Independence, interactions and inference in partial factorial trials

Helen A Dakin,[1] Alastair M Gray,[1] Graeme S MacLennan,[2] Richard W Morris[3] and David W Murray[4]

[1] Health Economics Research Centre, University of Oxford
[2] Health Services Research Unit, University of Aberdeen, Aberdeen
[3] Department of Primary Care & Population Health, University College London
[4] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford

Abstract

Introduction: Partial factorial trials, such as the Women's Health Initiative and UK Prospective Diabetes Study, compare ≥2 pairs of treatments on overlapping patient groups, randomising some patients to >1 comparison. They are generally analysed as several overlapping trials addressing multiple independent questions. However, those patients randomised to ≥2 treatments can also be analysed "inside-the-table" to investigate interactions and inform joint decisions between mutually-exclusive combinations of treatment strategies.

Aims: To explore the implications of partial factorial design on the methods and results of economic evaluation and compare the implications of considering the two factors simultaneously rather than drawing independent conclusions about each comparison.

Case study: The Knee Arthroplasty Trial randomised 2252 patients undergoing knee replacement to one or more comparisons between surgery types: patella (kneecap) resurfacing vs. no resurfacing; mobile vs. fixed bearing; and all-polyethylene vs. metal-backing. Three hundred and thirty-eight patients were randomised in the patella comparison and one other comparison.

Methods: We estimated total costs and QALYs within 10 years of primary knee replacement. "At-the-margins" estimates of incremental costs, QALYs and net benefits for each comparison (assuming no interaction) across all patients randomised were compared against two types of "inside-the-table" analysis (including interactions). The statistical significance of interactions and their impact on Akaike's information criterion (AIC), expected net benefits and value of perfect information was assessed.

Results: Non-significant qualitative interactions for net benefits were observed between patella resurfacing and both metal-backing and mobile bearings (p=0.06): mobile bearings and metal-backing increased expected net benefits in patients randomised to patella resurfacing, but not for those with no resurfacing. For the mobile bearing comparison, making a joint decision between the four mutually-exclusive strategies gave the same cost-effectiveness conclusions as treating the two factors as independent, although the two approaches gave very different conclusions for metal-backing. Considering interactions for net benefit decreased AIC in both cases.

Discussion: There is some evidence of interactions between patella resurfacing and other design aspects, although these were not expected a priori and could be explained by chance. The importance of interactions may be more usefully assessed by the bias or opportunity loss from ignoring interactions, than by standard statistical inference. All partial factorial trials should include sensitivity analyses exploring interactions. However, for partial factorial trials mirroring routine clinical practice, "at-the-margins" analysis may be the most useful way to estimate average costs and benefits on the whole trial population, even in the presence of interactions.

Introduction

Partial factorial trials[1] evaluate 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 (Figure 1). High-profile examples include the Women's Health Initiative (which compared HRT vs no HRT, dietary modification vs. no modification and vitamin D/calcium supplementation vs. placebo in overlapping groups) and the UK prospective diabetes study (which compared alternative strategies for glycaemic and blood pressure control in overlapping groups).

Figure 1: Schematics of full and partial factorial designs

Full factorial trial (n=400)

	Placebo of A	**Drug A**
Placebo of B	105 pts (**00**)	97 pts (**A0**)
Drug B	98 pts (**0B**)	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. 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 factor C (e.g. the difference between HRT and no HRT) differs depending on whether D (e.g. dietary modification) is also given. Partial factorial designs are often used when economic, geographic, or clinical constraints restrict the comparisons to which patients can be randomised (1), since this design facilitates flexible recruitment strategies, such as letting patients (2) or clinicians (6) choose which comparisons to be randomised into, or recruiting only those patients in certain countries (7) or those with specific comorbidites, clinical characteristics or laboratory findings to a second (or even third) comparison (2, 8).

Despite extensive research and established guidelines on factorial trials in general, only one review article has discussed partial factorial trials (1) and no other research teams have investigated economic evaluation of full factorial trials. As we discussed previously (9), full

[1] Here, we definite "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 (1-3). 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 (4, 5). Our literature reviews suggest that most studies meeting our definition of partial factorial trials describe their design as simply "factorial".

factorial trials can be analysed “at-the-margins” or “inside-the-table”. “At-the-margins” analysis assumes that the effect of the two factors is purely additive (with no interaction) and treats the factorial trial as two overlapping two-arm trials, evaluating the effect of treatment A by comparing outcomes in cells A0 and AB combined with those in 00 and 0B combined (Figure 1). It is equivalent to regression without an interaction term and has greater statistical power than inside-the-table analysis unless interactions are very large (10), but gives biased, misleading results whenever any interaction exists (5, 10, 11). In inside-the-table analysis, we analyse outcomes for cells 00, A0, 0B and AB separately and make pairwise comparisons between cells. Inside-the-table analysis is equivalent to regression with an interaction term and assumes that interactions exist: i.e. that the effect of factor A differs depending on factor B. It is unbiased, but generally has lower statistical power than at-the-margins analysis unless interactions are very large.

The type of analysis used and the importance and impact of interactions is likely to be of particular importance for costs, QALYs and net benefits, where large interactions are likely to be relatively common (9, 12). Furthermore, economic evaluation focuses on estimation of incremental cost-effectiveness ratios (ICERs) rather than statistical inference, which may mean that inside-the-table analysis is preferable to at-the-margins analysis, despite the latter’s greater statistical power (9, 12). Additionally, the two approaches mirror the two types of decision rule used in cost-effectiveness analysis (13). Inside-the-table analysis naturally considers the decision problem as a joint decision between the mutually exclusive treatments 00, A0, 0B and AB, where we maximise health gains from the healthcare budget by adopting the treatment with highest expected net benefits (9).[2] By contrast, at-the-margins analysis assumes that the decision between A and not-A is independent of that between B and not-B and would lead us to adopt A if it has positive incremental net benefits vs. not-A and (separately) adopt B if it has positive incremental net benefits (9).

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. We 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 “inside-the-table” like a full factorial trial. This analysis

[2] In this paper, we assume that inference is irrelevant to decisions about which treatment we adopt for use in the NHS based on current information (14) and that the treatment with the highest expected net benefit should be adopted. We discuss the appropriateness of this assumption and its implications for factorial trials in the discussion.

ensures unbiased estimation of interactions, but excludes large numbers of patients (41% in the case of Figure 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 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 with per protocol analysis (15). Furthermore, 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 some aspect of baseline 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 subgroup analysis of two-arm trials (16, 17). By contrast, full factorial randomisation with intention to treat (ITT) analysis and concealment of allocation avoid selection bias and ensure that the only systematic difference between randomised groups is the allocated treatment.

Although inside-the-table analysis is less prone to bias, at-the-margins analysis 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 when the ratio of B to not-B in the trial reflects that in the setting of interest (e.g. routine clinical practice) even in the presence of interactions (5, 18, 19). 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 may still be benefits from evaluating interactions and exploring whether outcomes differ with B.

All five of the partial factorial trials with economic evaluations that we are aware of have been analysed at-the-margins as overlapping trials addressing independent questions (e.g. evaluating Drug C by comparing outcomes across cells C0, CD and C- with those in cells 00, 0D and 0-) (20-24). Two studies evaluated interactions, of which one analysed the subsets of patients randomised to each combination of factors separately (24) and the other did not clearly describe their methods (20). The three remaining studies reported the different factors in different papers (21-23).

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 (1) at-the-margins analysis, (2) inside-the-table analysis on patients randomised to >1 comparison and (3) inside-the-table subgroup analysis.

Case study

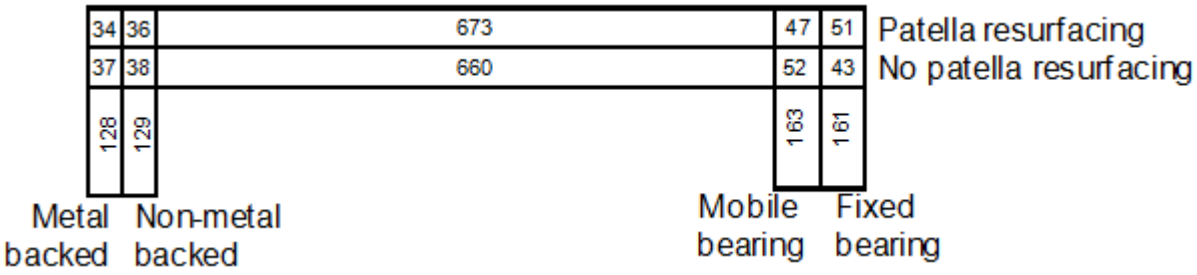
The Knee Arthroplasty Trial (KAT) is a pragmatic partial factorial randomised controlled trial evaluating three[3] aspects of knee prosthesis design (6, 25):

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 the patella resurfacing comparison as well as either Comparison A or Comparison B (Figure 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. 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. Given that there was no previous evidence on interactions and no clinical reason why one would be expected, the primary clinical and economic analyses were conducted “at-the-margins” and the sample size was calculated on that basis, although exploration of interactions was planned in the protocol (3).

[3] A fourth comparison comparing total knee replacement with unicompartmental replacement was also evaluated but is not discussed in this paper since it did not overlap with other comparisons.

Figure 2: Design of the Knee Arthroplasty Trial (n=2252)



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

Cost-utility analysis was conducted on results at a median of 10 years' post-operation follow-up. Annual EQ-5D utility measurements were used to calculate QALYs; costing methodology has been reported previously (25). Costs were discounted at 3.5% per annum and are presented in 2011/12 pounds.

Multiple imputation (MI) (26) was used to impute missing data on utilities and resource use. Partial factorial designs raise particular challenges for MI. Firstly, MI could be conducted once on the entire trial population, or separately for each comparison. We 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. Secondly, the optimal coding of treatment variables is unclear. To avoid bias, the variables, functional forms and interaction terms in the MI model should match those used in the analysis model (26). For full factorial trials, variables for all potential analysis models can be captured by including one dummy for each comparison and one for each interaction term. However, for partial factorial trials, having one dummy per comparison 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). We therefore included six treatment dummies (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.

Inverse probability weighting (IPW) (27) was used to allow for administrative censoring of patients who had not reached 10 years when the database was closed. Ordinary least squares (OLS) regression was used to adjust QALYs for imbalance in baseline utility (28). Bootstrapping was used to quantify uncertainty and was done 100 times for each of the 100 imputed datasets; point estimates are based on means for raw data (with no bootstrapping). Results for the 100 imputed datasets were combined using Rubin's rule (26). Total net monetary benefit (NMB) for each treatment was calculated for point estimates and for each

bootstrap replicate in each dataset. The proportion of all bootstrap replicates (across all datasets) where each treatment had highest NMBs was plotted on cost-effectiveness acceptability curves (CEACs). The expected value of perfect information (EVPI) was calculated as the mean maximum NMB accrued in each bootstrap replicate minus the mean NMBs for the treatment with highest expected NMB. Value of sample information and population EVPI were not calculated as the aim was to explore factors increasing or decreasing EVPI within this patient group, rather than make research recommendations.

Analytical methods

Bootstrapping and analysis of results using IPW were repeated using the following three approaches to evaluate the impact on the results. Unless otherwise specified, all analyses were conducted in Stata version 12 (StataCorp LP, College Station, TX), all p-values are two-sided and NMBs are reported at a £20,000 per QALY ceiling ratio (29).

Analysis 1 (base case): 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 is 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 (and therefore likely to have a 50% chance of having patella resurfacing in routine clinical practice), 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.

Analysis 2: Primary analysis of interactions on patients randomised to >1 comparison

Analysis 2 analysed the subset of patients randomised to >1 comparison as a small full-factorial trial on an ITT basis. 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 from a dummy indicating randomised allocation in comparison A or B, a dummy for whether patients were randomised to patella resurfacing, an interaction term equaling treatment*patella, 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.

Analysis 3: Secondary analysis of interactions as subgroup analysis

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 patella resurfacing. 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 used to predict costs or QALYs accrued in each year using OLS. 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.

Evaluating impact of interactions in Analyses 2 and 3

The impact of interactions was estimated using several measures:

- Statistical significance: The interaction between treatment allocations was calculated for total costs, total QALYs and total NMBs in each bootstrap replicate (interaction = 00-A0-0B+AB) and results across imputed datasets were pooled using Rubin's rule (26) to estimate the mean, standard error (SE) and two-sided p-value.
- Akaike information criterion (AIC): Mixed models were estimated to predict QALYs and (separately) NMBs for each patient in each year based on random constants for each patient, baseline utility, a dummy for randomised to patella resurfacing and a dummy for randomised allocation in comparison A or B, with or without an interaction term. Mixed models were used to obtain a single AIC estimate across all 10 study years, allowing for administrative censoring. Since mixed models of annual costs did not converge due to excess zeros, generalised linear models with log-link and gamma family were used to predict total cost over the period from randomisation to study exit with the same explanatory variables (excluding baseline utility). AIC for models with and without interactions were compared in each of the 100 imputed datasets. Since AIC values

cannot be combined across imputed datasets (26), we present the proportion of datasets in which AIC was lower with interactions.

- Expected opportunity loss from ignoring interactions: We propose this concept to estimate the additional value of allowing for interactions and making a joint decision about the two factors simultaneously, rather than considering each independently and assuming no interaction. This concept is analogous to the value of stratification proposed previously (30, 31). The opportunity loss from ignoring interactions under current information equals the difference between the NMB of the treatment we would adopt based on inside-the-table analysis and the NMB of the treatment we would adopt based on conducting at-the-margins analysis on each factor. Both NMB estimates were based on inside-the-table analysis of the same dataset to ensure consistent, unbiased estimation. To calculate the opportunity loss under perfect information, we repeated this calculation for each bootstrap replicate and averaged across replicates. The proportion of bootstrap replicates in which the opportunity loss is >0 gives the probability that at-the-margins analysis would give a misleading conclusion. The opportunity loss under current information (which is analogous to the static value of heterogeneity proposed previously (31)) equals zero whenever inside-the-table and at-the-margin analyses give the same conclusion, whereas the loss under perfect information may differ.
- Estimated bias from ignoring interactions: The bias in main effect size estimates caused by ignoring interaction terms is estimated as half the size of the interaction term (10).

Results

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

Analysis 1 evaluated each of the three comparisons as independent decisions on all patients randomised to that comparison. 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[4] (Table 1). 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

[4] Notably, this analysis provides a real-life example of the situation identified hypothetically previously (32, 33), in which net benefits differ significantly (p=0.96 on a one-tailed test), despite non-significant differences in both costs and effects.

independent options, we would therefore recommend patella resurfacing, metal-backing and (more hesitantly) mobile bearing.

Table 1: Base case results for all three comparisons. Numbers in brackets are standard errors.

		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.

Analysis 2 results: Primary analysis of interactions in factorial manner

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 2) and all SEs were at least twice as large as those in Analysis 1 due to the substantially smaller sample size.

Table 2: Primary analysis of interactions: including only those patients randomised to >1 comparison. Numbers in brackets are standard errors.

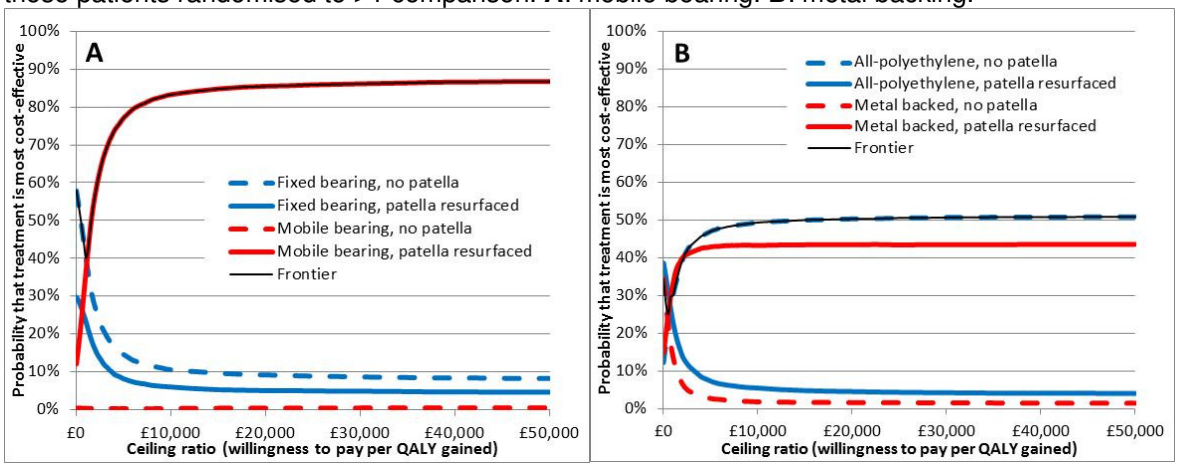
		Mobile bearing	Fixed bearing	Metal backing	All-polyethylene
Patella resurfaced	No. pts	47	51	34	36
	Cost	£9,068 (£466)	£9,169 (£1,165)	£8,036 (£411)	£7,833 (£567)
	QALYs	5.559 (0.264)	4.959 (0.289)	5.518 (0.337)	5.046 (0.330)
	NMB*	£102,110 (£5,372)	£90,015 (£6,256)	£102,327 (£6,870)	£93,087 (£6,837)
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, whereas allowing for interactions in NMB decreased AIC in all datasets and interactions for costs decreased AIC in 91% of cases. This suggests that

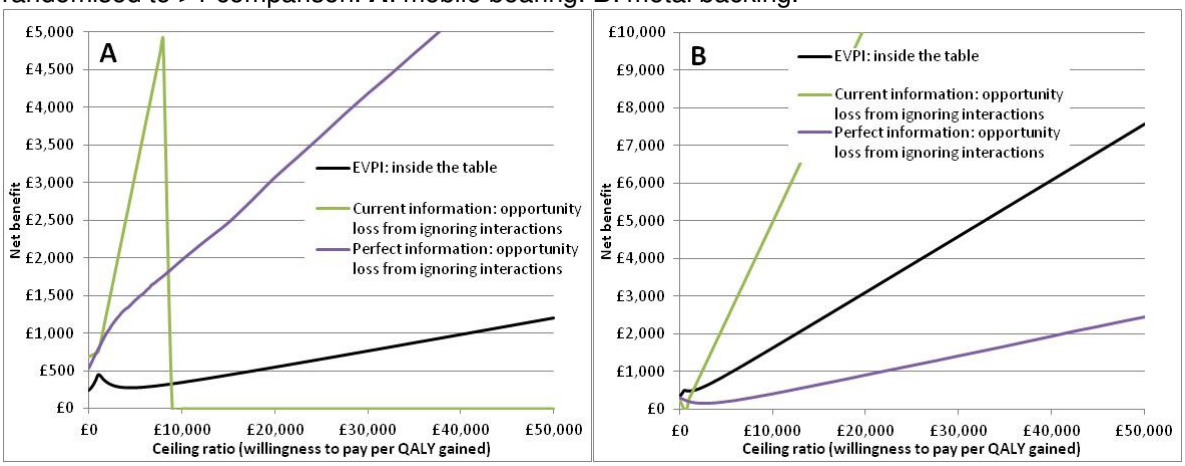
interactions in costs and NMB improve model fit but that interactions in QALYs worsen it; the interpretation of this conflicting finding is unclear. All three interactions were qualitative, with the incremental effect of mobile bearings changing sign depending on whether patients were allocated to patella resurfacing or no resurfacing (Table 2). 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 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 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 above £8,000/QALY gained, there is no opportunity loss from ignoring interactions under current information (Figure 4A). However, the two analyses gave different conclusions in 37% of bootstrap replicates, giving an opportunity loss of £3,071 under perfect information.

Figure 4: EVPI and expected opportunity loss from 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=0.047) and non-significant qualitative interactions for costs (£506 [SE: £907]; p=0.577) and NMB (£20,788 [SE: £11,094]; p=0.060). This suggests that the bias within at-the-margins estimates of incremental NMB could be £10,394 (£20,788/2). In all imputations, allowing for interactions for costs and QALYs increased AIC, whereas allowing for interactions in NMB improved model fit. 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 2). However, the subset of patients randomised in these two comparisons tended to accrue lower costs and higher QALYs than those randomised to just comparison B (Tables 1-2), suggesting that this patient population may not be typical. 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.

Nonetheless, there is substantial uncertainty around this finding (Figure 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 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 these two comparisons suggested that all-polyethylene and patella resurfacing were cost-effective at a £20,000/QALY ceiling ratio. On this basis, the opportunity loss from 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 value of making a joint decision under perfect information was £4,176.

Analysis 3 results: 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 3). 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.

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 5A; cf. 86% in Analysis 2) and the EVPI for the whole decision was £1,901/patient (Figure 6A; cf. £545 in Analysis 2).

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 >£1000/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.

Table 3: 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.

Figure 5: 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.

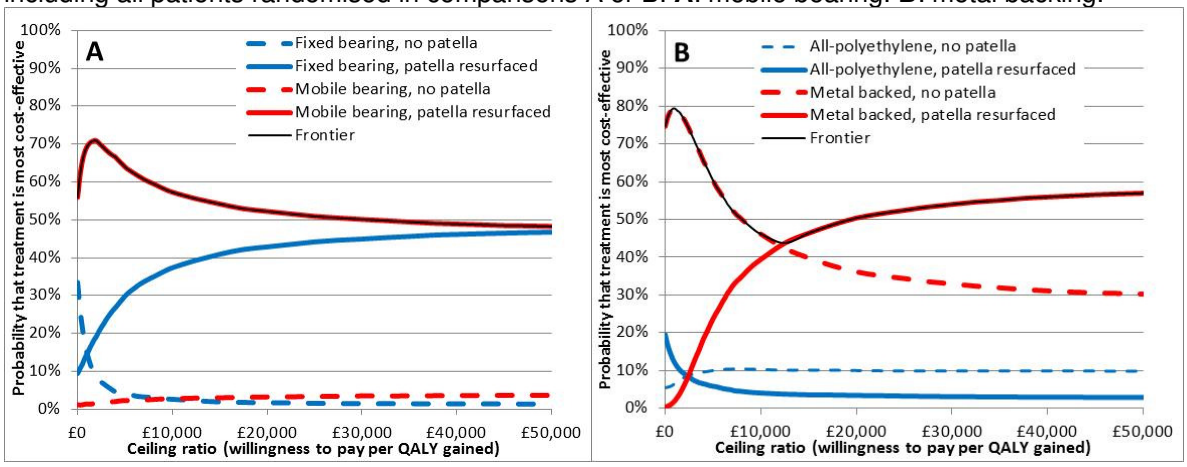
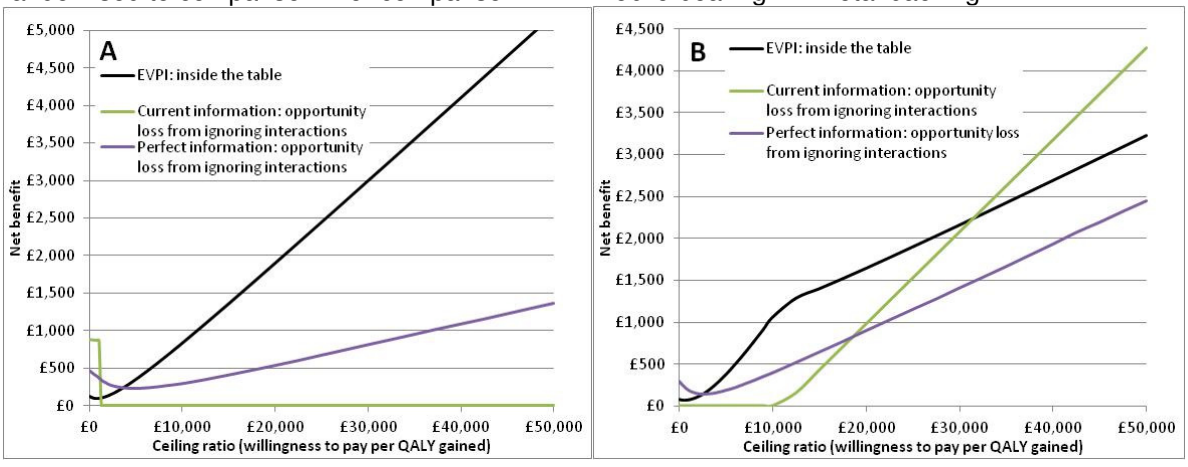


Figure 6: EVPI and expected opportunity loss from ignoring interactions under current and perfect information based on the secondary (subgroup) analysis of interactions: including all patients randomised to comparison A or comparison B. **A**: mobile bearing. **B**: metal backing.



For the metal backing comparison, SEs estimated in Analysis 3 (Table 3) 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 5B). The EVPI for the joint decision was £1,645/patient (Figure 6B): 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 loss from 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%.

Discussion

Our study suggests that there is evidence that the costs, QALYs and cost-effectiveness of mobile bearings and metal backing may be affected by whether or not patients also have their patella resurfaced. 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' quality of life 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.

Which analysis?

Our three analyses provided different estimates of the importance and magnitude of interactions and differences in costs, QALYs and NMB. Conclusions for the metal backing comparison also varied between analyses: Analysis 2 also suggested that all-polyethylene with no patella resurfacing was best, whereas Analyses 1 and 3 found metal backing with patella resurfacing to be best.

Analysing the subset of patients randomised to >1 comparison like a full factorial trial (Analysis 2) provides an unbiased estimation 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 higher 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, many partial factorial trials may provide a larger sample size for this analysis, mitigating some of these disadvantages: ≥39% of participants in the other partial factorial trials with economic evaluations were in >1 comparison, compared with only 15% (338/2252) in KAT.

By contrast, Analysis 3 included almost all patients recruited to comparisons A or B. This larger sample size increased statistical power and reduced the risk of chance fluctuations in outcomes between treatment combinations, giving smaller SEs and smaller interactions. However, in order to evaluate interactions in patients randomised to only one comparison, it was necessary to evaluate 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. We therefore cannot necessarily infer a causal relationship between patella resurfacing and Comparisons A or B. More fundamentally, we also cannot draw unbiased conclusions about the causative relationship between patella resurfacing and costs or QALYs because this comparison is prone to similar selection bias as observational studies. 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 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 anti-hypertensive treatment or placebo. As result, analysing the partial factorial trials using subgroup analyses is not an appropriate way to evaluate interactions or make a joint decision about the two factors simultaneously. Analysing the subset of patients randomised to >1 comparison like a full factorial trial 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 or not; this hypothesis is partially supported by National Joint Registry 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 taking account of the entire trial population.

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 (34) 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 (35, 36), which may provide a compromise between including or excluding interaction terms.

Inference and interactions

Claxton previously argued that statistical inference is irrelevant to healthcare adoption decisions as it is based on arbitrary cut-offs, favours treatments already in routine use and fails to maximise health gains produced from the healthcare budget (14). Our analysis 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, we 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 underpowered to detect interaction effects for primary outcomes (11), 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 weighs 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, finding interactions to be important on a NMB scale, but not for costs and QALYs. We also estimated AIC using a separate mixed model analysis, since it is unclear whether it is possible to estimate information criteria directly from IPW bootstrap results. Estimates of the likely bias and opportunity loss 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 incremental NMB estimates by up to £10,000. We would be grateful for feedback on the usefulness of these measures for weighing up the risks and benefits of allowing for interactions in partial factorial trials.

Conclusions

We demonstrate that partial factorial trials can be used to evaluate interactions, although they are best left for situations where interactions are expected to be negligible. However, 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 (5, 10, 11), so the potential for interactions should always be evaluated in sensitivity analysis. Interactions should be evaluated using ITT 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, without making any assumptions about interactions.

Acknowledgements

We would like to thank the KAT Trial Group for their role in designing and running the KAT trial. The KAT trial was funded by the NIHR Health Technology Assessment Programme (project number 95/10/01) and will be published in full as a Health Technology Assessment. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the HTA programme, NIHR, NHS or the Department of Health.

References

1. Allore HG, Murphy TE. An examination of effect estimation in factorial and standardly-tailored designs. Clin Trials. 2008;5(2):121-30.
2. 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.
3. KAT: Knee Arthroplasty Trial Protocol. Version 6 - March 2009. 2009; Available from: http://www.hta.ac.uk/protocols/199500100001.pdf.
4. Meinhert CL. Clinical trials: Design, conduct, and analysis. New York and Oxford: Oxford University Press; 1986.
5. 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.
6. 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.
7. 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.
8. 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.
9. Dakin HA, Gray A. Economic evaluation of factorial randomised controlled trials: Why the method of analysis matters Presented at the Health Economists' Study Group meeting 23-25 June 2010, Cork, Ireland, . 2010.
10. 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.
11. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol. 2003;3:26.
12. 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.
13. Karlsson G, Johannesson M. The decision rules of cost-effectiveness analysis. Pharmacoeconomics. 1996;9(2):113-20. Epub 1996/01/08.
14. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. J Health Econ. 1999;18(3):341-64.
15. Montori VM, Guyatt GH. Intention-to-treat principle. CMAJ. 2001;165(10):1339-41. Epub 2002/01/05.

16. 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.
17. 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.
18. Cox DR. Planning of experiments. Wiley Classics Edition published 1992 ed. New York: Wiley; 1958.
19. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied Linear Statistical Models. 4th ed. Chicago, IL: Irwin; 1996.
20. 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.
21. 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.
22. UK Prospective Diabetes Study (UKPDS). VIII. Study design, progress and performance. Diabetologia. 1991;34(12):877-90. Epub 1991/12/01.
23. 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.
24. 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.
25. 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).
26. 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.
27. 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.
28. 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.
29. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. June 2008 [13th July 2010]; Available from: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf.
30. 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.
31. 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.
32. Dakin H, Wordsworth S. Cost-Minimisation Analysis Versus Cost-Effectiveness Analysis, Revisited. Health Econ. 2011. Epub 2011/11/24.
33. 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.
34. Sadatsafavi M, Marra C, McCandless L, Bryan S. The challenge of incorporating external evidence in trialbased cost-effectiveness analyses: the use of resampling method. HEDG Health, Econometrics and Data Group, University of York, working paper. 2012;WP 12/24.
35. Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. Biometrics. 1997;53(2):456-64.
36. 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.