the rate of recovery after the categories of SAE that this patient has experienced to draw a line through the post-SAE 2 measurement and estimate EQ-5D utility on the day SAE 2 starts and the time and EQ-5D utility at which the patient is expected to have recovered from the SAE and returned to the EQ-5D utility trend between visits three and 12. The patient died five days after SAE 3; their EQ-5D utility was therefore assumed to follow the linear trend observed between visit 12 and the value imputed at visit 24 up until the day before SAE 3, and then fall linearly to zero between that date and the date of death.



Statistical methods

The economic evaluation used linear regression models with nonparametric bootstrapping, Kaplan-Meier sample averaging and Rubin's rule to combine the quarterly costs and QALYs accrued by each patient to estimate mean total costs and mean QALYs for each of the four study arms.

Thirty-two ordinary least squares regression analyses[86] were used to predict the drug costs, administration/monitoring costs, medication/medical service use costs and QALYs accrued in each quarter conditional on treatment regimen and drug. Interactions between drug and treatment regimen were included as additional independent variables for quarters 2-8 if

[86] 32 = four variables multiplied by eight quarters.

they were either statistically significant or larger than main effects.[87] Since all patients received monthly injections at visits 0-2, we assumed no interaction and no impact of treatment regimen during quarter 1. Analyses of QALYs also controlled for baseline utility to eliminate any bias that could result from imbalance in baseline utility.[4]

We used non-parametric bootstrapping to quantify the uncertainty around quarterly costs and QALYs, allowing for the skewed, heteroskedastic distributions and correlations between outcomes.[5] Bootstrapping involved sampling patients with replacement from each randomised group and estimating all regressions on each bootstrap sample. We also allowed for uncertainty around multiple imputation by generating 100 imputed datasets, each with different values drawn from the imputation model. Uncertainty around consultation costs and the rate of recovery from SAEs was taken into account by randomly sampling values from the relevant distributions for each imputed data set. Bootstrap samples were drawn 130 times for each of the 100 imputed datasets, generating 13,000 bootstrap estimates of mean quarterly costs and QALYs for each of the four study groups, which allow for uncertainty around imputed utilities, the rate of recovery from SAEs and consultation costs.

We also allowed for patients withdrawing early from the trial using Kaplan-Meier sample averaging, whereby costs and outcomes in each quarter are multiplied by Kaplan-Meier estimates of the probability of patients remaining alive at the start of each quarter and summed over all four quarters.[5 6] Kaplan-Meier estimates were adapted to prevent chance differences in numbers of deaths unrelated to treatment[88] affecting incremental QALYs by adding the overall probability of deaths unlikely/not related to study medication (averaged across all four arms) to the probability of potentially-drug related deaths that was observed in each arm. After weighting quarterly costs and QALYs by the Kaplan-Meier estimate of the proportion of patients alive at the start of the quarter and discounting costs incurred in Year 2 by 3.5%, quarterly costs and QALYs were added up to give the total cost and total QALYs accrued in each treatment group over the two-year trial period. The 100 imputed datasets were

[87] Analyses were replicated with and without interactions for drug costs, administration/monitoring costs, medication/medical service use costs and QALYs to identify any interactions that were statistically significant or had an absolute magnitude larger than either the main effect for treatment regimen or the main effect for drug. Interactions that were either statistically significant or larger than either main effect were included in the base case analysis to ensure that the bias associated with omitting qualitative interactions did not change the conclusions.

[88] The five causality groups that study investigators classified all SAEs into were used to categorise deaths into those definitely/probably/possibly related to study medication (referred to as potentially drug-related deaths) and those unlikely to be/not related to study medication (referred to as unrelated deaths).

combined using Rubin's rule[7] to estimate total and incremental costs, QALYs and net benefits and their standard errors (SE). Rubin's rule was implemented in Microsoft Excel, while all other statistical analyses were conducted in Stata version 12.

References

1. The IVAN Study Investigators, Chakravarthy U, Harding SP, et al. Ranibizumab versus Bevacizumab to Treat Neovascular Age-related Macular Degeneration: One-Year Findings from the IVAN Randomized Trial. Ophthalmology 2012;119(7):1399-411.
2. Luyten J, Marais C, Hens N, et al. Imputing QALYs from single time point health state descriptions on the EQ-5D and the SF-6D: a comparison of methods for hepatitis a patients. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research 2011;14(2):282-90.
3. Spencer M, Briggs AH, Grossman RF, et al. Development of an economic model to assess the cost effectiveness of treatment interventions for chronic obstructive pulmonary disease. PharmacoEconomics 2005;23(6):619-37.
4. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ 2005;14(5):487-96.
5. Gray A, Clarke P, Wolstenholme J, et al. Applied Methods of Cost-Effectiveness Analysis in Health Care Oxford: Oxford University Press, 2011.
6. Hallstrom AP, Sullivan SD. On estimating costs for economic evaluation in failure time studies. Med Care 1998;36(3):433-6.
7. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30(4):377-99.

Appendix J: Results of the inside-the-table subgroup analysis on the KAT trial

Methods for Analysis 3: Secondary analysis of interactions as subgroup analysis

In addition to analysing KAT at-the-margins (Analysis 1) and inside-the-table on those patients randomised to >1 comparison (Analysis 2), a sensitivity analysis was conducted using the subgroup approach described in Section 7.1.

Analysis 3 comprised a subgroup analysis of all patients randomised to either Comparison A or B, evaluating the impact of mobile bearings and metal backing in patients with and without randomisation in the patella resurfacing comparison (C). Whereas Analyses 1 and 2 were conducted on a strict ITT basis, for Analysis 3, it was necessary to subdivide patients based on actual patella treatment received, rather than randomised allocation, since patella resurfacing was not randomly allocated in 63% (581/919) of patients in Comparisons A or B. Patella treatment received was identified from the components used in the procedure; seven patients (0.76%) missing all component codes were excluded from this analysis as it was impossible to assess whether or not the patella was resurfaced.

A dummy for randomised allocation in Comparison A or B, a dummy for whether or not the patella was resurfaced, an interaction between these two dummies and (for QALYs) baseline utility comprised explanatory variables in the OLS models used to predict costs or QALYs accrued in each year. Total costs and QALYs were calculated from those predicted in each year using IPW. The two comparisons were analysed separately, with regression analyses being repeated on 100 bootstrap replicates of each of the 100 imputed datasets.

The impact of interactions was evaluated in the same way as for Analysis 2 (Section 7.3.2). The opportunity cost of ignoring interactions and the probability that at-the-margins analysis gives misleading conclusions compared the results of Analysis 3 with at-the-margins analysis conducted on the same set of patients included in Analysis 3.

Results of Analysis 3: Subgroup analysis of interactions

Analysis 3 evaluated interactions in a larger patient group by comparing the incremental effect of metal backing or mobile bearings in patients who received patella resurfacing with the incremental effect in patients who received no resurfacing. Since this analysis included all but seven patients randomised in Comparison A, the at-the-margins estimate of the effect of mobile vs. fixed bearings (averaged across patients with patella resurfacing and those without) in this sample was almost identical to that from Analysis 1 (Table J1). However, since 19 patients did not receive their allocated patella treatment and this analysis includes only 12% (193/1671) of patients randomised in Comparison C, the costs and benefits for the patella comparison differ substantially from the base case analysis. Most SEs were slightly larger than those observed in Analysis 1, but smaller than those in Analysis 2.

Table J1. Subgroup analysis of interactions: including all patients randomised to comparisons A or B

		Mobile bearing	Fixed bearing	Metal backing	All-polyethylene
Patella resurfaced	No. pts	121	125	99	98
	Cost	£8,413 (£234)	£9,289 (£763)	£8,843 (£426)	£8,134 (£548)
	QALYs	5.131 (0.187)	5.141 (0.179)	5.290 (0.170)	4.873 (0.202)
	NMB*	£94,202 (£3,815)	£93,526 (£3,857)	£96,949 (£3,527)	£89,319 (£4,247)
No patella resurfacing	No. pts	140	128	98	103
	Cost	£9,522 (£545)	£8,555 (£349)	£7,620 (£321)	£8,329 (£436)
	QALYs	4.872 (0.182)	4.768 (0.178)	5.179 (0.202)	5.025 (0.186)
	NMB*	£87,918 (£3,882)	£86,810 (£3,651)	£95,969 (£4,132)	£92,164 (£3,826)

* At a £20,000/QALY ceiling ratio.

Qualitative interactions between Comparisons A and C were observed for costs (-£1,842 [SE: £1,024]; p=0.72) and QALYs (-0.11 [SE: 0.31]; p=0.07), in addition to quantitative interactions for NMB (-£433 [SE: £6,680]; p=0.95). This analysis therefore suggests that the bias within at-the-margins estimates is substantially smaller than is suggested by Analysis 2 (-£216, vs. £10,334). However all interaction terms were substantially smaller than those estimated in Analysis 1 and had smaller SEs. In all imputed datasets, allowing for interactions in costs and QALYs increased AIC, while allowing for interactions in NMB decreased AIC. The interactions meant that mobile bearings were less costly and less effective than fixed bearings in patients who had patella resurfacing, but more costly and more effective in those without resurfacing. However, mobile bearings with patella resurfacing nonetheless had highest expected NMB at a £20,000/QALY ceiling ratio. Since this treatment had lower expected NMB than in Analysis 2, the uncertainty about which treatment was best was substantially greater, despite the larger sample size and smaller SEs: there was a 52% chance that mobile bearings with patella resurfacing was the best treatment (Figure J1A; cf. 86% in Analysis 2) and the EVPI for the whole decision was £1,901/patient (Figure J2A; cf. £545 in Analysis 2).

Figure J1. CEACs for multiple comparisons based on secondary (subgroup) analysis of interactions: including all patients randomised in comparisons A or B. **A**: mobile bearing. **B**: metal backing.

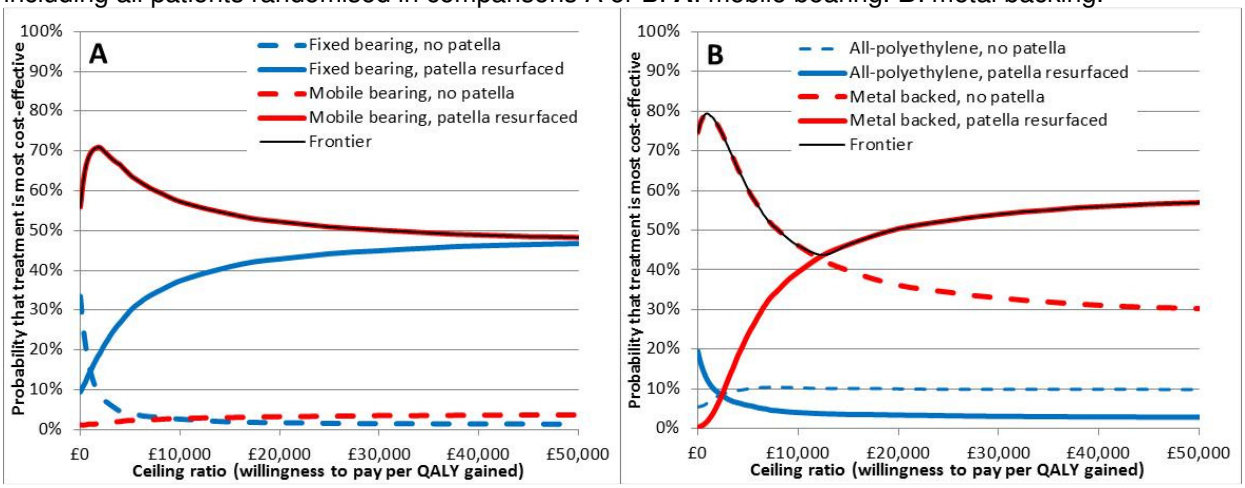
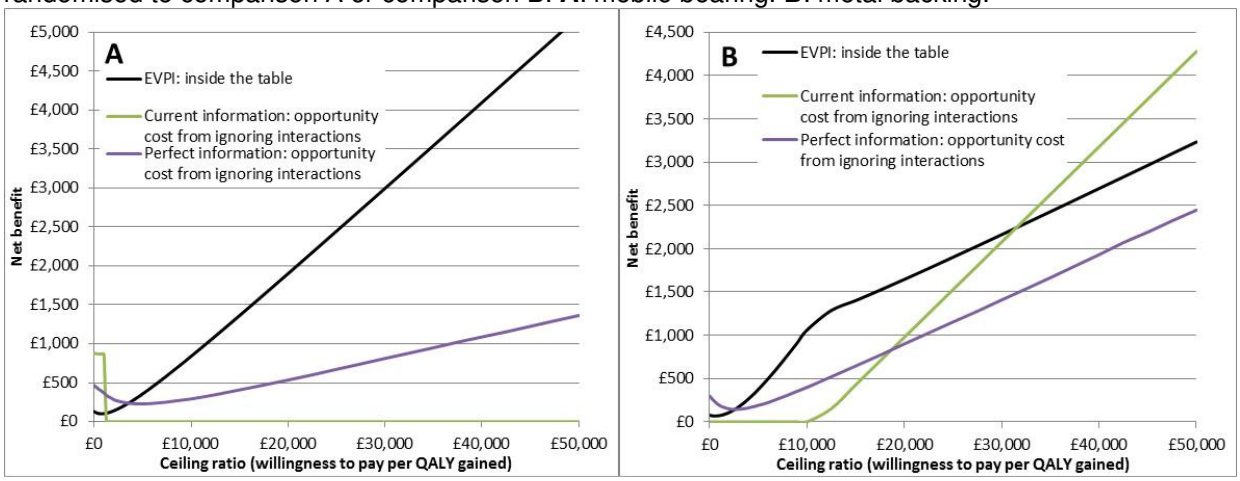


Figure J2. EVPI and expected opportunity cost of ignoring interactions under current and perfect information based on the secondary (subgroup) analysis of interactions: including all patients randomised to comparison A or comparison B. **A**: mobile bearing. **B**: metal backing.



Inside-the-table analysis and at-the-margins analysis gave the same conclusions and the value of making a joint decision was zero at ceiling ratios >£1,000/QALY, with the two analyses differing in 24% of cases at a £20,000/QALY ceiling ratio. The expected value of making a joint decision at this ceiling ratio was £525.

For the metal backing comparison, SEs estimated in Analysis 3 (Table J1) were substantially larger than those observed in Analysis 1, but generally smaller than those from Analysis 2. While the incremental costs and outcomes for metal backing vs. no metal backing matched Analysis 1, very different results were obtained for the patella resurfacing comparison: in the subset of patients randomised in Comparison B, those patients who received a patella accrued slightly fewer QALYs on average than those who did not undergo resurfacing.

There were qualitative interactions for costs (£1,419 [SE: £878]; p=0.11), QALYs (0.26 [SE: 0.32]; p=0.41) and NMB (£3,824 [SE: £6,698]; p=0.57), although these were generally smaller than those in Analysis 2 and not statistically significant, despite the substantially larger sample size. This analysis suggests that the bias within at-the-margins NMB estimates could be £1,912 (£3,824/2). The qualitative interaction for NMB meant that patella resurfacing was cost-effective in patients having metal backing, but not in those with all-polyethylene components, although metal backing was cost-effective regardless of patella resurfacing.

Metal backing with patella resurfacing had highest expected NMB; this conclusion matches the conclusion that was drawn from separate consideration of the two comparisons in the base case analysis, but differs from the conclusions of Analysis 2 (which suggested that all-polyethylene with no patella resurfacing was best). However, there was also substantial uncertainty around the best treatment, which was largely driven by uncertainty about patella resurfacing: the probability that metal backing with patella resurfacing had highest NMB was just 50%, with a 36% probability that the treatment with second highest expected NMB (metal backing without patella resurfacing) was best (Figure J1B). The EVPI for the joint decision was £1,645/patient (Figure J1B): just over half of that for Analysis 2 (£3,094). Since at-the-margins analysis on this sample found metal backing and no patella resurfacing to be cost-effective, the opportunity cost of ignoring interactions was £980 (£96,949-£95,969) under current information and £900 under perfect information. The probability that at-the-margins analysis gave different conclusions was 35%.

However, this analysis analysed the patella comparison as-treated, not as-randomised, which introduces substantial selection bias. As result, any observed interactions may be due

to patient characteristics (e.g. age, physical activity or severity of bone damage) that affect the chance of surgeons choosing to undertake patella resurfacing as well as the costs and QALYs accrued over the time horizon. More fundamentally, we also cannot draw unbiased conclusions about the causative relationship between patella resurfacing and costs or QALYs because this analysis is per-protocol and in most patients the decision to resurface (or not) was based on patient characteristics and surgeon preference, rather than random allocation. This bias and the large interactions mean that Analysis 3 finds patella resurfacing to be (on average) dominated by no patella resurfacing in those patients randomised in Comparison B, whereas Analyses 1 and 2 find patella resurfacing to be dominant. The main effect for the patella comparison is prone to similar selection bias as observational studies, which means that this analysis cannot usefully inform joint decisions about patella resurfacing and tibial design. The impact of this bias could be even greater for partial factorial designs in other clinical areas, such as those where only patients with high blood pressure are randomised to receive antihypertensive treatment or placebo.

Appendix K: Further details on the AFORRD economic evaluation

Type II diabetes is characterised by polypharmacy: in addition to requiring medications to improve glycaemic control, guidelines also recommend that moderate/high-risk patients also receive medications, such as statins and antihypertensives, to reduce the risk of cardiovascular events (399, 400). Although there is extensive evidence demonstrating that statins are highly effective and cost-effective at reducing LDL levels and preventing cardiovascular events (297, 298, 401), there was limited evidence in diabetic patients typical of those managed in primary care at the time the trial was funded. Omega-3 is not currently recommended (400), although there is evidence that it significantly reduces triglyceride levels in patients with type 2 diabetes (402).

Methods for estimating within-trial costs and QALYs

Costing analyses used an index year of 2012. List prices (403) for study medication (Table K1) were applied. The within-trial cost analysis included NHS costs accrued in the first 16 weeks after randomisation. Participants' GPs provided details of concomitant medication at baseline and reported all changes in concomitant medication occurring during the trial. Concomitant medications received between randomisation and the 16-week visit were valued using current prices (404); this was based on the average cost of specific medications used to treat diabetes and diabetes-related complications (e.g. cardiovascular disease), while the cost of all other medications was based on the average cost per prescription. Trial participants completed resource use questionnaires providing numbers of consultations with general practitioners, practice nurses and outpatient clinics over the 52-week trial period; consultation counts were divided by 3.25 (52÷16) to estimate resource use during the first 16 weeks and were valued using nationally-available costs (Table K1). All hospitalisations were identified through AE case report forms; hospitalisations occurring between randomisation and the 16-week visit were included in the analysis and valued using Department of Health reference costs for the relevant inpatient spell (405).

Table K1. Resource use and unit costs accrued during the 16-week trial period

Description	Unit cost	Source of unit cost
20 mg atorvastatin (non-proprietary, per day)	£0.06	£1.60 per 28-tablet pack (403)
2 g omega-3 EE90 (Omacor®, per day)	£1.02	£50.84 per 100 1g capsules (403)
Behavioural leaflet	£5.27	The leaflet comprised one A4 sheet printed on one side. Within the trial, it was posted to participants along with other questionnaires two weeks after start of treatment. However, for the purposes of the economic evaluation, I assumed that a clinical support worker or receptionist would post the leaflet to patients separately using a second-class stamp and enclosing a second stamped addressed envelope for the patient to return their action plan. The leaflet intervention was therefore assumed to cost £5.27: the price of two second-class stamps from April 2012 (£1.00), plus an estimate of the price of the two envelopes and printing (£0.10) and 10 minutes of patient-related work for a community clinical support worker (£4.17 (406)). The cost of the leaflet was assumed to be incurred regardless of what if any active

Description	Unit cost	Source of unit cost
		treatment patients were receiving.
GP home visit	£104.00	Cost per out of surgery visit lasting 23.4 minutes: includes qualification costs, and travel, but excludes direct staff costs (since these are captured in nurse visits) (406).
GP surgery consultation	£41.00	Cost per patient contact lasting 11.7 minutes (mean length of surgery consultation): includes qualification costs, but excludes direct staff costs, since nurse visits are costed separately (406).
GP telephone consultation	£25.00	Cost per telephone consultation lasting 7.1 minutes (mean length of telephone consultation): includes qualification costs, but excludes direct staff costs since nurse visits are costed separately (406).
Nurse home visit	£22.29	Cost per hour of face-to-face contact (£52 including qualifications) for a GP nurse, multiplied by the mean length of GP home visits (23.4 minutes) (406). Added cost of 3 miles' travel at £0.67/mile (3 miles is assumption, which is consistent with the travel cost of £1.50 in PSSRU 2010 (407)).
Nurse consultation	£13.43	Cost per hour of face-to-face contact (£52 including qualifications) for a GP nurse, multiplied by the mean length of GP nurse surgery visits (15.5 minutes) (406).
Nurse telephone consultation	£6.15	Cost per hour of face-to-face contact (£52 including qualifications) for a GP nurse, multiplied by the mean length of GP phone consultation (7.1 minutes) (406).
Health visitor home visit	£49.87	Assumed to be the same as a standard health visitor consultation, since health visitors most commonly do home visits.
Health visitor consultation	£49.87	Department of Health reference costs (405), CHSHVM sheet: national average cost for Health Visitor Core Services, Face to Face, One to One
Health visitor telephone consultation	£23.23	Department of Health reference costs (405), CHSHVM sheet: national average cost for Health Visitor Core Services, Non Face to Face
Dietician home visit	£70.74	Assumed to be the same as a standard dietician consultation
Dietician consultation	£70.74	Department of Health reference costs (405), CHSAHP sheet: national average cost for dietician
Dietician telephone consultation	£43.14	Assumed to cost 61% of the price of a surgery consultation, since GP telephone consultations cost 61% of a surgery consultation
Chiropodist home visit	£41.92	Assumed to be the same as a standard podiatry consultation
Chiropodist consultation	£41.92	Department of Health reference costs (405), CHSAHP sheet: national average cost for podiatrist
Chiropodist telephone consultation	£25.56	Assumed to cost 61% of the price of a surgery consultation, since GP telephone consultations cost 61% of a surgery consultation
Optician visit	£20.90	NHS sight test fee from 1st April 2013 to 31st March 2014 (http://www.fodo.com/resource-categories/nhs-sight-test-fees)
Telephone optician	£0.00	Applied no cost based on the assumption that high-street opticians will not be able to claim any money back from the NHS for telephone conversations
Hospital eye clinic visit	£85.96	Weighted average of all non-admitted face-to-face outpatient attendances for "Ophthalmology", "Medical Ophthalmology", or "Optometry": both consultant-lead and non-consultant lead. CL and NCL sheets from

Description	Unit cost	Source of unit cost
		Department of Health reference costs (405).
Telephone hospital eye clinic	£62.03	Weighted average of all non-admitted non-face-to-face outpatient attendances for "Ophthalmology", "Medical Ophthalmology", or "Optometry": both consultant-lead and non-consultant lead. CL and NCL sheets from Department of Health reference costs (405).
Outpatient visit	£105.15	Weighted average of all non-admitted face-to-face outpatient attendances outside of paediatric or child/adolescent departments: both consultant-led and non-consultant led. CL and NCL sheets from Department of Health reference costs (405).
Telephone outpatient	£57.81	Weighted average of all non-admitted non-face-to-face outpatient attendances outside of paediatric or child/adolescent departments: both consultant-led and non-consultant led. CL and NCL sheets from Department of Health reference costs (405).
Hospitalisations by indication (cost per hospital stay unless otherwise specified)		
Atrioventricular block	£3,051.04	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "Pace 1: Single Chamber or Implantable Diagnostic Device" OR "Pace 2: Dual Chamber"
Cataract	£2,074.47	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the word "cataract"
Cataract nuclear	£2,074.47	
Cataract operation	£2,074.47	
Circumcision	£1,468.62	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "Minor Penis Procedures, 19 years and over"
Chest pain	£2,686.34	Department of Health reference costs (405). NEI_L sheet. Weighted average of all HRGs
Cough	£2,686.34	
Diarrhoea	£2,686.34	
Joint swelling	£2,686.34	
Rectal haemorrhage	£2,686.34	
Femur fracture	£3,567.87	Department of Health reference costs (405). NEI sheet. Weighted average of all HRGs containing the phrase "Pathological Fractures"
Headache	£2,686.34	Department of Health reference costs (405). NEI_L sheet. Weighted average of all HRGs
Hip arthroplasty	£6,268.21	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "Major Hip Procedures for Non-Trauma"
Inguinal hernia	£2,520.36	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "hernia"
Excision of malignant [skin] tumour or basal cell carcinoma	£1,232.84	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "Minor Skin Procedures, 13 years and over"
Myocardial infarction	£2,154.95	Department of Health reference costs (405). NEI_L sheet. Weighted average of all HRGs containing the phrase "Actual or Suspected Myocardial Infarction
Knee replacement for osteoarthritis	£5,514.80, plus £320.99/day after day 11	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "Major Knee Procedures for Non-Trauma"
Bowel cancer	£7,035.78	Department of Health reference costs (405). EI sheet. Weighted average of all HRGs containing the phrase "Large Intestine Procedures", excluding those on under 19's

Conditional mean imputation was used to impute the cost of within-trial ambulatory consultations for the 122 patients who attended the week-16 visit but did not complete the 52-week questionnaire on numbers of consultations.[89] This simplification resulted in little/no efficiency loss, since all but one such patients were missing all 16-week data. Ex-smokers were less likely to have missing 16-week data (p=0.038), although neither costs the year before randomisation nor other baseline data used as prognostic variables in the UKPDS model significantly predicted missingness in logistic regression (p≥0.096). However, patients randomised to omega-3 (p=0.045) and/or leaflet (p=0.038) were significantly less likely to have missing data at week 16. Although it is plausible that the leaflet increased patients' motivation to continue trial medication, it is unclear why allocation to omega-3 would affect missing data, or to what extent this introduced bias.

Trial participants completed EQ-5D questionnaires at baseline and 52 weeks. Outcomes at 52 weeks were not included in QALY calculations, since these will be affected by the addition of atorvastatin and/or behavioural reinforcement beyond Week 16. Within-trial QALYs for each patient were calculated by multiplying patients' baseline utility by 16/52, then centring by subtracting the mean QALYs for the relevant cell of the 2x2x2 factorial design and adding on the grand mean across the whole sample. Within-trial QALYs were assumed to equal the grand mean for the 36 patients with missing baseline EQ-5D utility. This centring method removes any baseline imbalance in EQ-5D utilities, while ensuring that within-trial QALYs vary between patients and that the analysis uses a 40.3-year time horizon for both QALYs and costs. However, it means that all study arms have the same mean QALYs during the 16-week within-trial period.

Methods for extrapolating outcomes and analysing results

Costs, QALYs and life-years accrued beyond Week 16 were discounted at 3.5% per annum (145). HDL, total cholesterol, glycated haemoglobin (HbA1c), systolic blood pressure and smoking at week 16 were entered as current values; values recorded two weeks before randomisation were used as proxies for values at diagnosis. Age and smoking status at diagnosis were inferred from the year of diagnosis, date of birth and smoking history. Patients were assumed to have no history of vascular events or peripheral vascular disease at randomisation, although the presence/absence of atrial fibrillation was based on case report forms. The analyses presented assumed that all risk factors would remain constant over the 40-year extrapolated period on the assumption that medication doses will be adjusted to maintain risk.

Costs and utilities without diabetic events and the costs and HRQoL reductions associated with diabetic events were based on those observed in UKPDS (408, 409); costs were inflated to 2012 values using the Hospital and Community Health Services pay and prices index (410).

[89] For the 106 patients who completed resource use questionnaires at baseline but not week 52, the cost of consultations attended by each patient in the year before enrolment was added to the mean change in consultation costs between the baseline and 52-week questionnaires specific to each study arm. The overall mean consultation costs for that study arm was applied to the 16 patients who were missing resource use questionnaires at both baseline and week 52. For all analyses, consultation costs were imputed using mean costs in each of the eight study arms of the 2x2x2 design. Missing data on individual questions was imputed as zero based on the assumption that patients would have entered a non-zero value if they had visited this type of healthcare professional.

Primary analyses without covariates

The UKPDS Outcomes Model estimates lifetime costs and QALYs for M sets of risk equation parameters, which were estimated by bootstrapping the original UKPDS sample. The model records the mean costs, QALYs and life-years for each bootstrap replicate, averaged across patients, and the mean extrapolated value for each patient ($\bar{y}_i$) and its 95% CI ($\bar{y}_{i\pm} z_{0.05}\sqrt{\mathrm{var}_B(\bar{y}_i)}$, which captures uncertainty around risk equation parameters) (292).

For each patient, I estimated the mean discounted cost of study medication accrued during the extrapolated period ($Cost_{Drug,i}$) by multiplying the annual cost of medication within that study arm ($Cost_{\mathrm{Annual}Drug,i}$) by the expected number of discounted life-years in the extrapolated period (LE_i). The SE around extrapolated drug costs for each patient was assumed to equal the SE around the discounted life-years, multiplied by $Cost_{\mathrm{Annual}Drug,i}$.

Total cost for each patient ($Cost_{Total,i}$) was then calculated as the sum of within-trial cost ($Cost_{WT,i}$), complication cost during the 40-year extrapolation ($Cost_{Comp}$) and $Cost_{Drug,i}$

$$\begin{aligned} Cost_{Total,i} &= Cost_{WT,i} + Cost_{Comp,i} + Cost_{\mathrm{Annual}Drug,i}\cdot LE_i \\ Var\left(Cost_{WT,i}\right) &= Var\left(Cost_{Comp,i}\right) + Cost^2_{\mathrm{Annual}Drug,i}\cdot Var\left(LE_i\right) \\ &+ 2\rho_{CostLE}\sqrt{Var\left(Cost_{Comp,i}\right)\cdot Cost^2_{\mathrm{Annual}Drug,i}\cdot Var\left(LE_i\right)} \end{aligned} \qquad \text{(K1)}$$

where the correlation between complication costs and life expectancy (ρ_{CostLE}) was assumed to equal 0.3116 (the correlation between these parameters across the means from the 100 bootstraps).

Total QALYs for each patient equalled the sum of within-trial QALYs plus QALYs over the extrapolated period, with the SE equal to that around extrapolated QALYs.[90] The expected number of vascular events per patient over the 40-year trial period was calculated by summing event rates over all years and events.

NMB for each patient was calculated from the estimated costs and QALYs at a range of ceiling ratios, with the primary analysis interpreted at £20,000/QALY (130). SEs around NMB were calculated as

$$\begin{aligned} \mathrm{var}\left(NMB_{Rc,i}\right) &= Rc^2\cdot\mathrm{var}\left(\mathrm{QALYs_i}\right) + \mathrm{var}\left(Cost_{Total,i}\right) \\ &- 2\cdot\rho_{QC}\cdot Rc\cdot\sqrt{\mathrm{var}\left(\mathrm{QALYs_i}\right)\cdot\mathrm{var}\left(Cost_{Total,i}\right)} \end{aligned} \qquad \text{(K2)}$$

Where the correlation between costs and QALYs (ρ_{QC}) was assumed to equal 0.2559, the correlation between complication costs and QALYs across all 100 bootstraps.[91]

[90] The variance around within-trial costs and QALYs was assumed to be zero since these were directly observed.

[91] This is a simplification that assumes a 1:1 correlation between complication costs and extrapolated drug costs.

For each of the three analyses, group means for each of the variables of interest (*QALYs*, LE, $Cost_{Drug}$, $Cost_{Comp}$, $Cost_{WT}$, $Cost_{Total}$, $NMB_{Rc,i}$ and numbers of events) were calculated by averaging the mean per-patient values ($\bar{y}_i$) over all patients within that group. For example, in Analysis 1 mean QALYs for the atorvastatin-only group equalled the mean QALYs in patients randomised to atorvastatin, no omega-3 and no leaflet, while for Analysis 3 mean QALYs for patients receiving atorvastatin were calculated as the crude mean across all patients randomised to atorvastatin (regardless of whether they received omega-3 or leaflet).

SEs around each group mean ($\sqrt{\text{var}(y)}$) were calculated in Microsoft Excel 2010 using Rubin's rule. Rubin's rule normally combines uncertainty around imputation with sampling uncertainty by adding the mean within-imputation variance to the mean between-imputation variance multiplied by a correction factor (256, 258),

$$\text{var}(\bar{y}) = \frac{1}{M}\sum_{m=1}^{M}\text{var}(y_m) + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{m=1}^{M}(y_m-\bar{y})^2 \quad \text{(K3)}$$

where M is the number of imputed datasets.

Rubin's rule has previously been used to combine parameter uncertainty and sampling uncertainty within previous applications of UKPDS (210). However, the current Excel UKPDS Outcomes Model does not report extrapolated outcomes for each bootstrap on each patient. I therefore adapted Rubin's rule to calculate the mean between-bootstrap variance (B) for any given group, by averaging the patient-specific between-bootstrap variances ($\text{var}_B(\bar{y}_i)$) across all patients in that group.

$$B = \frac{1}{N}\sum_{i=1}^{N}\text{var}_B(\bar{y}_i) \quad \text{(K4)}$$

Between-bootstrap variance was assumed to be zero for within-trial costs and QALYs and for numbers of events.[92]

The within-bootstrap variance (W) was estimated by taking the SD across the mean extrapolated values ($\bar{y}_i$) for each of the N patients in the group and dividing the result by N to give the SE^2.

$$W = \left[\frac{1}{(N-1)}\sum_{i=1}^{N}(\bar{y}_i-\bar{y})^2\right] \div N \quad \text{(K5)}$$

This is assumed to be equivalent to the mean of the between-patient variances across all bootstraps.

The overall variance (taking into account parameter uncertainty and sampling uncertainty) therefore equals

$$\text{var}(\bar{y}) = W + \left(1+\frac{1}{M}\right)B \quad \text{(K6)}$$

where M is the number of bootstraps.

After estimation of group means (y_k) and SEs ($\sqrt{\text{var}(y_k)}$) for QALYs, costs, NMB and numbers of events were calculated using Rubin's rule; the relevant between-

[92] For simplicity, parameter uncertainty around the extrapolation was ignored within analyses of the numbers of diabetic events, since estimating between-bootstrap variance around the sum of 160 correlated random variables (rates of four different events over 40 years) is analytically difficult.

group differences (e.g. Δy_{a-0}) and any interaction terms (I_{AB}) were calculated for each analysis.

$$\Delta y_{a-0} = y_a - y_0 \qquad I_{AB} = y_0 - y_a - y_b + y_{ab}$$
$$I_{ABC} = y_a + y_b + y_c + y_{abc} - y_0 - y_{ab} - y_{ac} - y_{bc} \qquad \text{(K7)}$$

The variances around between-group differences and interactions were calculated from the between-patient variances (W_k) and the between-bootstrap variances (B_k) for the treatment groups used to calculate that difference or interaction term. Between-patient variances (W_k) for different groups were assumed to be independent since they are based on different patients, while the between-bootstrap variances (B_k) were assumed to be correlated between groups since they are based on the same set of risk equation parameters.

In order to simplify calculations and base results on just one run of the UKPDS model, the correlation between bootstraps for any pair of groups was based on the between-bootstrap correlation between mean outcomes for all patients randomised to atorvastatin and the mean outcomes for all patients randomised to no atorvastatin ($\rho_{A\text{-noA}}$).[93] $\rho_{A\text{-noA}}$ equalled:

- 0.9968 for life-years and study medication (the correlation between life-years in the two groups)
- 0.9964 for QALYs (the correlation between QALYs)
- 0.9934 for complication costs and total costs (the correlation for complication costs)
- 0.9955 for NMB (the average of the three aforementioned correlation coefficients).

The variance around differences between any two groups was therefore calculated as:

$$\text{var}(\Delta y_{a-0}) = \text{W}_a + \text{W}_0 + \left(1 + \frac{1}{\text{M}}\right)\left(B_a + B_0 - 2\rho_{A\text{-noA}}\sqrt{B_a \cdot B_0}\right) \qquad \text{(K8)}$$

Two-way interactions equal the difference between two sums of normally distributed variables $I_{AB} = (y_0 + y_{ab}) - (y_a + y_b)$ and three-way interactions equal the difference between two sums of two sums of normally distributed variables $I_{ABC} = ((y_a + y_b) + (y_{abc} + y_c)) - ((y_0 + y_{ab}) + (y_{ac} + y_{bc}))$. We can therefore calculate the variance around a two-way interaction in two steps. Firstly by calculating *B* for each sum of two correlated variables, such as the sum of groups *a* and *b*, as:

$$B_{a+b} = \left(B_a + B_b + 2\rho_{A\text{-noA}}\sqrt{B_a \cdot B_b}\right) \qquad \text{(K9)}[94]$$

[93] The UKPDS Outcomes Model only allows for two treatments groups and presents pairs of bootstrap means for each of these two groups. I therefore ran the model once, defining the groups by atorvastatin allocation, and estimated the correlation between mean outcomes for atorvastatin and mean outcomes for no atorvastatin across all 100 bootstrap replicates using the Excel function CORREL.

[94] Similarly, the sum of groups *0* and *ab* can be calculated as $B_{0+ab} = \left(B_0 + B_{ab} - 2\rho_{A\text{-noA}}\sqrt{B_0 \cdot B_{ab}}\right)$.

Secondly, the between-bootstrap variances for each pair of groups can then be combined one another and with *W* to give the variance around the interaction. For the two-way interaction between A and B, this was calculated as:

$$\mathrm{var}(I_{\mathrm{AB}}) = \mathrm{W}_0 + \mathrm{W}_a + \mathrm{W}_b + \mathrm{W}_{ab} + \left(1+\frac{1}{\mathrm{M}}\right)\left(B_{a+b} + B_{0+ab} - 2\rho_{A\text{-noA}}\sqrt{B_{a+b}\cdot B_{0+ab}}\right) \quad \text{(K10)}$$

For a three-way interaction, I first calculated the between-bootstrap variance for all four sums of two groups using Equation K9, then combined the variances of each of the two pairs of groups ($(y_a + y_b)+(y_{a\mathrm{bc}} + y_c)$ and $(y_0 + y_{ab})+(y_{ac} + y_{bc})$) using Equation K9 and then estimated the variance around three-way interaction as:

$$\begin{aligned}\mathrm{var}(I_{\mathrm{ABC}}) &= \mathrm{W}_0 + \mathrm{W}_a + \mathrm{W}_b + \mathrm{W}_c + \mathrm{W}_{ab} + \mathrm{W}_{ac} + \mathrm{W}_{bc} + \mathrm{W}_{abc} \\ &+ \left(1+\frac{1}{\mathrm{M}}\right)\left(B_{a+b+\mathrm{abc}+\mathrm{c}} + B_{0+ab+\mathrm{ac}+\mathrm{bc}} - 2\rho_{A\text{-noA}}\sqrt{B_{a+b+\mathrm{abc}+\mathrm{c}}\cdot B_{0+ab+\mathrm{ac}+\mathrm{bc}}}\right)\end{aligned} \quad \text{(K11)}$$

Although analytical methods exist for calculating CEACs (16) and EVPI (203) from the INB and its variance for two-arms, analytical methods have not yet been developed for multi-arm trials. However, in principle CEACs and EVPI could be calculated using Monte Carlo. Although it was not possible to implement within the time frames of this thesis, this could be conducted by manually generating variance-covariance matrices for the NMB for each of the three analyses, by entering the SE^2 along the diagonal and the covariance $\mathrm{cov}ar(\Delta y_{a-0})$ between any pair of treatment effect estimates in the remainder of the matrix.

$$\mathrm{cov}\,ar(\Delta y_{a-0}) = \rho_{A\text{-noA}}\left(1+\frac{1}{\mathrm{M}}\right)\left(\sqrt{B_a\cdot B_0}\right) \quad \text{(K12)}$$

The variance-covariance matrix could then be used to generate a large number of normally distributed correlated estimates of the NMB for each of the eight treatment groups considered in Analysis 1, each of the four treatment groups considered in Analysis 2 and each of the three main effects considered in Analysis 3. EVPI and the probability of each treatment having highest NMB could then be calculated for each ceiling ratio and repeated at a range of ceiling ratios.

In principle, the same method could be used to calculate the opportunity cost of ignoring interactions under perfect information and the probability that ignoring interactions changes the conclusions. This would require estimation of a large variance-covariance matrix showing the correlation between outcomes for each treatment-combination within different analyses. In principle, this matrix could be calculated by summing covariances for the variables that make up each treatment-effect estimator. However, this is likely to be difficult to conduct in practice and it was not possible to do so within the timeframes of this thesis.

Meta-regression analyses controlling for covariates

Post hoc analyses controlling for prognostic variables were conducted to evaluate the impact of baseline imbalance. Since there was no easy way to allow for covariates in the primary analysis, meta-regression was used as a simple alternative method for combining between-bootstrap parameter uncertainty and between-patient sampling uncertainty that readily facilitates consideration of covariates. Meta-regression was conducted in Stata using the user-written `metareg` command (411). This command is normally used to control for covariates while pooling results (e.g. log-odds ratios) of RCTs. However, I use it here to pool costs and QALYs across patients. Meta-

regression comprises a form of weighted regression, which pools observations conditional on covariates, weighting observations by their precision and estimating SEs around regression coefficients that take account of the variability between observations and the precision of each observation.

Two random effects meta-regression models were estimated for each of the three analyses: one that included only the relevant treatment indicators and interactions (without covariates); and another that also considered a fixed set of covariates. Only one set of covariates was considered; this comprised all of the UKPDS model inputs that were measured at baseline rather than at Week 16 (namely: white race; female gender; age at diagnosis; duration of diabetes; weight; height; atrial fibrillation; current smoker; total cholesterol; high-density lipoprotein; systolic blood pressure; and glycated haemoglobin, HbA1c). Meta-regression was estimated using a non-iterative moment estimator, since this was the only estimation option that converged with all models.

Meta-regression results should be interpreted with caution, since they give less weight to patients with high between-bootstrap SEs (who tend to be those with long life expectancy) and implicitly assume that outcomes for each patient are independent (ignoring the fact that all patients are estimated using the same sets of bootstrapped risk equations and therefore have highly correlated outcomes).

Meta-regression coefficients were used to estimate mean predicted costs and QALYs for each study arm, which enabled estimation of NMB and ICERs and identification of the treatment maximising NMB.

The opportunity cost of ignoring covariates under current information was calculated in the same way as the opportunity cost of ignoring interactions. In other words for each of the three analyses, I identified the treatment combination that would be adopted based on meta-regression with covariates and the treatment combination that would be adopted based on the primary analysis without covariates. I then subtracted the covariate-adjusted estimate of the NMB for the treatment adopted based on the primary analysis from the maximum covariate-adjusted NMB. This was repeated at a range of ceiling ratios to give the results in Figure 8.4.

References for all chapters and appendices

1. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA. 2003;289(19):2545-53. Epub 2003/05/22.

2. Lubsen J, Pocock SJ. Factorial trials in cardiology: pros and cons. Eur Heart J. 1994;15(5):585-8.

3. Hung HM. Two-stage tests for studying monotherapy and combination therapy in two-by-two factorial trials. Stat Med. 1993;12(7):645-60. Epub 1993/04/15.

4. Hung HM, Chi GY, O'Neill RT. Efficacy evaluation for monotherapies in two-by-two factorial trials. Biometrics. 1995;51(4):1483-93.

5. Ng T. The impact of a preliminary test for interaction in a 2 x 2 factorial trial. Proceedings of the Biopharmaceutical Section of the American Statistical Association, Alexandria, VA. 1991:220-7.

6. Armitage P, Berry G, Mathews JNS. Statistical methods in medical research. 4th ed. Malden, MA: Blackwell Science Ltd.; 2002.

7. Couper DJ, Hosking JD, Cisler RA, Gastfriend DR, Kivlahan DR. Factorial designs in clinical trials: Options for combination treatment studies. Journal of Studies on Alcohol. 2005;66(SUPPL. 15):24-32.

8. Blyth K, Gebski V. Factorial designs: a graphical aid for choosing study designs accounting for interaction. Clinical trials. 2004;1:315-25.

9. Finney DJ. Experimental Design and Its Statistical Basis. De Bruyn PPH, editor. London: Cambridge University Press; 1955.

10. Cox DR. Planning of experiments. Wiley Classics Edition published 1992 ed. New York: Wiley; 1958.

11. Montgomery DC. Design and Analysis of Experiments. Sixth ed. New York: John Wiley & Sons, Inc; 2005.

12. Kuntz KM, Weinstein MC. Chapter 7: Modelling in economic evaluation. In: Drummond M, McGuire A, editors. Economic evaluation in health care - merging theory with practice. New York, NY: Oxford University Press; 2001.

13. Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. Gray A, Briggs A, editors. Oxford: Oxford University Press; 2006.

14. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. BMJ. 2011;342:d1766. Epub 2011/04/13.

15. Doshi JA, Glick HA, Polsky D. Analyses of cost data in economic evaluations conducted alongside randomized controlled trials. Value Health. 2006;9(5):334-40. Epub 2006/09/12.

16. Gray A, Clarke P, Wolstenholme J, Wordsworth S. Applied Methods of Cost-Effectiveness Analysis in Health Care Gray A, Briggs A, editors. Oxford: Oxford University Press; 2011.

17. Glick HA, Doshi JA, Sonnad SS, Polsky D. Economic evaluation in clinical trials. Gray A, Briggs A, editors. Oxford: Oxford University Press; 2007.

18. Ramsey S, Willke R, Briggs A, Brown R, Buxton M, Chawla A, et al. Good Research Practices for Cost-Effectiveness Analysis Alongside Clinical Trials: The ISPOR RCT-CEA Task Force Report. Value Health. 2005;8(5):521-33.

19. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ. 2011;342:d1548. Epub 2011/04/09.

20. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol. 2003;3:26.

21. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ. 2010;340:c723. Epub 2010/03/25.

22. Silagy CA, Jewell D. Review of 39 years of randomized controlled trials in the British Journal of General Practice. Br J Gen Pract. 1994;44(385):359-63. Epub 1994/08/01.

23. Holman RR, Paul S, Farmer A, Tucker L, Stratton IM, Neil HA. Atorvastatin in Factorial with Omega-3 EE90 Risk Reduction in Diabetes (AFORRD): a randomised controlled trial. Diabetologia. 2009;52(1):50-9. Epub 2008/11/13.

24. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.2 ed: The Cochrane Collaboration; September 2009.

25. Yates F, Mather K. Ronald Aylmer Fisher. Biographical Memoirs of Fellows of the Royal Society of London. 1963;9:91-120.

26. Yates F. Complex experiments (with discussion). J Roy Statist Soc. 1935;B2:181-247.

27. Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain. 1926;33:503-13.

28. Fisher RA. The Design of Experiments. Eighth edition (first edition published 1935) ed. Edinburgh, London: Oliver and Boyd; 1966.

29. Chalmers TC, Eckhardt RD, Reynolds WE, Cigarroa JG, Jr., Deane N, Reifenstein RW, et al. The treatment of acute infectious hepatitis. Controlled studies of the effects of diet, rest, and physical reconditioning on the acute course of the disease and on the incidence of relapses and residual abnormalities. J Clin Invest. 1955;34(7, Part II):1163-235. Epub 1955/07/01.

30. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. Stat Med. 1989;8(2):161-71. Epub 1989/02/01.

31. Green S, Liu PY, O'Sullivan J. Factorial design considerations. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2002;20(16):3424-30. Epub 2002/08/15.

32. Bernaerts K, Gysemans KP, Nhan Minh T, Van Impe JF. Optimal experiment design for cardinal values estimation: guidelines for data collection. Int J Food Microbiol. 2005;100(1-3):153-65.

33. Piantadosi S. Chapter 19: Factorial trials. Clinical trials: a methodological perspective. Second ed. Hoboken, NJ: John Wiley & Sons Inc; 2005. p. 501-13.

34. Stampfer MJ, Buring JE, Willett W, Rosner B, Eberlein K, Hennekens CH. The 2 x 2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians. Stat Med. 1985;4(2):111-6.

35. Byar DP, Piantadosi S. Factorial designs for randomized clinical trials. Cancer Treat Rep. 1985;69(10):1055-63. Epub 1985/10/01.

36. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied Linear Statistical Models. 4th ed. Chicago, IL: Irwin; 1996.

37. Byar DP. Factorial and reciprocal control designs. Stat Med. 1990;9(1-2):55-63; discussion -4.

38. Sheikh A, Smeeth L, Ashcroft R. Randomised controlled trials in primary care: scope and application. Br J Gen Pract. 2002;52(482):746-51.

39. Korttila K, Apfel CC. Factorial design provides evidence to guide practice of anaesthesia. Acta Anaesthesiol Scand. 2005;49(7):927-9.

40. Slud EV. Analysis of factorial survival experiments. Biometrics. 1994;50(1):25-38.

41. Bria E, Di Maio M, Nistico C, Cuppone F, Terzoli E, Cognetti F, et al. Factorial design for randomized clinical trials. Ann Oncol. 2006;17(10):1607-8.

42. Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. Biometrics. 1997;53(2):456-64.

43. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. Cochrane Database Syst Rev. 2008(2):CD007176. Epub 2008/04/22.

44. Kenyon S, Pike K, Jones DR, Brocklehurst P, Marlow N, Salt A, et al. Childhood outcomes after prescription of antibiotics to pregnant women with spontaneous preterm labour: 7-year follow-up of the ORACLE II trial. Lancet. 2008;372(9646):1319-27. Epub 2008/09/23.

45. Muller KE, Fetterman BA. Regression and ANOVA: An Integrated Approach Using SAS Software. Cary, NC: SAS Institute Inc; 2002.

46. Yates F. The design and analysis of factorial experiments. Harpenden, England: Imperial Bureau of Soil Science. Technical communication No. 35; 1937.

47. Fox Z, Nitsch D, Wang D, Bakhai A. Chapter 11: Factorial design. In: Wang D, Bakhai A, editors. Clinical trials: a practical guide to design, analysis, and reporting. London, UK: Remedica; 2006.

48. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Econ. 2002;11(5):415-30. Epub 2002/07/12.

49. Willan AR, Briggs AH. Statistical analysis of cost-effectiveness data. Senn S, Scott M, Bloomfield P, Barnett V, editors. Chichester: John Wiley & Sons Ltd; 2006.

50. Jones AM. Models for Health Care (http://www.york.ac.uk/res/herc/documents/wp/10_01.pdf). Health Econometrics and Data group (HEDG) Working Paper 10/01. 2010.

51. Zethraeus N, Johannesson M, Jonsson B, Lothgren M, Tambour M. Advantages of using the net-benefit approach for analysing uncertainty in economic evaluation studies. Pharmacoeconomics. 2003;21(1):39-48. Epub 2002/12/18.

52. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Making. 1998;18(2 Suppl):S68-80.

53. Briggs A, Fenn P. Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. Health Econ. 1998;7(8):723-40.

54. Kilian R, Matschinger H, Loeffler W, Roick C, Angermeyer MC. A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. J Ment Health Policy Econ. 2002;5(1):21-31. Epub 2003/01/17.

55. Moran JL, Solomon PJ, Peisach AR, Martin J. New models for old questions: generalized linear models for cost prediction. J Eval Clin Pract. 2007;13(3):381-9. Epub 2007/05/24.

56. The IVAN Study Investigators, Chakravarthy U, Harding SP, Rogers CA, Downes SM, Lotery AJ, et al. Alternative treatments to inhibit VEGF in age-related choroidal neovascularisation: two-year findings of the IVAN randomised controlled trial. The Lancet. 2013;382(9900):1258-67.

57. The IVAN Study Investigators, Chakravarthy U, Harding SP, Rogers CA, Downes SM, Lotery AJ, et al. Ranibizumab versus Bevacizumab to Treat Neovascular Age-related Macular Degeneration: One-Year Findings from the IVAN Randomized Trial. Ophthalmology. 2012;119(7):1399-411. Epub 2012/05/15.

58. Murray DW, MacLennan GS, Breeman S, Dakin HA, Johnston L, Campbell MK, et al. A randomised controlled trial of the clinical effectiveness and cost-effectiveness of different knee prostheses: the Knee Arthroplasty Trial (KAT). Health Technology Assessment. 2014;18(19).

59. Johnston L, MacLennan G, McCormack K, Ramsay C, Walker A. The Knee Arthroplasty Trial (KAT) design features, baseline characteristics, and two-year functional outcomes after alternative approaches to knee replacement. J Bone Joint Surg Am. 2009;91(1):134-41. Epub 2009/01/06.

60. Breeman S, Campbell C, Dakin H, Fiddian N, Fitzpatrick R, Grant A, et al. Patellar resurfacing in total knee replacement: five year clinical and economic results of a large randomized controlled trial. The Journal of Bone and Joint Surgery (American). 2011;(In press).

61. Allore HG, Murphy TE. An examination of effect estimation in factorial and standardly-tailored designs. Clin Trials. 2008;5(2):121-30.

62. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials. 1998;19(1):61-109. Epub 1998/03/11.

63. Redmond C, K., Colton T. Factorial designs. Biostatistics in Clinical trials. Chichester: John Wiley & Sons Ltd; 2001.

64. Lammertse D, Tuszynski MH, Steeves JD, Curt A, Fawcett JW, Rask C, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: Clinical trial design. Spinal Cord. 2007;45(3):232-42.

65. Green S. Factorial designs with time to event endpoints. In: Crowley J, Ankerst DP, editors. Handbook of Statistics in Clinical Oncology. Second ed. Boca Raton, FL: Chapman & Hall/CRC; 2006. p. 181-9.

66. Bowling A. Research methods in health: investigating health and health services. Third ed. New York: Open University Press; 2009.

67. Portney LG, Watkins MP. Foundations of clinical research: applications to practice (Chapters 10 and 21). Upper Saddle River, NJ: Pearson Education Inc; 2009.

68. Meinhert CL. Clinical trials: Design, conduct, and analysis. New York and Oxford: Oxford University Press; 1986.

69. Jadad AR. Randomised controlled trials: a user's guide. London: BMJ Books; 1998.

70. Festing MF. Guidelines for the design and statistical analysis of experiments in papers submitted to ATLA. Altern Lab Anim. 2001;29(4):427-46.

71. Torgerson DJ, Torgerson CJ. Designing randomised trials in health, education and the social sciences: an introduction. Basingstoke, UK: Palgrave Macmillan; 2008.

72. Last JM, editor. A dictionary of epidemiology. 2nd ed. Oxford: Oxford University Press; 1988.

73. Natarajan R, Turnbull BW, Slate EH, Clark LC. A computer program for sample size and power calculations in the design of multi-arm and factorial clinical trials with survival time endpoints. Comput Methods Programs Biomed. 1996;49(2):137-47. Epub 1996/03/01.

74. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 2 x k factorial design when time-to-failure is the outcome [corrected]. Control Clin Trials. 1993;14(6):511-22. Epub 1993/12/01.

75. Stolberg HO, Norman G, Trop I. Randomized controlled trials. AJR Am J Roentgenol. 2004;183(6):1539-44. Epub 2004/11/18.

76. Earl-Slater A. The handbook of clinical trials and other research Abingdon: Radcliffe Medical Pres Ltd.; 2002. p. 137-8.

77. Ram CV. Antihypertensive efficacy of angiotensin receptor blockers in combination with hydrochlorothiazide: a review of the factorial-design studies. J Clin Hypertens (Greenwich). 2004;6(10):569-77.

78. Kochar M, Guthrie R, Triscari J, Kassler-Taub K, Reeves RA. Matrix study of irbesartan with hydrochlorothiazide in mild-to-moderate hypertension. Am J Hypertens. 1999;12(8 Pt 1):797-805. Epub 1999/09/10.

79. Campbell MK, Carr C, Devellis B, Switzer B, Biddle A, Amamoo MA, et al. A randomized trial of tailoring and motivational interviewing to promote fruit and vegetable consumption for cancer prevention and control. Annals of Behavioral Medicine. 2009;38(2):71-85.

80. Dolor RJ, Yancy WS, Jr., Owen WF, Matchar DB, Samsa GP, Pollak KI, et al. Hypertension Improvement Project (HIP): study protocol and implementation challenges. Trials. 2009;10:13. Epub 2009/02/28.

81. Byar DP, Herzberg AM, Tan WY. Incomplete factorial designs for randomized clinical trials. Stat Med. 1993;12(17):1629-41.

82. Hung HM. Global tests for combination drug studies in factorial trials. Stat Med. 1996;15(3):233-47.

83. Nair V, Strecher V, Fagerlin A, Ubel P, Resnicow K, Murphy S, et al. Screening experiments and the use of fractional factorial designs in behavioral intervention research. Am J Public Health. 2008;98(8):1354-9.

84. Hung HM. Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. Stat Med. 2000;19(16):2079-87.

85. Buring JE, Hennekens CH. Cost and efficiency in clinical trials: the U.S. Physicians' Health Study. Stat Med. 1990;9(1-2):29-33. Epub 1990/01/01.

86. Monninkhof E, van der Valk P, van der Palen J, van Herwaarden C, Zielhuis G. Effects of a comprehensive self-management programme in patients with chronic obstructive pulmonary disease. Eur Respir J. 2003;22(5):815-20.

87. Wald NJ. The epidemiological approach: An introduction to epidemiology in medicine. Fourth ed. London: Royal Society of Medicine Press Ltd; 2004.

88. Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 3rd ed. New York: Springer; 1998.

89. Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, et al. Factors that limit the quality, number and progress of randomised controlled trials. Health Technology Assessment. 1999;3(20):iii-139.

90. Simon R. A decade of progress in statistical methodology for clinical trials. Stat Med. 1991;10(12):1789-817. Epub 1991/12/01.

91. Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. Health Technol Assess. 2005;9(26):1-134, iii-iv.

92. Orme M, Dakin H, Kelly S, Loftus J, Collins S. Mixed treatment comparison and meta-regression of the efficacy and safety of prostaglandin analogues and comparators for primary open-angle glaucoma and ocular hypertension. Current Medical Research and Opinion. 2010;26(3):511-28.

93. National Collaborating Centre for Acute Care. Glaucoma: Diagnosis and management of chronic open angle glaucoma and ocular hypertension. April 2009 [28 September 2010]; Available from: http://www.nice.org.uk/nicemedia/pdf/CG85FullGuideline.pdf.

94. Xiang AH, Sather HN, Azen SP. Power considerations for testing an interaction in a 2 x k factorial design with a failure time outcome. Control Clin Trials. 1994;15(6):489-502.

95. Gray R, Collins R, Peto R, Wheatley K. Large-scale randomized evidence: trials and overviews. In: Price P, Karol Sikora K, Illidge T, editors. Treatment of Cancer. Fifth ed. London: Hodder Arnold; 2008. p. 138-51.

96. Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, et al. Recruitment to randomised trials: strategies for trial enrollment and participation study. The STEPS study. Health Technol Assess. 2007;11(48):iii, ix-105. Epub 2007/11/15.

97. Hung HMJ, Ng TH, Chi GYH, Lipicky RJ. Response surface and factorial designs for combination antihypertensive drugs. Drug Information Journal. 1990;24:371-8.

98. Hung HM. On identifying a positive dose-response surface for combination agents. Stat Med. 1992;11(5):703-11. Epub 1992/03/01.

99. Hung HM, Chi GY, Lipicky RJ. Testing for the existence of a desirable dose combination. Biometrics. 1993;49(1):85-94. Epub 1993/03/01.

100. Chow S-C, Liu J-P. Section 7.6: Combination trials. Design and analysis of clinical trials: Concepts and methodologies. Second ed. Hoboken, NJ: John Wiley & Sons Inc; 2004.

101. Hung HMJ. Testing for global superiority of a combination agent in the presence of baseline imbalance. Proceedings of the Biopharmaceuitical Section of the American Statistical Association, Alexandria, Virginia: American Statistical Association. 1990:85-90.

102. Hung HMJ, Ng TH, Chi GYH, Lipicky RJ. Testing for the existence of a dose combination beating its components. Proceedings Of the Biopharmaceuitical Section of the American Statistical Association, Alexandria, Virginia: American Statistical Association. 1989:53-9.

103. Hafner KB, Ruberg SJ. Factorial dose-response studies using frequency and magnitude of dose. J Biopharm Stat. 1996;6(3):253-62.

104. Wing LM, Chalmers JP, West MJ. Cross-over factorial studies with antihypertensive drugs. Nephron. 1987;47 Suppl 1:94-8.

105. Lindblad AS. The effects of treatment noncompliance on power in a 2 x 2 factorial design. J Biopharm Stat. 1996;6(3):263-70.

106. Ottenbacher KJ. Interpretation of interaction in factorial analysis of variance design. Stat Med. 1991;10(10):1565-71.

107. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985;41(2):361-72. Epub 1985/06/01.

108. StataCorp. Stata Base Reference Manual Volume 1 A-H. Release 11. College Station, TX: Stata Press; 2009.

109. Patel HI, Tsai KT. Analysis of paired ordered categorical data in a factorial design. J Biopharm Stat. 1996;6(1):15-36.

110. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol. 2004;57(5):454-60. Epub 2004/06/16.

111. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. 2002;21(19):2917-30. Epub 2002/09/27.

112. Stewart WH. Application of response surface methodology and factorial designs to clinical trials for drug combination development. J Biopharm Stat. 1996;6(3):219-30. Epub 1996/07/01.

113. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess. 2001;5(33):1-56. Epub 2001/11/10.

114. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. Stat Med. 1993;12(13):1239-48. Epub 1993/07/15.

115. Welton NJ, Ades AE, Caldwell DM, Peters TJ. Research prioritization based on expected value of partial perfect information: a case-study on interventions to increase uptake of breast cancer screening. J R Statist Soc A. 2008;171(Part 4):807-41.

116. Brown J, Welton NJ, Bankhead C, Richards SH, Roberts L, Tydeman C, et al. A Bayesian approach to analysing the cost-effectiveness of two primary care interventions aimed at improving attendance for breast screening. Health Econ. 2006;15(5):435-45.

117. Miller MA, Seaman JW. A Bayesian approach to assessing the superiority of a dose combination. Biometrical Journal. 1998;40:43-55.

118. Welton NJ, Madan JJ, Caldwell DM, Peters TJ, Ades AE. Expected Value of Sample Information for Multi-Arm Cluster Randomized Trials with Binary Outcomes. Med Decis Making. 2013.

119. Armitage P, Berry G, Mathews JNS. Chapter 18: Clinical trials. Statistical methods in medical research. 4th ed. Malden, MA: Blackwell Science Ltd.; 2002. p. 591-647.

120. European Medicines Agency (EMEA). ICH Topic E 9 Statistical Principles for Clinical Trials: Note for Guidance on Statistical Principles for Clinical Trial (CPMP/ICH/363/96). September 1998 [18th April 2011]; Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf.

121. Downes N. The clinical trial protocol. In: Di Giovanna I, Hayes G, editors. Principles of Clinical Research. Stroud, UK: Wrightson Biomedical Publishing Ltd; 2001. p. 181-96.

122. Pocock SJ. Clinical trials: A practical approach. Chichester: John Wiley & Sons Ltd; 1984.

123. Hackshaw A. A concise guide to clinical trials. Chichester, UK: BMJ Publishing Group Ltd.; 2009.

124. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869. Epub 2010/03/25.

125. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 3rd ed. New York: Oxford University Press; 2005 2005.

126. Dakin H, Wordsworth S. Cost-minimisation analysis versus cost-effectiveness analysis, revisited. Health Econ. 2013;22(1):22-34. Epub 2011/11/24.

127. McGuire A. Chapter 1: Theoretical concepts in the economic evaluation of health care. In: Drummond M, McGuire A, editors. Economic evaluation in health care - merging theory with practice. New York, NY: Oxford University Press; 2001. p. 1-21.

128. Johannesson M, O'Conor RM. Cost-utility analysis from a societal perspective. Health Policy. 1997;39(3):241-53. Epub 1997/02/06.

129. Weinstein MC. Chapter 5: From cost-effectiveness ratios to resource allocation: where to draw the line? In: Sloan FA, editor. Valuing health care: Costs, benefits and effectiveness of pharmaceuticals and other medical technologies. 1st paperback edition ed. Cambridge: Cambridge University Press; 1996. p. 77-97.

130. National Institute for Health and Clinical Excellence. Social value judgements: Principles for the development of NICE guidance. Second Edition. 2008 [31st May 2012]; Available from: http://www.nice.org.uk/media/C18/30/SVJ2PUBLICATION2008.pdf.

131. Tambour M, Zethraeus N, Johannesson M. A note on confidence intervals in cost-effectiveness analysis. International journal of technology assessment in health care. 1998;14(3):467-71. Epub 1998/10/22.

132. Phelps CE, Mushlin AI. On the (near) equivalence of cost-effectiveness and cost-benefit analyses. International journal of technology assessment in health care. 1991;7(1):12-21.

133. Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. Health Econ. 1996;5(6):513-24. Epub 1996/11/01.

134. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? Health Econ. 2006;15(7):677-87.

135. Gray AM. Cost-effectiveness analyses alongside randomised clinical trials. Clin Trials. 2006;3(6):538-42. Epub 2006/12/16.

136. Gray A, Clarke P, Wolstenholme J, Wordsworth S. Chapter 11: Presenting cost-effectiveness results. Applied Methods of Cost-Effectiveness Analysis in Health Care Oxford: Oxford University Press; 2011.

137. O'Brien B. Economic evaluation of pharmaceuticals. Frankenstein's monster or vampire of trials? Med Care. 1996;34(12 Suppl):DS99-108. Epub 1996/12/01.

138. Marshall DA, Hux M. Design and analysis issues for economic analysis alongside clinical trials. Med Care. 2009;47(7 Suppl 1):S14-20. Epub 2009/06/19.

139. Briggs A. Economic evaluation and clinical trials: size matters. Bmj. 2000;321(7273):1362-3.

140. Backhouse ME. Use of randomised controlled trials for producing cost-effectiveness evidence: potential impact of design choices on sample size and study duration. Pharmacoeconomics. 2002;20(15):1061-77. Epub 2002/11/29.

141. Hollingworth W, McKell-Redwood D, Hampson L, Metcalfe C. Cost utility analysis conducted alongside RCTs: Is inefficiency considered in sample size calculations? Poster presented at the MRC Clinical Trials Methodology Conference, Bristol, UK, 4th-5th October 2011. 2011.

142. Glick HA, Polsky DP, Schulman KA. Chapter 6: Trial-based economic evaluations: an overview of design and analysis. In: Drummond M, McGuire A, editors. Economic evaluation in health care - merging theory with practice. New York, NY: Oxford University Press; 2001.

143. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, et al. Modelling in economic evaluation: an unavoidable fact of life. Health Econ. 1997;6(3):217-27.

144. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis. 1967;20(8):637-48. Epub 1967/08/01.

145. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. 2013 [17th May 2013]; Available from: http://www.nice.org.uk/media/D45/1E/GuideToMethodsTechnologyAppraisal2013.pdf.

146. Anderson P, Lloyd A. Outcomes research and health economics. In: Di Giovanna I, Hayes G, editors. Principles of Clinical Research. Stroud, UK: Wrightson Biomedical Publishing Ltd; 2001. p. 291-316.

147. Lloyd-Parks J. Clinical trial design. In: Di Giovanna I, Hayes G, editors. Principles of Clinical Research. Stroud, UK: Wrightson Biomedical Publishing Ltd; 2001. p. 93-116.

148. Whitehurst DG, Bryan S. Trial-based clinical and economic analyses: the unhelpful quest for conformity. Trials. 2013;14:421. Epub 2013/12/07.

149. Mihaylova B, Briggs A, Armitage J, Parish S, Gray A, Collins R. Lifetime cost effectiveness of simvastatin in a range of risk groups and age groups derived from a randomised trial of 20,536 people. BMJ. 2006;333(7579):1145. Epub 2006/11/14.

150. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ. 2005;14(5):487-96. Epub 2004/10/22.

151. Noble SM, Hollingworth W, Tilling K. Missing data in trial-based cost-effectiveness analysis: the current state of play. Health Econ. 2012;21(2):187-200.

152. Bland JM. The tyranny of power: is there a better way to calculate sample size? BMJ. 2009;339:b3985. Epub 2009/10/08.

153. Kay R. Statistical thinking for clinical trials. In: Di Giovanna I, Hayes G, editors. Principles of Clinical Research. Stroud, UK: Wrightson Biomedical Publishing Ltd; 2001. p. 339-54.

154. Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? BMJ. 2000;320(7243):1197-200. Epub 2000/04/28.

155. Gray A, Clarke P, Wolstenholme J, Wordsworth S. Chapter 7: Analysing costs. Applied Methods of Cost-Effectiveness Analysis in Health Care Oxford: Oxford University Press; 2011.

156. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. J Health Serv Res Policy. 1998;3(4):233-45. Epub 1998/09/04.

157. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. Journal of health economics. 1999;18(3):341-64.

158. Briggs AH, O'Brien BJ. The death of cost-minimization analysis? Health Econ. 2001;10(2):179-84. Epub 2001/03/17.

159. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 ed: The Cochrane Collaboration; March 2011.

160. O'Brien BJ. A tale of two (or more) cities: geographic transferability of pharmacoeconomic data. The American journal of managed care. 1997;3 Suppl:S33-9.

161. Eckermann S, Coory M, Willan AR. Indirect comparison: relative risk fallacies and odds solution. J Clin Epidemiol. 2009;62(10):1031-6. Epub 2009/01/31.

162. Barber J, Thompson S. Multiple regression of cost data: use of generalised linear models. J Health Serv Res Policy. 2004;9(4):197-204. Epub 2004/10/29.

163. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Chapter 8: Methods for obtaining health state values: generic preference-based measures of health and alternatives. In: Brazier J, editor. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.

164. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002;40(2):113-28. Epub 2002/01/22.

165. Hu B, Fu AZ. Predicting Utility for Joint Health States: A General Framework and a New Nonparametric Estimator. Med Decis Making. 2010;30(5):E29–E39.

166. Dale W. What is the best model for estimating joint health states utilities? Comparing the linear index model to the proportional decrement model. Med Decis Making. 2010;30(5):531-3. Epub 2010/10/21.

167. Fu AZ, Luo N. Predicting utility for joint health conditions: A comparison of different estimators using EQ-5D index in the United States. Presented at the 28th plenary meeting of the EuroQol group, Oxford UK, September 2011. 2011.

168. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-108.

169. Karlsson G, Johannesson M. The decision rules of cost-effectiveness analysis. Pharmacoeconomics. 1996;9(2):113-20. Epub 1996/01/08.

170. Johannesson M. Theory and Methods of Economic Evaluation of Health Care. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1996.

171. Weinstein MC. Decision rules for incremental cost-effectiveness analysis. In: Jones A, editor. The Elgar companion to health economics. Cheltenham, UK: Edward Elgar; 2006. p. 469-78.

172. Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, et al. Decision making in health and medicine: integrating evidence and values. Cambridge, UK: Cambridge University Press; 2001.

173. Johannesson M, Weinstein MC. On the decision rules of cost-effectiveness analysis. Journal of health economics. 1993;12(4):459-67. Epub 1993/11/04.

174. Weinstein MC. Principles of cost-effective resource allocation in health care organizations. International journal of technology assessment in health care. 1990;6(1):93-103. Epub 1990/01/01.

175. Johannesson M. On the estimation of cost-effectiveness ratios. Health Policy. 1995;31(3):225-9. Epub 1995/02/07.

176. Johannesson M. The relationship between cost-effectiveness analysis and cost-benefit analysis. Soc Sci Med. 1995;41(4):483-9. Epub 1995/08/01.

177. Johannesson M. A note on the depreciation of the societal perspective in economic evaluation of health care. Health Policy. 1995;33(1):59-66. Epub 1995/06/07.

178. Stinnett AA, Paltiel AD. Mathematical programming for the efficient allocation of health care resources. Journal of health economics. 1996;15(5):641-53. Epub 1996/09/04.

179. Earnshaw SR, Dennett SL. Integer/linear mathematical programming models: a tool for allocating healthcare resources. Pharmacoeconomics. 2003;21(12):839-51.

180. Sendi P, Al MJ. Revisiting the decision rule of cost-effectiveness analysis under certainty and uncertainty. Soc Sci Med. 2003;57(6):969-74.

181. Birch S, Gafni A. Cost effectiveness/utility analyses. Do current decision rules lead us to where we want to be? Journal of health economics. 1992;11(3):279-96.

182. Weinstein MC, Zeckhauser RJ. Critical ratios and efficient allocation. Journal of Public Economics. 1973;2(2):147-57.

183. Murray CJ, Evans DB, Acharya A, Baltussen RM. Development of WHO guidelines on generalized cost-effectiveness analysis. Health Econ. 2000;9(3):235-51. Epub 2000/05/03.

184. Barton GR, Briggs AH, Fenwick EA. Optimal cost-effectiveness decisions: the role of the cost-effectiveness acceptability curve (CEAC), the cost-effectiveness acceptability frontier (CEAF), and the expected value of perfection information (EVPI). Value Health. 2008;11(5):886-97. Epub 2008/05/21.

185. Tan-Torres Edejer T, Baltussen R, Adam T, Hutubessy R, Acharya A, Evans DB, et al. Making choices in health: WHO guide to cost-effectiveness analysis. Geneva, Switzerland: World Health Organization; 2003.

186. Devlin N, Parkin D. Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. Health Econ. 2004;13(5):437-52. Epub 2004/05/06.

187. Heshmat S. Chapter 9: Connecting value and cost: cost-effectiveness analysis. An overview of managerial economics in the health care system. Albany, NY: Delmar; 2001.

188. Briggs AH. Handling uncertainty in cost-effectiveness models. Pharmacoeconomics. 2000;17(5):479-500.

189. Phillips C. What is cost-effectiveness? (http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/Cost-effect.pdf). February 2009.

190. Evans DB, Edejer TT, Adam T, Lim SS. Methods to assess the costs and health effects of interventions for improving health in developing countries. Bmj. 2005;331(7525):1137-40. Epub 2005/11/12.

191. Manning W. Dealing with skewed data on costs and expenditures. In: Jones A, editor. The Elgar companion to health economics. Cheltenham, UK: Edward Elgar; 2006. p. 439-46.

192. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. The New England journal of medicine. 1977;296(13):716-21.

193. Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. Health Econ. 2005;14(12):1217-29. Epub 2005/06/10.

194. Gomes M, Grieve R, Nixon R, Ng ES, Carpenter J, Thompson SG. Methods for covariate adjustment in cost-effectiveness analysis that use cluster randomised trials. Health Econ. 2012;21(9):1101-18.

195. Koopmanschap MA, Touw KC, Rutten FF. Analysis of costs and cost-effectiveness in multinational trials. Health Policy. 2001;58(2):175-86. Epub 2001/09/12.

196. Nixon RM, Wonderling D, Grieve RD. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. Health Econ. 2010;19(3):316-33. Epub 2009/04/21.

197. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Econ. 2004;13(5):461-75. Epub 2004/05/06.

198. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003;89(2):232-8.

199. Hernandez Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. Value Health. 2012;15(3):550-61.

200. Basu A, Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. Med Decis Making. 2012;32(1):56-69. Epub 2011/10/20.

201. van Hout BA, Al MJ, Gordon GS, Rutten FF. Costs, effects and C/E-ratios alongside a clinical trial. Health Econ. 1994;3(5):309-19.

202. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. Health Econ. 2001;10(8):779-87.

203. Willan AR, Pinto EM. The value of information and optimal clinical trial design. Stat Med. 2005;24(12):1791-806. Epub 2005/04/05.

204. Eckermann S, Willan AR. Expected value of information and decision making in HTA. Health Econ. 2007;16(2):195-209.

205. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. Health Econ. 1997;6(4):327-40. Epub 1997/07/01.

206. Briggs A. Statistical methods for cost-effectiveness analysis alongside clinical trials. In: Jones A, editor. The Elgar companion to health economics. Cheltenham, UK: Edward Elgar; 2006. p. 503-13.

207. StataCorp. sureg — Zellner's seemingly unrelated regression. Stata Base Reference Manual: Release 11. College Station, TX: StataCorp LP; 2009. p. 2100-7.

208. StataCorp. Suest - seemingly unrelated estimation. Stata Base Reference Manual: Release 11. College Station, TX: StataCorp LP; 2009. p. 1800-18.

209. Sutton AJ, Cooper NJ, Abrams KR, Lambert PC, Jones DR. A Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. J Clin Epidemiol. 2005;58(1):26-40. Epub 2005/01/15.

210. Clarke PM, Gray AM, Briggs A, Farmer AJ, Fenn P, Stevens RJ, et al. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). Diabetologia. 2004;47(10):1747-59.

211. Gray AM, Clarke P. The economic analyses of the UK prospective diabetes study. Diabet Med. 2008;25 Suppl 2:47-51.

212. O'Hagan A, Luce BR. A primer on Bayesian Statistics in Health Economics and Outcomes Research: Bayesian Initiative in Health Economics & Outcomes Research and Centre for Bayesian Statistics in Health Economics; 2003.

213. Sadatsafavi M, Marra C, McCandless L, Bryan S. The challenge of incorporating external evidence in trial-based cost-effectiveness analyses: the use of resampling method. HEDG Health, Econometrics and Data Group, University of York, working paper. 2012;WP 12/24.

214. Rubin DB. The Bayesian bootstrap. The Annals of Statistics. 1981;9(1):130-4.

215. White IR, Moodie E, Thompson SG, Croudace T. A modelling strategy for the analysis of clinical trials with partly missing longitudinal data. International journal of methods in psychiatric research. 2003;12(3):139-50.

216. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 2000;10:325-37.

217. Claxton K, Palmer S, Longworth L, Bojke L, Griffin S, McKenna C, et al. Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. Health Technol Assess. 2012;16(46):1-323.

218. Espinoza MA, Manca A, Claxton K, Sculpher MJ. The value of identifying heterogeneity: a framework for subgroup cost effectiveness analysis. Paper presented at the HESG Conference Winter 2011 (York). 2011.

219. National Institute for Health and Clinical Excellence. Tenofovir disoproxil for the treatment of chronic hepatitis B. 2009 [23 October 2013]; Available from: http://www.nice.org.uk/nicemedia/live/12184/44890/44890.pdf.

220. National Institute for Health and Clinical Excellence. Entecavir for the treatment of chronic hepatitis B. 2008 [23 October 2013]; Available from: http://www.nice.org.uk/nicemedia/live/12046/41667/41667.pdf.

221. Dakin H, Bentley A, Dusheiko G. Cost-utility analysis of tenofovir disoproxil fumarate in the treatment of chronic hepatitis B. Value Health. 2010;13(8):922-33.

222. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14(6):1523-32.

223. Roset M, Badia X, Mayo NE. Sample size calculations in studies using the EuroQol 5D. Qual Life Res. 1999;8(6):539-49.

224. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med. 2005;60(7):1571-82.

225. Gudex C, editor. Time trade-off user manual: props and self-completion methods. York: University of York, Centre for Health Economics (http://www.york.ac.uk/inst/che/pdf/op20.pdf); April 1994.

226. Williams A. The measurement and valuation of health: A chronicle. The University of York Discussion paper 136. 1995.

227. Siegel JE, Weinstein MC, Torrance GW. Chapter 9: Reporting cost-effectiveness studies and results. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, editors. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 276-303.

228. Burnham KP, Anderson DR. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods and Research. 2004;33(2):261-304.

229. Lancelot R, Lesnoff M, McDermott JJ. Use of Akaike information criteria for model selection and inference: An application to assess prevention of gastrointestinal parasitism and respiratory mortality of Guinean goats in Kolda, Senegal. Preventive Veterinary Medicine. 2002;55(4):217-40.

230. Coyle D, Buxton MJ, O'Brien BJ. Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. Health Econ. 2003;12(5):421-7. Epub 2003/04/30.

231. Sever PS, Dahlof B, Poulter NR, Wedel H, Beevers G, Caulfield M, et al. Rationale, design, methods and baseline demography of participants of the Anglo-Scandinavian Cardiac Outcomes Trial. ASCOT investigators. J Hypertens. 2001;19(6):1139-47. Epub 2001/06/14.

232. Sullivan MD, Anderson RT, Aron D, Atkinson HH, Bastien A, Chen GJ, et al. Health-Related Quality of Life and Cost-Effectiveness Components of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial: Rationale and Design. The American Journal of Cardiology. 2007;99(12, Supplement 1):S90-S102.

233. Mihaylov S, Stark C, McColl E, Steen N, Vanoli A, Rubin G, et al. Stepped treatment of older adults on laxatives. The STOOL trial. Health Technol Assess. 2008;12(13):iii-iv, ix-139. Epub 2008/05/09.

234. Thomas KS, Muir KR, Doherty M, Jones AC, O'Reilly SC, Bassey EJ. Home based exercise programme for knee pain and knee osteoarthritis: randomised controlled trial. Bmj. 2002;325(7367):752.

235. Boyle J, McCartney E, Forbes J, O'Hare A. A randomised controlled trial and economic evaluation of direct versus indirect and individual versus group modes of speech and language therapy for children with primary language impairment. Health Technol Assess. 2007;11(25):iii-iv, xi-xii, 1-139. Epub 2007/07/06.

236. Pinto DS, Stone GW, Shi C, Dunn ES, Reynolds MR, York M, et al. Economic Evaluation of Bivalirudin With or Without Glycoprotein IIb/IIIa Inhibition Versus Heparin With Routine Glycoprotein IIb/IIIa Inhibition for Early Invasive Management of Acute Coronary Syndromes. Journal of the American College of Cardiology. 2008;52(22):1758-68.

237. Thompson RS, Michnich ME, Gray J, Friedlander L, Gilson B. Maximizing compliance with hemoccult screening for colon cancer in clinical practice. Med Care. 1986;24(10):904-14. Epub 1986/10/01.

238. Campbell AJ, Robertson MC, La Grow SJ, Kerse NM, Sanderson GF, Jacobs RJ, et al. Randomised controlled trial of prevention of falls in people aged >=75 with severe visual impairment: the VIP trial. BMJ. 2005;331(7520):817.

239. Sevick MA, Miller GD, Loeser RF, Williamson JD, Messier SP. Cost-effectiveness of exercise and diet in overweight and obese adults with knee osteoarthritis. Med Sci Sports Exerc. 2009;41(6):1167-74. Epub 2009/05/23.

240. Salize HJP, Merkel SD-R, Reinhard ID-M, Twardella DP, Mann KMD, Brenner HMDMPH. Cost-effective Primary Care-Based Strategies to Improve Smoking Cessation: More Value for Money. Archives of Internal Medicine. 2009;169(3):230-5.

241. Lamb SE, Gates S, Underwood MR, Cooke MW, Ashby D, Szczepura A, et al. Managing Injuries of the Neck Trial (MINT): design of a randomised controlled trial of treatments for whiplash associated disorders. BMC Musculoskelet Disord. 2007;8:7. Epub 2007/01/30.

242. Foster EM, Jensen PS, Schlander M, Pelham WE, Jr., Hechtman L, Arnold LE, et al. Treatment for ADHD: is more complex treatment cost-effective for more complex cases? Health Serv Res. 2007;42(1 Pt 1):165-82. Epub 2007/03/16.

243. Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico. Lancet. 1999;354(9177):447-55. Epub 1999/08/28.

244. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: cost effectiveness of physical treatments for back pain in primary care. BMJ. 2004;329(7479):1381. Epub 2004/11/24.

245. Reddy P, Kalus JS, Caron MF, Horowitz S, Karapanos A, Coleman CI, et al. Economic analysis of intravenous plus oral amiodarone, atrial septal pacing, and both strategies to prevent atrial fibrillation after open heart surgery. Pharmacotherapy. 2004;24(8):1013-9. Epub 2004/09/02.

246. Hollinghurst S, Sharp D, Ballard K, Barnett J, Beattie A, Evans M, et al. Randomised controlled trial of Alexander technique lessons, exercise, and massage (ATEAM) for chronic and recurrent back pain: economic evaluation. BMJ. 2008;337:a2656. Epub 2008/12/17.

247. Hollis JF, McAfee TA, Fellows JL, Zbikowski SM, Stark M, Riedlinger K. The effectiveness and cost effectiveness of telephone counselling and the nicotine patch in a state tobacco quitline. Tobacco Control. 2007;16 Supplement(1):i53-i9.

248. Sullivan FM, Swan IR, Donnan PT, Morrison JM, Smith BH, McKinstry B, et al. A randomised controlled trial of the use of aciclovir and/or prednisolone for the early treatment of Bell's palsy: the BELLS study. Health Technol Assess. 2009;13(47):iii-iv, ix-xi 1-130. Epub 2009/10/17.

249. Lindgren P, Buxton M, Kahan T, Poulter NR, Dahlof B, Sever PS, et al. The lifetime cost effectiveness of amlodipine-based therapy plus atorvastatin compared with atenolol plus atorvastatin, amlodipine-based therapy alone and atenolol-based therapy alone: results from ASCOT1. Pharmacoeconomics. 2009;27(3):221-30. Epub 2009/04/10.

250. Cantor JC, Morisky DE, Green LW, Levine DM, Salkever DS. Cost-effectiveness of educational interventions to improve patient outcomes in blood pressure control. Prev Med. 1985;14(6):782-800. Epub 1985/11/01.

251. Brandon TH, Meade CD, Herzog TA, Chirikos TN, Webb MS, Cantor AB. Efficacy and cost-effectiveness of a minimal intervention to prevent smoking relapse: dismantling the effects of amount of content versus contact. J Consult Clin Psychol. 2004;72(5):797-808. Epub 2004/10/16.

252. Barnett PG, Masson CL, Sorensen JL, Wong W, Hall S. Linking opioid-dependent hospital patients to drug treatment: health care use and costs 6 months after randomization. Addiction. 2006;101(12):1797-804.

253. Saillour-Glenisson F, Michel P, Daucourt V. [Medico-economic assessment of two methods for implementing thyroid testing guidelines]. Rev Epidemiol Sante Publique. 2005;53 Spec No 1:1S79-88. Epub 2005/12/06. Evaluation medico-economique de la mise en oeuvre de recommandations de prescription des examens biologiques explorant la fonction thyroidienne.

254. Waterhouse JC, Walters SJ, Oluboyede Y, Lawson RA. A randomised 2 x 2 trial of community versus hospital pulmonary rehabilitation, followed by telephone or conventional follow-up. Health Technol Assess. 2010;14(6):i-v, vii-xi, 1-140. Epub 2010/02/12.

255. Dakin HA, Wordsworth S, Gray A, Rogers C, Abangma G, Reeves B. Why consider interactions in trial-based economic evaluation? A case study of a factorial trial. Presented at the Health Economists' Study Group meeting 25th-27th June 2012, Oxford. 2012.

256. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99. Epub 2010/12/02.

257. British Medical Association. British National Formulary 62 (http://bnf.org/bnf/index.htm). London: Pharmaceutical Press; September 2011.

258. Briggs A, Clark T, Wolstenholme J, Clarke P. Missing... presumed at random: cost-analysis of incomplete data. Health Econ. 2003;12(5):377-92.

259. Ng E, Diaz-Ordaz K, Grieve R, Nixon R, Thompson S, Carpenter J. Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: An approach to model choice. Stat Methods Med Res (in press). 2013.

260. Chakravarthy U, Harding S, Rogers C, Downes S, Lotery A, Dakin H, et al. A randomised controlled trial to assess the effectiveness and cost-effectiveness of alternative treatments to Inhibit VEGF in Age-related choroidal Neovascularisation (IVAN). Health Technol Assess. 2015 (in press).

261. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009;12(4):409-18. Epub 2009/11/11.

262. KAT: Knee Arthroplasty Trial Protocol. Version 6 - March 2009. 2009; Available from: http://www.hta.ac.uk/protocols/199500100001.pdf.

263. UK Prospective Diabetes Study (UKPDS). VIII. Study design, progress and performance. Diabetologia. 1991;34(12):877-90. Epub 1991/12/01.

264. Mehta SR, Yusuf S, Diaz R, Zhu J, Pais P, Xavier D, et al. Effect of glucose-insulin-potassium infusion on mortality in patients with acute ST-segment elevation myocardial infarction: the CREATE-ECLA randomized controlled trial. JAMA. 2005;293(4):437-46. Epub 2005/01/27.

265. Yusuf S, Mehta SR, Chrolavicius S, Afzal R, Pogue J, Granger CB, et al. Effects of fondaparinux on mortality and reinfarction in patients with acute ST-segment elevation myocardial infarction: the OASIS-6 randomized trial. JAMA. 2006;295(13):1519-30. Epub 2006/03/16.

266. Montori VM, Guyatt GH. Intention-to-treat principle. CMAJ. 2001;165(10):1339-41. Epub 2002/01/05.

267. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117. Epub 2010/04/01.

268. Cui L, Hung HM, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. J Biopharm Stat. 2002;12(3):347-58. Epub 2002/11/27.

269. McKenna C, Bojke L, Manca A, Adebajo A, Dickson J, Helliwell P, et al. Shoulder acute pain in primary health care: is retraining GPs effective? The SAPPHIRE randomized trial: a cost-effectiveness analysis. Rheumatology (Oxford). 2009;48(5):558-63. Epub 2009/03/05.

270. Janzon M, Levin LA, Swahn E. Cost-effectiveness of an invasive strategy in unstable coronary artery disease; results from the FRISC II invasive trial. The Fast Revascularisation during InStability in Coronary artery disease. Eur Heart J. 2002;23(1):31-40. Epub 2001/12/14.

271. Lindgren P, Buxton M, Kahan T, Poulter NR, Dahlof B, Sever PS, et al. Cost-effectiveness of atorvastatin for the prevention of coronary and stroke events: an economic analysis of the Anglo-Scandinavian Cardiac Outcomes Trial--lipid-lowering arm (ASCOT-LLA). Eur J Cardiovasc Prev Rehabil. 2005;12(1):29-36. Epub 2005/02/11.

272. Lindgren P, Buxton M, Kahan T, Poulter NR, Dahlof B, Sever PS, et al. Economic evaluation of ASCOT-BPLA: antihypertensive treatment with an amlodipine-based regimen is cost effective compared with an atenolol-based regimen. Heart. 2008;94(2):e4. Epub 2007/10/06.

273. Dahlof B, Sever PS, Poulter NR, Wedel H, Beevers DG, Caulfield M, et al. Prevention of cardiovascular events with an antihypertensive regimen of amlodipine adding perindopril as required versus atenolol adding bendroflumethiazide as required, in the Anglo-Scandinavian Cardiac Outcomes Trial-Blood Pressure Lowering Arm (ASCOT-BPLA): a multicentre randomised controlled trial. Lancet. 2005;366(9489):895-906. Epub 2005/09/13.

274. Sever PS, Dahlof B, Poulter NR, Wedel H, Beevers G, Caulfield M, et al. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial--Lipid Lowering Arm (ASCOT-LLA): a multicentre randomised controlled trial. Lancet. 2003;361(9364):1149-58. Epub 2003/04/11.

275. Sever P, Dahlof B, Poulter N, Wedel H, Beevers G, Caulfield M, et al. Potential synergy between lipid-lowering and blood-pressure-lowering in the Anglo-Scandinavian Cardiac Outcomes Trial. Eur Heart J. 2006;27(24):2982-8. Epub 2006/12/06.

276. Janzon M, Levin LA, Swahn E. Cost effectiveness of extended treatment with low molecular weight heparin (dalteparin) in unstable coronary artery disease: results from the FRISC II trial. Heart. 2003;89(3):287-92. Epub 2003/02/20.

277. Wallentin L, Lagerqvist B, Husted S, Kontny F, Stahle E, Swahn E. Outcome at 1 year after an invasive compared with a non-invasive strategy in unstable coronary-artery disease: the FRISC II invasive randomised trial. FRISC II Investigators. Fast Revascularisation during Instability in Coronary artery disease. Lancet. 2000;356(9223):9-16. Epub 2000/07/13.

278. Moller BH, Janzon M, Levin LA, Swahn E. Comment concerning 'cost-effectiveness of an invasive strategy in unstable coronary artery disease' [2] (multiple letters). European Heart Journal. 2002;23(20):1634-5.

279. Invasive compared with non-invasive treatment in unstable coronary-artery disease: FRISC II prospective randomised multicentre study. FRagmin and Fast Revascularisation during InStability in Coronary artery disease Investigators. Lancet. 1999;354(9180):708-15. Epub 1999/09/04.

280. Long-term low-molecular-mass heparin in unstable coronary-artery disease: FRISC II prospective randomised multicentre study. FRagmin and Fast Revascularisation during InStability in Coronary artery disease. Investigators. Lancet. 1999;354(9180):701-7. Epub 1999/09/04.

281. Watson J, Helliwell P, Morton V, Adebajo A, Dickson J, Russell I, et al. Shoulder acute pain in primary healthcare: is retraining effective for GP principals? SAPPHIRE--a randomized controlled trial. Rheumatology (Oxford). 2008;47(12):1795-802. Epub 2008/09/30.

282. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. UK Prospective Diabetes Study Group. Bmj. 1998;317(7160):703-13.

283. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). UK Prospective Diabetes Study (UKPDS) Group. Lancet. 1998;352(9131):854-65.

284. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. Lancet. 1998;352(9131):837-53.

285. Cost effectiveness analysis of improved blood pressure control in hypertensive patients with type 2 diabetes: UKPDS 40. UK Prospective Diabetes Study Group. Bmj. 1998;317(7160):720-6.

286. Gray A, Raikou M, McGuire A, Fenn P, Stevens R, Cull C, et al. Cost effectiveness of an intensive blood glucose control policy in patients with type 2 diabetes: economic analysis alongside randomised controlled trial (UKPDS 41). United Kingdom Prospective Diabetes Study Group. Bmj. 2000;320(7246):1373-8.

287. Clarke P, Gray A, Adler A, Stevens R, Raikou M, Cull C, et al. Cost-effectiveness analysis of intensive blood-glucose control with metformin in overweight patients with type II diabetes (UKPDS No. 51). Diabetologia. 2001;44(3):298-304.

288. Gray A, Clarke P, Raikou M, Adler A, Stevens R, Neil A, et al. An economic evaluation of atenolol vs. captopril in patients with Type 2 diabetes (UKPDS 54). Diabetic Medicine. 2001;18(6):438-44.

289. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. Journal of Statistical Software. 2011;42(7):1-52.

290. Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical science : a review journal of the Institute of Mathematical Statistics. 2010;25(1):1-21.

291. Farmer AJ, Paul S, Tucker L, Griffin SJ, Hardeman W, Hammond T, et al. Do action plans improve adherence to medication in a 16 week clinical trial? Diabetic Medicine. 2008;25(Suppl 1):115.

292. Diabetes Trial Unit. UKPDS Outcomes Model User Manual: Version 1.3. ISIS Innovation Ltd; 2011 [10th March 2014]; Available from: https://www.dtu.ox.ac.uk/outcomesmodel/UKPDSOutcomesManual.pdf.

293. Diabetes Trial Unit. Atorvastatin in Factorial with Omega-3 Fatty Acids Risk Reduction in Diabetes. [8th October 2013]; Available from: http://www.dtu.ox.ac.uk/aforrd/protocol.php [summary].

294. Gomes M, Ng ES, Grieve R, Nixon R, Carpenter J, Thompson SG. Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. Med Decis Making. 2012;32(2):350-61.

295. Stevens RJ, Kothari V, Adler AI, Stratton IM, United Kingdom Prospective Diabetes Study G. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). Clinical science. 2001;101(6):671-9.

296. UKPDS Risk Engine version 2.0. [6th October 2008]; Available from: http://www.dtu.ox.ac.uk/riskengine/index.php.

297. Taylor F, Huffman MD, Macedo AF, Moore TH, Burke M, Davey Smith G, et al. Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2013;1:CD004816. Epub 2013/02/27.

298. Mihaylova B, Emberson J, Blackwell L, Keech A, Simes J, Barnes EH, et al. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. Lancet. 2012;380(9841):581-90. Epub 2012/05/23.

299. Elbasha EH, Messonnier ML. Cost-effectiveness analysis and health care resource allocation: decision rules under variable returns to scale. Health Econ. 2004;13(1):21-35. Epub 2004/01/16.

300. Lauer JA, Rohrich K, Wirth H, Charette C, Gribble S, Murray CJ. PopMod: a longitudinal population model with two interacting disease states. Cost effectiveness and resource allocation : C/E. 2003;1(1):6. Epub 2003/05/30.

301. Luce BR, Manning W, Siegel JE, Lipscomb J. Estimating costs in cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, editors. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press; 1996. p. 176-213.

302. Murray CJ, Lauer JA, Hutubessy RC, Niessen L, Tomijima N, Rodgers A, et al. Effectiveness and costs of interventions to lower systolic blood pressure and cholesterol: a

global and regional analysis on reduction of cardiovascular-disease risk. Lancet. 2003;361(9359):717-25. Epub 2003/03/07.

303. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. Bmj. 2005;331(7521):897-900.

304. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Making. 2013;33(5):607-17.

305. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 1: introduction. Med Decis Making. 2013;33(5):597-606.

306. Bafeta A, Trinquart L, Seror R, Ravaud P. Reporting of results from network meta-analyses: methodological systematic review. BMJ. 2014;348:g1741.

307. Caldwell DM, Gibb DM, Ades AE. Validity of indirect comparisons in meta-analysis. Lancet. 2007;369(9558):270; author reply 1. Epub 2007/01/30.

308. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Med Decis Making. 2013;33(5):641-56.

309. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. Biostatistics. 2009;10(4):792-805.

310. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics. 2008;26(9):753-67.

311. Kahneman D. Thinking, fast and slow. London: Penguin Books Ltd; 2012.

312. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health. 2013;16(2):231-50. Epub 2013/03/30.

313. Latimer NR. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Making. 2013;33(6):743-54. Epub 2013/01/24.

314. Sun X, Briel M, Busse JW, Akl EA, You JJ, Mejza F, et al. Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials. Trials. 2009;10:101. Epub 2009/11/11.

315. Peto R. Current misconception 3: that subgroup-specific trial mortality results often provide a good basis for individualising patient care. Br J Cancer. 2011;104(7):1057-8. Epub 2011/03/31.

316. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. Med J Aust. 2004;180(6):289-91. Epub 2004/03/12.

317. Stevens W, Normand C. Optimisation versus certainty: understanding the issue of heterogeneity in economic evaluation. Soc Sci Med. 2004;58(2):315-20. Epub 2003/11/08.

318. Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. Pharmacoeconomics. 2008;26(9):799-806. Epub 2008/09/05.

319. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and

Research (CBER). Guidance for Industry Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products. May 1998 [12th August 2010]; Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm078749.pdf.

320. Akritas MG, Brunner E. Nonparametric methods for factorial designs with censored data. Journal of the American Statistical Association. 1997;92(438):568-76.

321. Fletcher DJ, Lewis SM, Matthews JN. Factorial designs for crossover clinical trials. Stat Med. 1990;9(10):1121-9.

322. Holms AG, Berrettoni JN. Chain-pooling ANOVA for two-level factorial replication-free experiments. Technometrics. 1969;11(4):725-46.

323. Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JA. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. Int J Epidemiol. 2003;32(5):840-6.

324. Pool JL, Cushman WC, Saini RK, Nwachuku CE, Battikha JP. Use of the factorial design and quadratic response surface models to evaluate the fosinopril and hydrochlorothiazide combination therapy in hypertension. Am J Hypertens. 1997;10(1):117-23.

325. Janosky JE. Interpretation of interaction in factorial analysis of variance design. Stat Med. 1992;11(10):1403.

326. Coronis Collaborative Group, Abalos E, Addo V, Brocklehurst P, El Sheikh M, Farrell B, et al. Caesarean section surgical techniques (CORONIS): a fractional, factorial, unmasked, randomised controlled trial. Lancet. 2013;382(9888):234-48.

327. Brocklehurst P, Abalos E, Addo V, Sharma JB, Matthews J, Oyieke J, et al. The CORONIS Trial. International study of caesarean section surgical techniques: A randomised fractional, factorial trial. BMC Pregnancy and Childbirth. 2007;7(24).

328. Apfel CC, Korttila K, Abdalla M, Biedler A, Kranke P, Pocock SJ, et al. An international multicenter protocol to assess the single and combined benefits of antiemetic interventions in a controlled clinical trial of a 2x2x2x2x2x2 factorial design (IMPACT). Controlled Clinical Trials. 2003;24(6):736-51.

329. Ram CV. Antihypertensive efficacy of olmesartan medoxomil or valsartan in combination with amlodipine: a review of factorial-design studies. Curr Med Res Opin. 2009;25(1):177-85.

330. Goldberg MR, Offen WW. Pinacidil with and without hydrochlorothiazide. Dose-response relationships from results of a 4 x 3 factorial design study. Drugs. 1988;36 Suppl 7:83-92.

331. Pordy RC. Cilazapril plus hydrochlorothiazide: improved efficacy without reduced safety in mild to moderate hypertension. A double-blind placebo-controlled multicenter study of factorial design. Cardiology. 1994;85(5):311-22. Epub 1994/01/01.

332. Eccles M, Grimshaw J, Steen N, Parkin D, Purves I, McColl E, et al. The design and analysis of a randomized controlled trial to evaluate computerized decision support in primary care: the COGENT study. Fam Pract. 2000;17(2):180-6. Epub 2000/04/12.

333. Eccles M, McColl E, Steen N, Rousseau N, Grimshaw J, Parkin D, et al. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. BMJ. 2002;325(7370):941. Epub 2002/10/26.

334. Stone GW, Bertrand M, Colombo A, Dangas G, Farkouh ME, Feit F, et al. Acute Catheterization and Urgent Intervention Triage strategY (ACUITY) trial: study design and rationale. Am Heart J. 2004;148(5):764-75. Epub 2004/11/04.

335. Stone GW, White HD, Ohman EM, Bertrand ME, Lincoff AM, McLaurin BT, et al. Bivalirudin in patients with acute coronary syndromes undergoing percutaneous coronary intervention: a subgroup analysis from the Acute Catheterization and Urgent Intervention Triage strategy (ACUITY) trial. Lancet. 2007;369(9565):907-19. Epub 2007/03/21.

336. Stone GW, McLaurin BT, Cox DA, Bertrand ME, Lincoff AM, Moses JW, et al. Bivalirudin for patients with acute coronary syndromes. The New England journal of medicine. 2006;355(21):2203-16. Epub 2006/11/25.

337. Messier SP, Loeser RF, Miller GD, Morgan TM, Rejeski WJ, Sevick MA, et al. Exercise and dietary weight loss in overweight and obese older adults with knee osteoarthritis: the Arthritis, Diet, and Activity Promotion Trial. Arthritis Rheum. 2004;50(5):1501-10. Epub 2004/05/18.

338. Miller GD, Rejeski WJ, Williamson JD, Morgan T, Sevick MA, Loeser RF, et al. The Arthritis, Diet and Activity Promotion Trial (ADAPT): design, rationale, and baseline results. Control Clin Trials. 2003;24(4):462-80. Epub 2003/07/17.

339. White CM, Caron MF, Kalus JS, Rose H, Song J, Reddy P, et al. Intravenous plus oral amiodarone, atrial septal pacing, or both strategies to prevent post-cardiothoracic surgery atrial fibrillation: the Atrial Fibrillation Suppression Trial II (AFIST II). Circulation. 2003;108 Suppl 1:II200-6. Epub 2003/09/13.

340. Bobadilla JF, Merikle E, Garcia M, Darba J, Sanchez C. Cost-effectiveness of atorvastatin plus amlodipine versus atorvastatin plus atenolol in hypertensive patients without previous coronary heart disease, normal to mildly elevated cholesterol levels and at least 3 cardiovascular risk factors [Abstract PCV34]. Value in Health. 2007;10(6):A415-A6.

341. Little P, Lewith G, Webley F, Evans M, Beattie A, Middleton K, et al. Randomised controlled trial of Alexander technique lessons, exercise, and massage (ATEAM) for chronic and recurrent back pain. Bmj. 2008;337:a884.

342. Bankhead C, Richards SH, Peters TJ, Sharp DJ, Hobbs FD, Brown J, et al. Improving attendance for breast screening among recent non-attenders: a randomised controlled trial of two interventions in primary care. J Med Screen. 2001;8(2):99-105. Epub 2001/08/02.

343. Hlatky MA, Boothroyd DB, Melsop KA, Kennedy L, Rihal C, Rogers WJ, et al. Economic outcomes of treatment strategies for type 2 diabetes mellitus and coronary artery disease in the Bypass Angioplasty Revascularization Investigation 2 Diabetes trial. Circulation. 2009;120(25):2550-8. Epub 2009/11/19.

344. Frye RL, August P, Brooks MM, Hardison RM, Kelsey SF, MacGregor JM, et al. A randomized trial of therapies for type 2 diabetes and coronary artery disease. The New England journal of medicine. 2009;360(24):2503-15. Epub 2009/06/09.

345. Brooks MM, Frye RL, Genuth S, Detre KM, Nesto R, Sobel BE, et al. Hypotheses, Design, and Methods for the Bypass Angioplasty Revascularization Investigation 2 Diabetes (BARI 2D) Trial. The American Journal of Cardiology. 2006;97(12, Supplement 1):9-19.

346. Sorensen JL, Masson CL, Delucchi K, Sporer K, Barnett PG, Mitsuishi F, et al. Randomized trial of drug abuse treatment-linkage strategies. J Consult Clin Psychol. 2005;73(6):1026-35. Epub 2006/01/06.

347. Sullivan FM, Swan IR, Donnan PT, Morrison JM, Smith BH, McKinstry B, et al. Early treatment with prednisolone or acyclovir in Bell's palsy. The New England journal of medicine. 2007;357(16):1598-607. Epub 2007/10/19.

348. Hernandez RA, Sullivan F, Donnan P, Swan I, Vale L, for the Bells Trial Group. Economic evaluation of early administration of prednisolone and/or aciclovir for the treatment of Bell's palsy. Family Practice. 2009;26(2):137-44.

349. Dickson K, Marshall M, Boyle J, McCartney E, O'Hare A, Forbes J. Cost analysis of direct versus indirect and individual versus group modes of manual-based speech-and-language therapy for primary school-age children with primary language impairment. Int J Lang Commun Disord. 2009;44(3):369-81. Epub 2008/09/30.

350. Brandon TH, Collins BN, Juliano LM, Lazev AB. Preventing relapse among former smokers: a comparison of minimal interventions through telephone and mail. J Consult Clin Psychol. 2000;68(1):103-13.

351. Bakhai A, Stone GW, Grines CL, Murphy SA, Githiora L, Berezin RH, et al. Cost-effectiveness of coronary stenting and abciximab for patients with acute myocardial infarction: results from the CADILLAC (Controlled Abciximab and Device Investigation to Lower Late Angioplasty Complications) trial. Circulation. 2003;108(23):2857-63. Epub 2003/11/12.

352. Stone GW, Grines CL, Cox DA, Garcia E, Tcheng JE, Griffin JJ, et al. Comparison of angioplasty with stenting, with or without abciximab, in acute myocardial infarction. The New England journal of medicine. 2002;346(13):957-66. Epub 2002/03/29.

353. Morisky DE, Levine DM, Green LW, Shapiro S, Russell RP, Smith CR. Five-year blood pressure control and mortality following health education for hypertensive patients. Am J Public Health. 1983;73(2):153-62.

354. Green LW, Levine DM, Deeds S. Clinical trials of health education for hypertensive outpatients: design and baseline data. Prev Med. 1975;4(4):417-25.

355. Johannesson M, Borgquist L, Jonsson B, Lindholm LH. The cost effectiveness of lipid lowering in Swedish primary health care. The CELL Study Group. J Intern Med. 1996;240(1):23-9. Epub 1996/07/01.

356. Lindholm LH, Ekbom T, Dash C, Isacsson AKE, Schersten B. Changes in cardiovascular risk factors by combined pharmacological and nonpharmacological strategies: the main results of the CELL Study. Journal of Internal Medicine. 1996;240(1):13-22.

357. Effing T, Kerstjens H, van der Valk P, Zielhuis G, van der Palen J. (Cost)-effectiveness of self-treatment of exacerbations on the severity of exacerbations in patients with COPD: the COPE II study. Thorax. 2009;64(11):956-62.

358. Clarkson JE, Turner S, Grimshaw JM, Ramsay CR, Johnston M, Scott A, et al. Changing clinicians' behavior: a randomized controlled trial of fees and education. J Dent Res. 2008;87(7):640-4. Epub 2008/06/25.

359. Tworoger SS, Yasui Y, Ulrich CM, Nakamura H, LaCroix K, Johnston R, et al. Mailing strategies and recruitment into an intervention trial of the exercise effect on breast cancer biomarkers. Cancer Epidemiol Biomarkers Prev. 2002;11(1):73-7. Epub 2002/01/30.

360. Franzosi MG, Maggioni AP, Santoro E, Tognoni G, Cavalieri E. Cost-effectiveness analysis of early lisinopril use in patients with acute myocardial infarction. Results from GISSI-3 trial. Pharmacoeconomics. 1998;13(3):337-46. Epub 1998/02/06.

361. GISSI-3: effects of lisinopril and transdermal glyceryl trinitrate singly and together on 6-week mortality and ventricular function after acute myocardial infarction. Gruppo Italiano per lo Studio della Sopravvivenza nell'infarto Miocardico. Lancet. 1994;343(8906):1115-22. Epub 1994/05/07.

362. Franzosi MG, Brunetti M, Marchioli R, Marfisi RM, Tognoni G, Valagussa F. Cost-effectiveness analysis of n-3 polyunsaturated fatty acids (PUFA) after myocardial infarction: results from Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto (GISSI)-Prevenzione Trial. Pharmacoeconomics. 2001;19(4):411-20. Epub 2001/06/01.

363. Halpern SD, Ubel PA, Berlin JA, Asch DA. Randomized Trial of $5 Versus $10 Monetary Incentives, Envelope Size, and Candy to Increase Physician Response Rates to Mailed Questionnaires. Medical Care. 2002;40(9):834-9.

364. Heart Protection Study Collaborative Group. Statin Cost-Effectiveness in the United States for People at Different Vascular Risk Levels. Circulation: Cardiovascular Quality & Outcomes. 2009;2(2):65-72.

365. Mihaylova B, Briggs A, Armitage J, Parish S, Gray A, Collins R. Cost-effectiveness of simvastatin in people at different levels of vascular disease risk: economic analysis of a randomised trial in 20,536 individuals. Lancet. 2005;365(9473):1779-85. Epub 2005/05/25.

366. MRC/BHF Heart Protection Study of cholesterol-lowering therapy and of antioxidant vitamin supplementation in a wide range of patients at increased risk of coronary heart disease death: early safety and efficacy experience. Eur Heart J. 1999;20(10):725-41. Epub 1999/05/18.

367. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet. 2002;360(9326):23-33. Epub 2002/07/13.

368. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet. 2002;360(9326):7-22. Epub 2002/07/13.

369. Jensen PS, Garcia JA, Glied S, Crowe M, Foster M, Schlander M, et al. Cost-effectiveness of ADHD treatments: findings from the multimodal treatment study of children with ADHD. Am J Psychiatry. 2005;162(9):1628-36. Epub 2005/09/02.

370. Wells KC, Pelham WE, Kotkin RA, Hoza B, Abikoff HB, Abramowitz A, et al. Psychosocial treatment strategies in the MTA study: rationale, methods, and critical issues in design and implementation. J Abnorm Child Psychol. 2000;28(6):483-505. Epub 2000/12/05.

371. A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. The MTA Cooperative Group. Multimodal Treatment Study of Children with ADHD. Arch Gen Psychiatry. 1999;56(12):1073-86. Epub 1999/12/11.

372. Atthobari J, Asselbergs FW, Boersma C, de Vries R, Hillege HL, van Gilst WH, et al. Cost-effectiveness of screening for albuminuria with subsequent fosinopril treatment to prevent cardiovascular events: A pharmacoeconomic analysis linked to the prevention of renal and vascular endstage disease (PREVEND) study and the prevention of renal and vascular endstage disease intervention trial (PREVEND IT). Clin Ther. 2006;28(3):432-44. Epub 2006/06/06.

373. Asselbergs FW, Diercks GF, Hillege HL, van Boven AJ, Janssen WM, Voors AA, et al. Effects of fosinopril and pravastatin on cardiovascular events in subjects with microalbuminuria. Circulation. 2004;110(18):2809-16. Epub 2004/10/20.

374. Richards SH, Bankhead C, Peters TJ, Austoker J, Hobbs FD, Brown J, et al. Cluster randomised controlled trial comparing the effectiveness and cost-effectiveness of two primary care interventions aimed at improving attendance for breast screening. J Med Screen. 2001;8(2):91-8. Epub 2001/08/02.

375. Rodewald LE, Szilagyi PG, Humiston SG, Barth R, Kraus R, Raubertas RF. A randomized study of tracking with outreach and provider prompting to improve immunization coverage and primary care. Pediatrics. 1999;103(1):31-8. Epub 1999/01/26.

376. Twardella D, Brenner H. Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. Tob Control. 2007;16(1):15-21.

377. Thomas KSP, Miller PM, Doherty MM, Muir KRP, Jones ACM, O'Reilly SCM. Cost Effectiveness of a Two-Year Home Exercise Program for the Treatment of Knee Pain. Arthritis & Rheumatism. 2005;53(3) Arthritis Care &(Research):388-94.

378. Desai AS, Solomon DH, Stone PH, Avorn J. Economic consequences of routine coronary angiography in low- and intermediate-risk patients with unstable angina pectoris. Am J Cardiol. 2003;92(4):363-7. Epub 2003/08/14.

379. Effects of tissue plasminogen activator and a comparison of early invasive and conservative strategies in unstable angina and non-Q-wave myocardial infarction. Results of the TIMI IIIB Trial. Thrombolysis in Myocardial Ischemia. Circulation. 1994;89(4):1545-56.

380. Tjan-Heijnen VCG, Caleo S, Postmus PE, Ardizzoni A, Burghouts JTM, Buccholz E, et al. Economic evaluation of antibiotic prophylaxis in small-cell lung cancer patients receiving chemotherapy: an EORTC double-blind placebo-controlled phase III study (08923). Annals of Oncology. 2003;14(2):248-57.

381. Tjan-Heijnen VC, Postmus PE, Ardizzoni A, Manegold CH, Burghouts J, van Meerbeeck J, et al. Reduction of chemotherapy-induced febrile leucopenia by prophylactic use of ciprofloxacin and roxithromycin in small-cell lung cancer patients: an EORTC double-blind placebo-controlled phase III study. Ann Oncol. 2001;12(10):1359-68. Epub 2002/01/05.

382. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. BMJ. 2004;329(7479):1377. Epub 2004/11/24.

383. Clarke PM, Gray AM, Briggs A, Stevens RJ, Matthews DR, Holman RR, et al. Cost-utility analyses of intensive blood glucose and tight blood pressure control in type 2 diabetes (UKPDS 72). Diabetologia. 2005;48(5):868-77.

384. Wesson J, Olawo A, Bukusi V, Solomon M, Pierre-Louis B, Stanback J, et al. Reaching providers is not enough to increase IUD use: a factorial experiment of 'academic detailing' in Kenya. J Biosoc Sci. 2008;40(1):69-82. Epub 2007/04/21.

385. Buse JB, Bigger JT, Byington RP, Cooper LS, Cushman WC, Friedewald WT, et al. Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial: design and methods. Am J Cardiol. 2007;99(12A):21i-33i.

386. Coulton S, Perryman K, Bland M, Cassidy P, Crawford M, Deluca P, et al. Screening and brief interventions for hazardous alcohol use in accident and emergency departments: A randomised controlled trial protocol. BMC Health Services Research. 2009;9(114).

387. Walker DG, Aedo C, Albala C, Allen E, Dangour AD, Elbourne D, et al. Methods for economic evaluation of a factorial-design cluster randomised controlled trial of a nutrition

supplement and an exercise programme among healthy older people living in Santiago, chile: The CENEX study. BMC Health Services Research. 2009;9(85).

388. Dangour AD, Albala C, Aedo C, Elbourne D, Grundy E, Walker D, et al. A factorial-design cluster randomised controlled trial investigating the cost-effectiveness of a nutrition supplement and an exercise programme on pneumonia incidence, walking capacity and body mass index in older people living in Santiago, Chile: The CENEX study protocol. Nutrition Journal. 2007;6(14).

389. Newbury-Birch D, Bland M, Cassidy P, Coulton S, Deluca P, Drummond C, et al. Screening and brief interventions for hazardous and harmful alcohol use in probation services: a cluster randomised controlled trial protocol. BMC Public Health. 2009;9:418.

390. Coombes BK, Bisset L, Connelly L, Brooks P, Vicenzino B. Optimising corticosteroid injection for lateral epicondylalgia with the addition of physiotherapy: A protocol for a randomised control trial with placebo comparison. BMC Musculoskelet Disord. 2009;10(1):76. Epub 2009/06/26.

391. Cals JW, Hopstaken RM, Butler CC, Hood K, Severens JL, Dinant GJ. Improving management of patients with acute cough by C-reactive protein point of care testing and communication training (IMPAC3T): study protocol of a cluster randomised controlled trial. BMC Fam Pract. 2007;8:15. Epub 2007/03/31.

392. Abbott JH, Clare MC, McKenzie JE, David GD, Theis JC, Campbell AJ. Exercise therapy, manual therapy, or both, for osteoarthritis of the hip or knee: A factorial randomised controlled trial protocol. Trials. 2009;10(11).

393. Kyriakides TC, Babiker A, Singer J, Cameron W, Schechter MT, Holodniy M, et al. An open-label randomized clinical trial of novel therapeutic strategies for HIV-infected patients in whom antiretroviral therapy has failed: rationale and design of the OPTIMA Trial. Control Clin Trials. 2003;24(4):481-500. Epub 2003/07/17.

394. Barber MD, Brubaker L, Menefee S, Norton P, Borello-France D, Varner E, et al. Operations and pelvic muscle training in the management of apical support loss (OPTIMAL) trial: design and methods. Contemp Clin Trials. 2009;30(2):178-89. Epub 2009/01/10.

395. Kaner E, Bland M, Cassidy P, Coulton S, Deluca P, Drummond C, et al. Screening and brief interventions for hazardous and harmful alcohol use in primary care: A cluster randomised controlled trial protocol. BMC Public Health. 2009;9(287).

396. Coleman T, McEwen A, Bauld L, Ferguson J, Lorgelly P, Lewis S. Protocol for the Proactive Or Reactive Telephone Smoking CeSsation Support (PORTSSS) trial. Trials. 2009;10:26. Epub 2009/04/30.

397. Andrews PJ, Avenell A, Noble DW, Campbell MK, Battison CG, Croal BL, et al. Randomised trial of glutamine and selenium supplemented parenteral nutrition for critically ill patients. Protocol Version 9, 19 February 2007 known as SIGNET (Scottish Intensive care Glutamine or seleNium Evaluative Trial). Trials. 2007;8(1):25. Epub 2007/09/22.

398. Ayles HM, Sismanidis C, Beyers N, Hayes RJ, Godfrey-Faussett P. ZAMSTAR, The Zambia South Africa TB and HIV Reduction study: Design of a 2 x 2 factorial community randomized trial. Trials. 2008;9(63).

399. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren M, et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). Eur Heart J. 2012;33(13):1635-701. Epub 2012/05/05.

400. National Institute for Health and Care Excellence. Type 2 diabetes: the management of type 2 diabetes. 2010 [18th October 2013]; Available from: http://www.nice.org.uk/nicemedia/live/12165/44320/44320.pdf.

401. National Institute for Health and Clinical Excellence. Statins for the prevention of cardiovascular events. 2006 [18 October 2013]; Available from: http://www.nice.org.uk/nicemedia/live/11564/33151/33151.pdf.

402. Hartweg J, Perera R, Montori V, Dinneen S, Neil HA, Farmer A. Omega-3 polyunsaturated fatty acids (PUFA) for type 2 diabetes mellitus. Cochrane Database Syst Rev. 2008(1):CD003205.

403. British Medical Association. British National Forumulary 66. September 2013 [8th October 2013]; Available from: http://www.bnf.org/bnf/index.htm.

404. Health & Social Care Information Centre. Prescription Cost Analysis - England, 2012 [NS]. 2013 [18th December 2013]; Available from: http://www.hscic.gov.uk/catalogue/PUB10610.

405. Department of Health. National Schedule of Reference Costs 2012-13 for NHS Trusts and PCTs combined. London, United Kingdom: Department of Health; November 2013 [11th February 2014]; Available from: https://www.gov.uk/government/publications/nhs-reference-costs-2012-to-2013.

406. Curtis L. Unit Costs of Health and Social Care 2013. Canterbury, UK: PSSRU Personal Social Services Research Unit; 2013.

407. Curtis L. Unit Costs of Health and Social Care 2010. Canterbury, UK: PSSRU Personal Social Services Research Unit; 2010.

408. Clarke P, Gray A, Legood R, Briggs A, Holman R. The impact of diabetes-related complications on healthcare costs: results from the United Kingdom Prospective Diabetes Study (UKPDS Study No. 65). Diabet Med. 2003;20(6):442-50. Epub 2003/06/06.

409. Clarke P, Gray A, Holman R. Estimating utility values for health states of type 2 diabetic patients using the EQ-5D (UKPDS 62). Med Decis Making. 2002;22(4):340-9.

410. Curtis L. Unit Costs of Health and Social Care 2012. Canterbury, UK: PSSRU Personal Social Services Research Unit; 2012.

411. Sterne J, Bradburn M, Egger M. Chapter 18: Meta-analysis in Stata. In: Egger M, Davey Smith G, Altman D, editors. Systematic reviews in health care: Meta-analysis in context: BMJ books; 2001.