Why consider interactions in trial-based economic evaluation? A case study of a factorial trial

Helen Dakin,[1] Sarah Wordsworth,[1] Alastair Gray,[1] Chris Rogers,[2] Giselle Abangma,[1] and Barney Reeves[2]
[1] Health Economics Research Centre, University of Oxford
[2] Clinical Trials and Evaluation Unit, School of Clinical Sciences, University of Bristol

Abstract

Background: Factorial trials make ≥2 treatment comparisons simultaneously and can evaluate interactions between treatments. Clinical outcomes are often analysed without interactions to maximise statistical power, as genuine interactions between treatment factors are rare. However, large interactions for costs and QALYs are more common due to diminishing marginal returns for utilities and multiplicative effects on costs or event rates.

Aim: To examine the effect of different assumptions about interactions within a factorial trial-based economic evaluation.

Case study: IVAN comprises a randomised non-inferiority trial comparing two dosing regimens (monthly and as-needed) of two drugs (Avastin and Lucentis) in age-related macular degeneration.

Hypotheses: Interactions are unlikely for QALYs and non-drug costs unless efficacy/safety differs between treatments. However, large interactions are expected for drug costs, as reducing dosing frequency will have a proportionately greater effect for Lucentis than Avastin (which is much cheaper). Ignoring interactions by presenting results for monthly vs. as-needed dosing, averaged over both drugs, would therefore be misleading. Conversely, allowing for unimportant interactions would reduce efficiency and bias CEACs.

Analytical methods: To maximise statistical power while minimising bias, we propose allowing for interactions only in those components of net benefit where interactions are statistically significant or change the conclusions. We use bootstrapping to combine costs and QALYs, while allowing for correlations between them and propagating uncertainty.

Results/conclusions: Assumptions about interactions alter ICERs, CEACs and value of information in our case study. Health economists should always consider interactions when analysing factorial trials and take account of any interactions that could change conclusions.

1. Introduction

While most randomised controlled trials (RCTs) compare two interventions, factorial RCTs simultaneously evaluate two or more "factors", for example, comparing Drug A vs. placebo and Drug B vs. placebo in all combinations, or comparing two alternative dosing regimens for two drugs. In addition to answering several questions simultaneously (1-5), factorial RCTs can also test whether factors interact: for example, whether A has a greater (or lesser) effect when given alongside B. Interactions may be super-additive/synergistic (whereby the effect of A and B together is more than the effect of each separately), sub-additive/ antagonistic (where the effect of A and B together is less than the sum of the parts), or qualitative (where A increases outcomes in the absence of B but decreases it in the presence of B, or vice versa).

Factorial trials may be analysed using different assumptions about interactions. Clinical endpoints are typically analysed using "at-the-margins" analysis or using regression without interaction terms (6-8), which treats the design as two overlapping trials, for example estimating the difference between Drug A and no Drug A, averaged over patients with/without B. This assumes that the factors have purely additive effects: for example, that A increases efficacy by the same amount regardless of whether B is given. If there is genuinely no interaction, ignoring interactions is statistically efficient, answering two questions with the same sample size required for one (9-11). However, this form of analysis gives biased or misleading results if there is any interaction (9, 10, 12-14), since we ignore the fact that A has a greater (or lesser) effect in patients who are also receiving B. By contrast, allowing for interactions by using regression with an interaction term or "inside-the-table" analysis (analysing each cell of the design separately) gives unbiased results (12, 13), but is less efficient than ignoring interactions unless interactions are very large (7, 13, 15), since treatment effects are calculated using only two cells of the factorial design, rather than averaging across the whole sample.

We previously explored the impact of including vs. excluding interactions on a simulated economic evaluation and showed that at-the-margins analysis can give misleading conclusions about which treatment maximises net benefits. At-the-margins analysis also treats the two factors as independent options rather than considering the cells of the factorial design as mutually exclusive alternatives (16). Furthermore, we proposed several

mechanisms that may cause interactions to be larger for costs, quality-adjusted life-years (QALYs) and net benefits than for clinical outcomes, including diminishing marginal returns for utilities and life expectancy and multiplicative effects on costs or event rates. Additionally, the choice of analytical approach is likely to be of greater importance for economic evaluation than for clinical outcomes as the conclusions of economic evaluation are based almost solely on estimation of point estimates, whereas most clinical conclusions are driven by statistical inference.

Although the methods for the economic evaluation of two-arm RCTs are well-established (17, 18), we are aware of no research on methods for economic evaluation of factorial RCTs. Despite growing numbers of such studies, many evaluations of factorial RCTs take no account of interactions. We therefore aim to test these theoretical predictions and examine the effect of making different assumptions about interactions in the context of a "real" factorial trial-based economic evaluation.

2. Case study

Age-related macular degeneration (AMD) is the UK's leading cause of blindness (19) reducing the acuteness of vision in the centre of the visual field, affecting patients' ability to read, recognise faces and carry out daily activities. Lucentis (ranibizumab) is a licensed biologic drug that prevents vision loss in 95% of patients at 1-year (20, 21) and reduces disease progression, but costs £742 per dose (22). Given budget pressures, the NHS is keen to find out whether it would be safe, effective and/or cost-effective to reduce dosing frequency and/or use a closely-related drug (Avastin, bevacizumab) that is much less costly, but is currently licensed only for certain cancers, not AMD.

IVAN comprises a factorial randomised trial comparing Avastin vs Lucentis and monthly injections vs. giving three monthly injections then monitoring outcomes and giving further courses of three injections to any patients showing disease progression or worsening vision (23) (Table 1). The trial recruited 610 patients and was powered to test whether Avastin is non-inferior to Lucentis for best corrected distance visual acuity and whether as-needed dosing is non-inferior to monthly therapy at two years. One-year results have been published (23), but follow-up is continuing.

Table 1: Design of the IVAN trial

		Factor 1: Drug	
		Lucentis	**Avastin**
Factor 2: Dosing regimen	**Monthly injections**	Monthly Lucentis	Monthly Avastin
	As-needed dosing	As-needed Lucentis	As-needed Avastin

One-year results suggest that as-needed and monthly dosing had equal efficacy, although the comparison between Lucentis and Avastin is inconclusive (23). The incidence of side-effects was low and similar for both drugs and dosing regimens, although cardiovascular events that have previously been linked to Lucentis or Avastin were more common among Avastin-treated patients and serious systemic adverse events were non-significantly more common in the Lucentis groups.

A within-trial economic evaluation is being conducted alongside the assessment of clinical outcomes to evaluate costs and cost-effectiveness from an NHS perspective. Preliminary results at one-year are presented here. EQ-5D utility was measured at 0, 3 and 12 months and after any serious adverse event (SAE), although for simplicity, quarterly QALYs for the current analyses are based on linear interpolation between routine measurements and exclude measurements after SAEs. Missing utility data were imputed using multiple imputation (24), which generated 25 imputed datasets. Resource use data were collected using questionnaires completed by nurses at monthly clinic visits. Resources were valued using national unit cost data (22, 25, 26) and microcosting estimates of resource use at clinic visits for monitoring and/or drug administration collected via questionnaires completed by 14 IVAN centres (27). Uncertainty and variability around microcosting estimates was captured by randomly sampling costs from the distribution of clinics providing cost data, with sampling weights determined by the number of NHS patients attending each clinic per week. Avastin was assumed to cost £49/dose, based on the price charged by the commercial supplier used in the trial, while Lucentis cost £742.17 (22). Resource use data and unit costs were combined to estimate quarterly costs of: Avastin/Lucentis; drug administration and ocular monitoring; hospitalisations and ambulatory consultations for SAEs or adverse events (AEs)[a] previously linked with Avastin/Lucentis treatment, and changes in medications licensed for treatment-related AEs. We used Kaplan-Meier sample averaging (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

[a] For clarity, the abbreviation "AEs" is used to describe all adverse events (serious or otherwise) throughout the remainder of this paper.

classed as unrelated to treatment occurred at the same rate in all four study arms when calculating the proportion of patients in each arm who were alive at the start of each quarter. This adapted Kaplan-Meier estimator was multiplied by mean costs and QALYs accrued in each quarter. Results presented here are preliminary as analyses modelling changes in quality of life after SAEs are ongoing.

Since IVAN is a factorial trial, it is important to consider the potential for interactions between factors: i.e. to examine whether the difference in costs, QALYs or net benefit between Avastin and Lucentis is affected by the dosing frequency used. Since IVAN is designed to demonstrate non-inferiority and no difference in efficacy is expected, we did not *a priori* expect any interaction between drug and dosing regimen for QALYs or for the cost of drug administration/monitoring or AEs unless the number of re-treatments differed between Avastin and Lucentis and/or differences in efficacy/safety were observed. In general, interactions rarely occur between factors that have little/no effect on outcomes (3, 9). However, we would expect a large interaction for drug costs, as reducing dosing frequency will have a proportionately greater effect for Lucentis than for the less costly Avastin, as dosing regimen and drug may have multiplicative (rather than additive) effects.

Ignoring the interaction by presenting results for monthly vs. as-needed dosing averaged over both drugs (an at-the-margins analysis) would therefore be misleading and introduce bias (9, 10, 12-14). For example, an "at-the-margins" estimate of the incremental cost of giving monthly rather than as-needed dosing 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 standard errors (SEs) and reduce statistical power (7, 13, 15). Although biased point estimates will distort clinical decision-making based on economic evaluation more than inflated SEs (16, 28), inefficient statistical analysis would nonetheless bias estimates of decision uncertainty, overestimating value of information and biasing cost-effectiveness acceptability curves (CEACs) towards 50% when two treatments are compared and towards 25% when four treatments are under consideration.

3. Analytical methods for dealing with factorial design

The anticipated interaction for drug costs means that conclusions about which drug has lower costs or higher net benefits cannot be made independently of conclusions about dosing regimen. We must therefore treat the four combinations of drug and dosing 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, when *estimating* the costs and QALYs for each treatment arm, we do not necessarily need to allow for interactions in those components of net benefit for which the effect of drug and dosing regimen is additive. By breaking down net benefit into four *components* (QALYs, drug costs, administration/monitoring costs and AE 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. Most analyses of clinical endpoints include only statistically significant interactions. However, this is less appropriate for economic evaluation, where interpretation is based on the magnitude of incremental cost-effectiveness ratios (ICERs), rather than statistical significance, and where the outcomes of interest (costs and QALYs) are often under-powered (29) and more likely to show large interactions (16). As a result, large or qualitative interactions that could change conclusions may not be statistically significant. Our statistical analysis plan therefore stated that we would include all interactions that were *either* statistically significant *or* which changed the ranking of treatment regimens differed with respect to total costs, total QALYs or net benefits at ceiling ratios between £10,000 and £40,000/QALY gained. Since inference is arguably irrelevant to decisions about adoption today (28), conclusions will be based on point estimates not uncertainty; we therefore included those interactions that could change the ranking of treatments, regardless of the statistical significance of interactions or differences between treatments. No interaction terms or indicators of dosing regimen were included for costs and QALYs in the first quarter (Q1) as all patients meet the retreatment criteria at baseline and therefore receive three monthly injections. However, for each component of net benefit,

outcomes in Q2, Q3 and Q4 were analysed in the same way, as there is no logical reason why interactions would exist for some but not all quarters. We therefore estimated the magnitude and statistical significance of the interaction terms for all net benefit components in all quarters using regression (ordinary least squares, OLS, for administration/monitoring costs and drug costs and generalised linear models, GLM, with identity link and gamma family for AE costs and QALYs). Those components for which the interaction was statistically significant at the conventional $\alpha=0.05$ level at ≥ 1 quarter were considered to have important interactions. We then conducted the bootstrapping analyses described below with all combinations of non-significant interactions to identify any other net benefit components where interactions changed treatment rankings.

It is already established that correct analysis of trial-based economic evaluations should allow for correlations between costs and QALYs and their typically skewed distributions (30). This is often done using bootstrapping (31), which is particularly convenient when we need to subdivide costs and QALYs into components that must be analysed with different assumptions about interactions, and is one of the only practical methods that also enables KMSA and combining multiple imputed datasets. Since there was marked imbalance in mean baseline utilities between groups, we adjusted QALYs for baseline utility to avoid bias (32).

We therefore conducted the following steps on each of 1000 bootstrap replicates drawn independently from each of the 25 imputed datasets, giving a total of 25,000 bootstrap replicates; each step was repeated using the original sample for each imputed dataset to get point estimates:

a) OLS regression was conducted separately on QALYs, drug costs, administration/ monitoring costs and AE 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 standard deviations across bootstrap replicates rather than regression output. For Q2-4, explanatory variables included a dummy indicating whether patients were randomised to monthly injections and one indicating drug allocation. For Q1, only the dummy for drug was included in the base case analysis, since all patients received monthly injections. Analyses on quarterly QALYs also controlled for baseline

EQ-5D utility. Each regression was analysed with and without an interaction term equal to drug*dosing regimen.

b) 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, using the number of deaths related to study medication observed in each quarter for each arm and the probability of unrelated deaths averaged across all arms. Due to difficulties with zero cell counts, the impact of excluding interactions from these analyses was not evaluated for this paper.
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.

Two main alternative scenarios were evaluated to assess the impact of changing the assumptions about interactions: excluding all interaction terms; and including all interactions for all components of net benefit, including in Q1. Alterative criteria for identifying important interactions were evaluated in sensitivity analyses.

Analyses were conducted in Stata version 12 (StataCorp, College Station, TX). Analyses with and without interactions were conducted on the same set of bootstrap replicates to eliminate chance variations between the sets of results presented. Costs and QALYs for each imputed dataset were combined using Rubin's rule (33) in Microsoft Excel 2007.

The expected value of perfect information (EVPI) was estimated as the mean maximum net benefit for each bootstrap replicate minus the expected net benefit for the treatment that would be adopted based on point estimates. For simplicity, EVPI calculations excluded evidence from other trials. CEACs were estimated as the proportion of bootstrap replicates (across all 25 datasets) where each treatment maximised net benefits. Base case conclusions assume a ceiling ratio of £20,000/QALY gained (34).

4. Results

Regression analyses on quarterly costs and QALYs suggested that there were no statistically significant interactions between drug and dosing regimens for any outcome other than drug cost, where large super-additive interactions were observed in Q2-4 ($p<0.001$; Table 2). When drug costs were analysed on a logarithmic scale (using GLM with log(link)), this

interaction became non-significant ($p \geq 0.48$) and negligible (≤ 0.28), validating the hypothesis that drug and dosing regimen have multiplicative effects on drug costs. We therefore allowed for interactions when analysing drug cost in Q2-4.

Table 2: Crude mean results for each treatment arm, adjusting for missing data, censoring and deaths unrelated to treatment.

Component of net benefit	Quarter	Crude mean (standard deviation) in each arm†				Interaction (SE)‡
		Monthly Lucentis (N=157¥)	As-needed Lucentis (N=155¥)	Monthly Avastin (N=148¥)	As-needed Avastin (N=146¥)	
QALYs (no BL adjustment)	1	0.206 (0.049)	0.204 (0.046)	0.208 (0.044)	0.208 (0.047)	0.0015 (0.0077)
	2	0.208 (0.049)	0.208 (0.05)	0.209 (0.044)	0.207 (0.052)	-0.0018 (0.0083)
	3	0.207 (0.045)	0.205 (0.048)	0.209 (0.041)	0.205 (0.046)	-0.0007 (0.0078)
	4	0.206 (0.050)	0.203 (0.052)	0.206 (0.047)	0.206 (0.045)	0.0042 (0.0086)
	Total	0.827§	0.820§	0.832§	0.826§	0.0003§
Total QALYs (adjusted for BL)		0.824§	0.829§	0.826§	0.816§	0.0144§
Drug cost	1	£2189 (£184)	£2207 (£118)	£144 (£14)	£145 (£12)	-£18 (£20)
	2	£2113 (£277)	£959 (£858)	£138 (£24)	£78 (£60)	£1094 (£112)*
	3	£2097 (£271)	£1164 (£812)	£137 (£21)	£80 (£56)	£877 (£99)*
	4	£2095 (£309)	£890 (£841)	£136 (£26)	£77 (£58)	£1145 (£119)*
	Total	£8494§	£5220§	£555§	£380§	£3098§
Adminis-tration & monitoring cost	1	£296 (£52)	£297 (£50)	£295 (£53)	£296 (£53)	£0.19 (£11.62)
	2	£251 (£57)	£245 (£79)	£247 (£65)	£239 (£73)	-£1.21 (£14.23)
	3	£250 (£55)	£224 (£71)	£246 (£59)	£228 (£72)	£7.06 (£13.30)
	4	£250 (£58)	£219 (£76)	£245 (£67)	£222 (£74)	£7.61 (£14.47)
	Total	£1047§	£985§	£1033§	£984§	£13.66§
AE cost	1	£25 (£113)	£28 (£71)	£32 (£87)	£15 (£41)	-£20 (£13)
	2	£38 (£140)	£52 (£229)	£106 (£1002)	£24 (£71)	-£97 (£67)
	3	£28 (£100)	£190 (£1906)	£23 (£66)	£51 (£227)	-£134 (£107)
	4	£31 (£83)	£30 (£86)	£50 (£374)	£47 (£256)	-£1 (£37)
	Total	£122§	£299§	£211§	£137§	-£250§
Total cost		£9663§	£6504§	£1799§	£1501§	£2862§
Total net benefit (£20k/QALY ceiling ratio)		£6886§	£9903§	£14,848§	£15,010§	-£2855§

BL, baseline

* Statistically significant interaction ($p<0.05$)

† Mean and standard deviations for each outcome were calculated by multiplying the adjusted Kaplan-Meier probability of being alive at the start of that quarter (averaged across all patients who were alive at the start of each quarter and were uncensored at the end of each quarter) by outcomes for each patient, calculating the mean and standard deviation across these values for each imputed dataset and applying Rubin's Rule to combine results for each dataset.

‡ Interaction terms were estimated using regression on those patients in the trial at the start of each quarter, combining imputed datasets using mim. Drug costs and administration and monitoring costs were estimated using OLS regression, since drug costs followed no standard distribution and administration/monitoring costs were not markedly skewed. Other endpoints were estimated using GLM with gamma family and identity link. QALYs were subtracted from 0.25 prior to analysis since some patients had negative QALYs (due to states worse than death); £100 was added to all AE costs as some patients stopping medication after baseline had negative AE costs.

§ Standard deviations around total QALYs and costs cannot be calculated from this analysis as patient numbers vary due to censoring and deaths. Statistical significance is not shown.

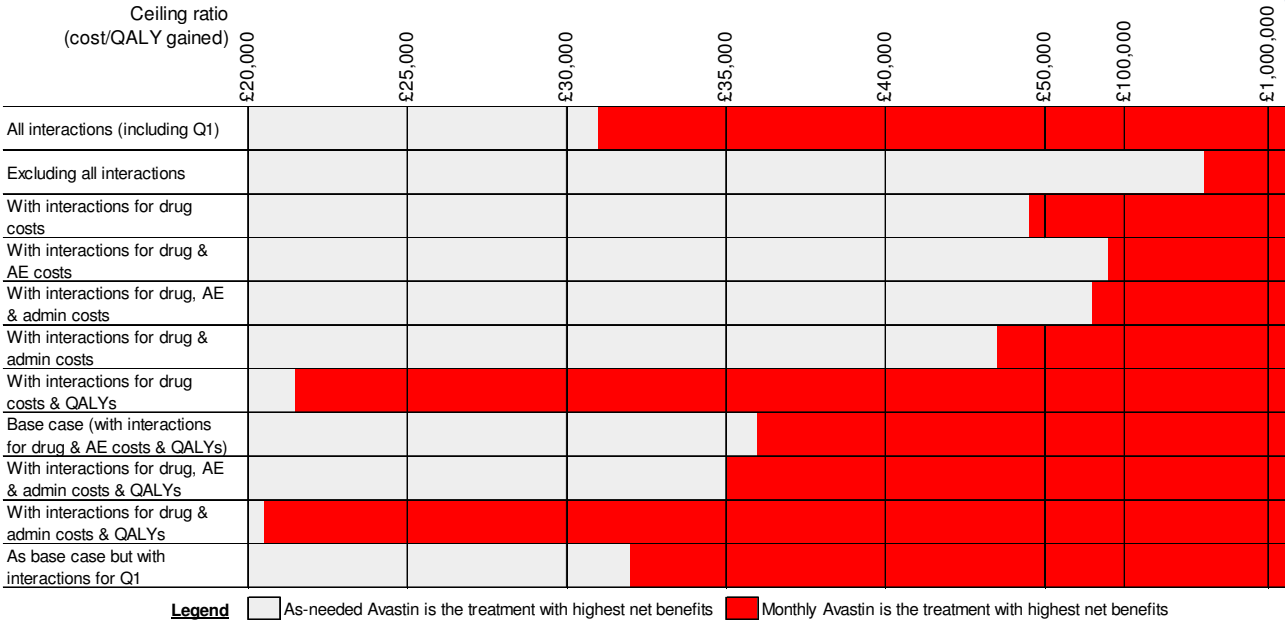
¥ Patient numbers include all patients who were alive at baseline and did not withdraw before visit 2 (regardless of whether they died before visit 2).

Interactions for administration/monitoring costs were negligible and non-significant. A large, qualitative (but non-significant) interaction was also observed for AE cost in Q2 ($p=0.152$), which suggested that using Avastin in place of Lucentis increases AE costs for monthly treatment, but decreases them for as-needed. A marked interaction for AE costs was also

observed in Q3 (p=0.213). Omitting two patients with AE costs above £10,000 reduced the absolute magnitude of both interactions by ≥£96. For QALYs, both the interaction and main effects were very small, suggesting that the choice of drug and dosing regimen had negligible effect on quality of life. Although analysis of crude means (Table 2) suggested that using Avastin in place of Lucentis increased QALYs for both dosing regimens, there was a non-significant qualitative interaction when we adjusted for baseline utility (p≥0.63).

We also estimated mean costs, QALYs and net benefits with all combinations of interactions to identify which interactions changed the rankings of treatments for costs, QALYs or net benefits at ceiling ratios between £10,000 and £40,000/QALY. By definition, allowing for the qualitative interaction for QALYs changed which treatment had highest QALYs (from as-needed Lucentis to monthly Lucentis). Furthermore, allowing for interactions in drug cost and AE cost markedly changed ICERs and affected which treatment had highest net benefit at a wide range of ceiling ratios (Figure 1): for example, as-needed Avastin cost £21,158/QALY compared with monthly Avastin when we allowed for interactions in drug cost and QALYs, but £35,665/QALY when we also allowed for interactions in AE cost.

Figure 1: Impact of including different combinations of interactions on the conclusions about which treatment maximises net benefits.



However, including interactions for one variable sometimes altered the impact of interactions in another variable. This means that although interactions for administration/monitoring costs are very small and generally have no impact on the conclusions, there are several narrow

ceiling ratio ranges where the inclusion of interactions for administration costs alongside other interaction terms changes the conclusions (Figure 1): for example, administration costs may change the conclusions at ceiling ratios between £34,650 and £35,700/QALY *if* we also include interactions for QALYs, drug cost and AE cost, and at ceiling ratios between £20,100 and £21,200/QALY if interactions for AE cost are ignored. However, the implications of this are unclear. Since the interactions around administration costs had less effect than interactions in Q1, the base case analysis therefore allowed for interactions in QALYs, AE costs and drug costs, but assumed that drug and dosing regimen had additive effects on administration/monitoring costs.

The base case results (Table 3) demonstrate that (as expected) monthly Lucentis is the most costly treatment and as-needed Avastin is the least costly. The as-needed Lucentis arm accrued highest mean QALYs. Cost differences between arms were driven largely by drug cost, with no significant differences in QALYs or AE costs ($p \geq 0.09$), although the cost of administration and monitoring was significantly higher for monthly treatment than as-needed ($p < 0.0001$).

Table 3: Base case results, allowing for interactions in QALYs, drug costs and costs of managing expected AEs, but assuming no interactions for administration/monitoring costs.

	Mean (SE) QALYs	**Mean (SE) drug cost**	**Mean (SE) administration & monitoring cost**	**Mean (SE) AE cost**	**Mean (SE) total cost**	**Mean (SE) net benefit at £20k ceiling ratio**
Monthly Lucentis	0.825 (0.011)	£8,503 (£58)	£1044 (£11)	£123 (£18)	£9670 (£66)	£6830 (£222)
As-needed Lucentis	0.829 (0.012)	£5,211 (£162)	£989 (£12)	£297 (£156)	£6497 (£221)	£10,075 (£331)
Monthly Avastin	0.825 (0.011)	£555 (£4)	£1038 (£11)	£202 (£90)	£1795 (£90)	£14,707 (£252)
As-needed Avastin	0.817 (0.012)	£379 (£11)	£980 (£13)	£146 (£30)	£1504 (£36)	£14,835 (£239)
Difference Lucentis vs Avastin	Monthly: 0.000 (0.013) As-needed: 0.012 (0.015)	Monthly: £7948 (£58)* As-needed: £4832 (£162)*	£8 (£14)	Monthly: -£79 (£92) As-needed: £152 (£159)	Monthly: £7875 (£110)* As-needed: £4993 (£223)*	Monthly: -£7877 (£298)* As-needed: -£4759 (£373)*
Difference Monthly vs as-needed	Lucentis: -0.004 (0.013) Avastin: 0.008 (0.013)	Lucentis: £3292 (£171)* Avastin: £176 (£12)*	£56 (£13)*	Lucentis: -£174 (£157) Avastin: £56 (£95)	Lucentis: £3173 (£229)* Avastin: £291 (£96)*	Lucentis: -£3246 (£347)* Avastin: -£128 (£290)
Interaction	-0.012 (0.019)	£3116 (£172)*	N/A	£-230 (£183)	£2882 (£245)*	-£3118 (£452)*

* Difference is significantly different from zero ($p < 0.05$)

Costing results are slightly different from those reported previously (23) due to sampling variation in bootstrapping and sampling from the distribution of administration costs, and as the current results use OLS regression rather than "at the margins" analysis benchmarked on monthly Lucentis and do not control for dosing regimen in Q1. Furthermore, interactions for AE costs were excluded from the costing analysis presented previously (since they had no effect on conclusions for costs), but are included here as they affect conclusions about net benefits when we consider the joint distribution of costs and QALYs.

Although patients receiving monthly Avastin injections accrued slightly (but not significantly) more QALYs than those on as-needed Avastin (p=0.13), as-needed Avastin had highest net benefits at a £20,000/QALY ceiling ratio of all four treatment strategies, with monthly Avastin costing £35,665/QALY gained vs. as-needed Avastin. Monthly Lucentis was strongly dominated by as-needed Lucentis and as-needed Lucentis cost £1.34 million/QALY compared with monthly Avastin and £428,457/QALY vs. as-needed Avastin.

We then evaluated the impact of ignoring all interactions and assuming additive effects for all components of net benefit. This gave quite different results (Table 4). In particular, monthly Lucentis became the treatment with highest QALYs (rather than as-needed Lucentis) and mean drug costs and AE costs were markedly different from the base case analysis. In particular, the interaction for drug cost was so large that OLS with no interaction term predicts that drug costs for as-needed Avastin will be negative, as well as giving unrealistic estimates of the cost of other dosing regimens.

Table 4: Results ignoring all interactions (at the margins analysis)

	Mean (SE) QALYs	Mean (SE) drug cost	Mean (SE) administration & monitoring cost	Mean (SE) AE cost	Mean (SE) total cost	Mean (SE) net benefit at £20k ceiling ratio
Monthly Lucentis	0.827 (0.010)	£7739 (£64)	£1044 (£11)	£180 (£47)	£8963 (£82)	£7581 (£214)
As-needed Lucentis	0.826 (0.011)	£5963 (£125)	£989 (£12)	£242 (£121)	£7194 (£171)	£9335 (£281)
Monthly Avastin	0.823 (0.010)	£1362 (£45)	£1038 (£11)	£143 (£79)	£2542 (£91)	£13,914 (£231)
As-needed Avastin	0.819 (0.011)	-£419 (£45)	£980 (£13)	£205 (£52)	£765 (£69)	£15,620 (£229)
Difference Lucentis vs Avastin	0.006 (0.011)	£6380 (£88)*	£8 (£14)	£37 (£93)	£6425 (£127)*	-£6309 (£251)*
Difference monthly vs as-needed	0.002 (0.009)	£1779 (£89)*	£56 (£13)*	£-62 (£93)	£1773 (£128)*	-£1730 (£228)*
Interaction	N/A	N/A	N/A	N/A	N/A	N/A

* Difference is significantly different from zero (p<0.05)

For QALYs and AE costs, ignoring interactions generates markedly narrower SEs than the base case analysis, which reflects the increased statistical efficiency demonstrated previously (7, 13, 15). However, the trend is less consistent for drug cost, total costs and net benefits due to the extreme heteroskedasticity in drug costs: for example, SEs around drug cost and total cost are smaller than the base case for as-needed Lucentis, but larger for other arms, as drug costs for as-needed treatment are much more variable than for monthly and as those for Avastin vary less than those for Lucentis.

When we include interactions for all types of cost and in all quarters (including interactions for administration/monitoring costs and in Q1 for all endpoints), point estimates are very similar to the base case (Table 5), with as-needed Lucentis accruing highest QALYs and with all outcomes mirroring those in Table 2. This demonstrates that ignoring interactions in Q1 and for administration/monitoring costs introduced relatively little bias into the analysis. However, including all interactions increased SEs for all outcomes other than total drug cost for monthly Lucentis, since all regression analyses include an additional variable that has negligible effect and as treatment effects are calculated using only two cells of the factorial design, not all four.

Table 5: Results including interactions for all components of net benefit in all quarters (inside the table analysis)

	Mean (SE) QALYs	**Mean (SE) drug cost**	**Mean (SE) administration & monitoring cost**	**Mean (SE) AE cost**	**Mean (SE) total cost**	**Mean (SE) net benefit at £20k ceiling ratio**
Monthly Lucentis	0.824 (0.012)	£8494 (£59)	£1047 (£13)	£122 (£19)	£9663 (£68)	£6826 (£234)
As-needed Lucentis	0.829 (0.013)	£5220 (£162)	£985 (£14)	£299 (£157)	£6504 (£222)	£10,079 (£341)
Monthly Avastin	0.826 (0.011)	£555 (£4)	£1033 (£13)	£211 (£90)	£1799 (£90)	£14,719 (£262)
As-needed Avastin	0.816 (0.013)	£380 (£11)	£984 (£15)	£137 (£31)	£1501 (£38)	£14,823 (£252)
Difference Lucentis vs Avastin	Monthly: -0.001 (0.014) As-needed: 0.013 (0.016)	Monthly: £7939 (£59)* As-needed: £4840 (£162)*	Monthly: £15 (£18) As-needed: £1 (£21)	Monthly: -£89 (£92) As-needed: £162 (£159)	Monthly: £7865 (£112)* As-needed: £5003 (£224)*	Monthly: -£7893 (£316)* As-needed: -£4744 (£390)*
Difference monthly vs as-needed	Lucentis: -0.005 (0.015) Avastin: 0.01 (0.015)	Lucentis: £3274 (£172)* Avastin: £175 (£12)*	Lucentis: £62 (£19)* Avastin: £48 (£19)*	Lucentis: -£176 (£158) Avastin: £74 (£95)	Lucentis: £3159 (£231)* Avastin: £298 (£98)*	Lucentis: -£3254 (£380)* Avastin: -£105 (£328)
Interaction	-0.014 (0.021)	£3098 (£172)*	£14 (£27)	-£250 (£184)	£2862 (£250)*	-£3149 (£502)*

* Difference is significantly different from zero ($p<0.05$)

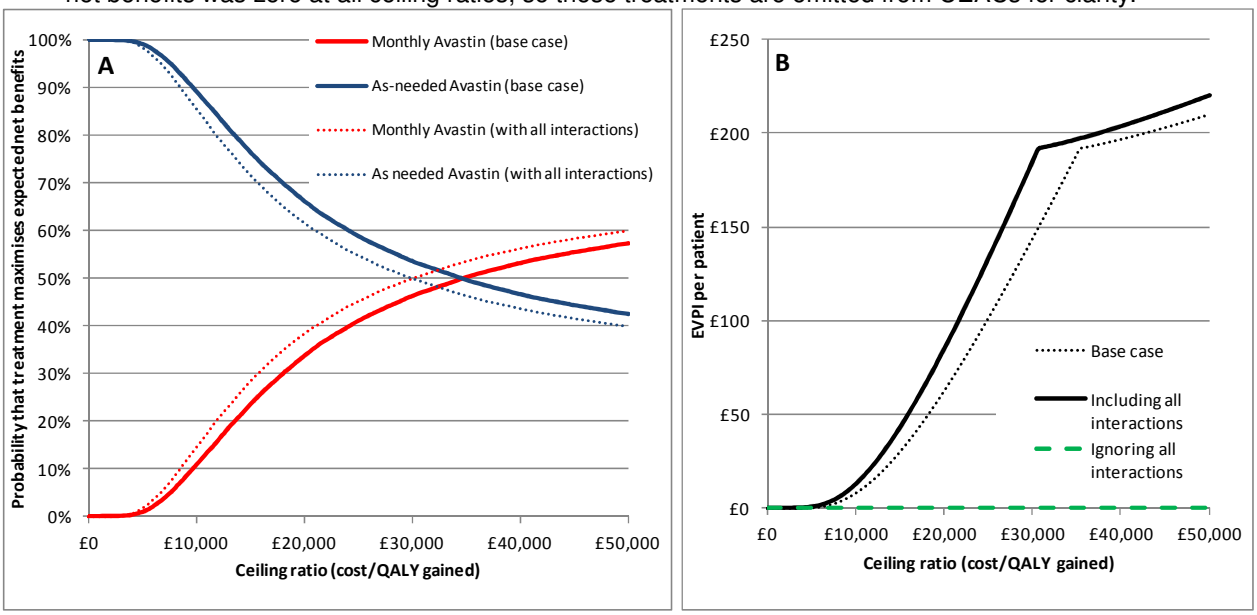
The increase in uncertainty associated with including additional interactions has a substantial effect on CEACs and EVPI. The base case analysis suggests that we can be 66% confident that as-needed Avastin maximises net benefit at a £20,000/QALY ceiling ratio (Table 6, Figure 2) and that the EVPI is £62 per patient. By contrast, if we exclude all interactions, EVPI remains below £0.14 and the probability of any treatment other than as-needed Avastin maximising net benefits remains below 0.1% at all ceiling ratios up to £50,000/QALY gained. When all interactions are included, EVPI is always higher than the base case (Figure

2). *In general*, the probability of each treatment being cost-effective is always closer to 50% when all interactions are included. However, at certain ceiling ratios (e.g. £35,000/QALY), the treatment maximising net benefits differs between the base case and the analysis including all interactions, due to the compound effect of small interactions and differences between dosing regimens in Q1 and for administration/monitoring costs. The probability of either Lucentis regimen maximising net benefits remained below 1% at all ceiling ratios below £147,000 for all analyses.

Table 6: Impact of assumptions about interactions on uncertainty measures

		£20,000/QALY ceiling ratio			£35,000/QALY ceiling ratio		
		Base case	Exclude all inter-actions	Include all inter-actions	Base case	Exclude all inter-actions	Include all inter-actions
EVPI per patient		£62	£0	£85	£188	£0	£197
Probability cost-effective	Monthly Lucentis	0%	0%	0%	0%	0%	0%
	As-needed Lucentis	0%	0%	0%	0%	0%	0%
	Monthly Avastin	34%	0%	38%	50%	0%	54%
	As-needed Avastin	66%	100%	62%	50%	100%	46%

Figure 2: Cost-effectiveness acceptability curves (A) and EVPI (B), considering all four cells of the factorial design as mutually exclusive options. The probability that either Lucentis regimen maximised net benefits was zero at all ceiling ratios, so these treatments are omitted from CEACs for clarity.



However, some of these findings are sensitive to the criteria we used to select which interactions were considered "important" or "genuine". In particular, it is unclear whether interactions for administration/monitoring costs should also have been included, since these affect which treatment maximises net benefits at some ceiling ratios. Including all interactions beyond Q1 would have reduced the ICER for monthly Avastin vs as-needed

Avastin by £1,060 to £34,605/QALY and increased SEs around net benefits by 2-6%. If we had included only those interactions that were statistically significant (at either α=0.05 or α=0.10), as is typically done for analyses of clinical endpoints, we would have included only interactions for drug cost, not for QALYs and AE cost. This would have reduced SEs around net benefit by 4-10% but suggested that monthly Avastin cost £48,594/QALY vs as-needed Avastin, rather than £35,665 as in the base case. However, we would have used the same base case model if our analysis plan were to include all interactions that were either significant or qualitative, or if we had included all interactions with an absolute magnitude larger than ≥1 main effect. Analysing Q1 outcomes in the same way as other quarters increased SEs around net benefit by 3-6% and reduced the ICER for monthly Avastin vs as-needed Avastin to £31,745/QALY. Since the effect of treatment allocation on drug cost is largely multiplicative, we also evaluated the effect of estimating drug costs using GLM with log-link and omitting the interaction term; however, this increased the ICER for monthly Avastin vs as-needed Avastin to £48,855/QALY and increased SEs around total costs and net benefit by up to 18%, in addition to markedly increasing computation time.

5. Discussion

In this particular trial, the assumptions about interactions had a marked effect on conclusions about which treatment maximises QALYs and net benefits and on measures of the decision uncertainty and the value of collecting further information. Analyses including all interactions have been shown to generate unbiased estimates of incremental outcomes (12, 13); as result, the conclusions of the analysis that allows for all interactions may be taken as a "gold standard" against which we can compare mean results for any given analysis. However, in addition to estimating much larger SEs, we show that allowing for unimportant interactions also increases estimates of the probability of adopting the wrong treatment and of the value of collecting further information. Ignoring all interactions suggested there was no decision uncertainty around the study results, as well as giving misleading point estimates. Although it is unclear which analysis gives the "correct" measure of decision uncertainty, we feel that the base case results are most realistic as it would be inappropriate to commission further research where a more efficient statistical analysis could have sufficed.

Tailoring assumptions about interactions to each component of net benefit 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 produced smaller SEs, lower EVPI and (generally) a lower probability of making the wrong decision. Drawing conclusions based on inside-the-table results 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 dosing regimen as though these decisions were independent. In this case study, presenting results inside the table suggested that monthly Lucentis is dominated by as-needed Lucentis, while monthly Avastin could be cost-effective compared with as-needed Avastin if the NHS were willing to pay more than £36,700/QALY gained. However, the implications of these findings for clinical practice are unclear. In particular, most ophthalmology clinics lack the staff and facilities to administer monthly injections and the differences in QALYs, net benefits and visual acuity (23) between monthly and as-needed Avastin are unlikely to be clinically or economically significant. Furthermore, the current results are preliminary and further analysis is ongoing to allow for quality of life measurements taken after SAEs.

However, this 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. Several questions remain about the most appropriate criteria and we would welcome feedback from HESG members on these questions and on what criteria are most appropriate. Firstly, theoretical debate remains about the importance of statistical significance, which is the main (if not only) criteria driving decisions about whether or not to include interactions in clinical analyses of factorial trials (35), although less stringent significance levels (e.g. α=0.10) are often used for interactions than for main effects (36). However, while hypothesis testing remains important for analysis of clinical endpoints, in economic evaluation, conclusions are primarily based on estimation and inference is argued to be irrelevant (28). Furthermore, economic endpoints may be more likely to have large interactions (16) but are underpowered to detect them due to large variances and skewed distributions.

We therefore planned to include interactions that were significant and those that changed conclusions about total costs, QALYs or net benefits. However, qualitative interactions for specific cost components that do not change the conclusions about total costs may also be important and many interactions that do change conclusions at some ceiling ratios could be spurious. In particular, our analysis observed a very large qualitative interaction for AE costs

that changed the conclusions about which treatment had highest AE costs but not which had highest costs overall. However, this interaction appears to be due to chance as the two patients with costs >£10,000 were in diametrically opposite cells of the design (as-needed Lucentis and monthly Avastin) and the interaction became substantially smaller when these patients were omitted. Similarly, the qualitative interaction seen for QALYs changes conclusions about which treatment is most effective, but would not have done if main effects were larger. We also found that even very small interactions in administration/monitoring costs and in Q1 could change the conclusions at a small range of ceiling ratios within the pre-specified range; this raises the question of whether any interaction would have been considered "unimportant" based on a strict application of our criteria. As result, alternative criteria, such as pre-specifying the maximum amount of bias permitted in total costs, QALYs or net benefit may be preferable to avoid judgements about interactions during analysis; however, this raises the additional question of how much bias we should be willing to accept. This issue also links in with the general problem of what conclusions should be drawn from economic evaluations observing small differences that could be due to chance: should we ignore uncertainty and adopt the treatment with highest expected net benefit, or should we conclude that the treatments are roughly equivalent?

It is also unclear whether it is possible (or advisable) to identify in advance which interactions could change the conclusions, without rerunning analyses with all combinations of interactions. In this case study, the results validated most of our hypotheses, although we also found marked interactions in QALYs and AE costs, as well as drug costs, and found that interactions in administration/monitoring costs might also change conclusions in certain circumstances. In practice, however, it is difficult to estimate the impact of interactions in costs and QALYs on ICERs and net benefit; as a result, re-estimating mean results with all combinations of interactions may be the simplest strategy, despite the complications that it introduces into the interpretation of such analyses.

As with all statistical analyses, it is essential to pre-specify the criteria used 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 Bayesian approach of attaching informative (or sceptical) priors to the interaction term (36, 37) provides an alternative to making black and white decisions on which interactions are "in" or "out". This approach could be particularly useful for outcomes such as AE costs where a qualitative interaction appears to have arisen by chance. 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 sub-additive but not qualitative (37)) might be suitable for QALYs and non-drug costs, but not for drug costs.

A further shortcoming of our approach is that it explicitly adjusts group means for assumptions about interactions. As result, the main results (Table 3) artificially remove interactions that are excluded from the analysis; we therefore recommend that unadjusted group means (e.g. Table 2) are also presented wherever possible. Furthermore, we were unable to assess the implications of ignoring interactions in Kaplan-Meier estimators for this paper due to the difficulties introduced by zero cell counts, but hope to explore methods for evaluating this in future work.

Our approach could be applied in other situations where interactions are expected for some types of costs or health benefits but not others. Our analysis uses a single case study that has a non-inferiority design and drug costs as the main cost driver and evaluates two drugs and two dosing 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 (9, 10, 12-14), so researchers should always consider the likely or potential magnitude of interactions in factorial trials and allow for any interactions that could affect the conclusions: particularly as large interactions are likely to be more common for costs and QALYs than for clinical endpoints (16).

6. Acknowledgements

We thank the IVAN study investigators who collected the data for the trial and thank James Raftery for his involvement in discussions on the IVAN statistical analysis. The IVAN trial is funded by the National Institute for Health Research (NIHR) Health Technology Assessment

(HTA) programme (project number 07/36/01) and will be published in full in Health Technology Assessment. Visit the HTA programme website 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.

7. References

1. Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain. 1926; 33: 503-13.
2. Armitage P, Berry G, Mathews JNS. Statistical methods in medical research. 4th ed. Malden, MA: Blackwell Science Ltd., 2002.
3. Byar DP, Piantadosi S. Factorial designs for randomized clinical trials. Cancer Treat Rep. 1985; 69: 1055-63.
4. Finney DJ. Experimental Design and Its Statistical Basis. London: Cambridge University Press 1955.
5. McAlister FA, Straus SE, Sackett DL, et al. Analysis and reporting of factorial trials: a systematic review. JAMA. 2003; 289: 2545-53.
6. Lubsen J, Pocock SJ. Factorial trials in cardiology: pros and cons. Eur Heart J. 1994; 15: 585-8.
7. Fox Z, Nitsch D, Wang D, et al. Chapter 11: Factorial design. In: Wang D, Bakhai A, eds., Clinical trials: a practical guide to design, analysis, and reporting. London, UK: Remedica, 2006.
8. Green S, Liu PY, O'Sullivan J. Factorial design considerations. J Clin Oncol. 2002; 20: 3424-30.
9. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. Stat Med. 1989; 8: 161-71.
10. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol. 2003; 3: 26.
11. Byar DP. Factorial and reciprocal control designs. Stat Med. 1990; 9: 55-63; discussion 63-4.
12. Hung HM, Chi GY, O'Neill RT. Efficacy evaluation for monotherapies in two-by-two factorial trials. Biometrics. 1995; 51: 1483-93.
13. Hung HM. Two-stage tests for studying monotherapy and combination therapy in two-by-two factorial trials. Stat Med. 1993; 12: 645-60.
14. Blyth K, Gebski V. Factorial designs: a graphical aid for choosing study designs accounting for interaction. Clinical trials. 2004; 1: 315-25.
15. 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.
16. 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.
17. Ramsey S, Willke R, Briggs A, et al. Good Research Practices for Cost-Effectiveness Analysis Alongside Clinical Trials: The ISPOR RCT-CEA Task Force Report. Value Health. 2005; 8: 521-33.
18. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ. 2011; 342: d1548.

19. Bunce C, Xing W, Wormald R. Causes of blind and partial sight certifications in England and Wales: April 2007-March 2008. Eye (Lond). 2010; 24: 1692-9.
20. Rosenfeld PJ, Brown DM, Heier JS, et al. Ranibizumab for neovascular age-related macular degeneration. N Engl J Med. 2006; 355: 1419-31.
21. Brown DM, Kaiser PK, Michels M, et al. Ranibizumab versus verteporfin for neovascular age-related macular degeneration. N Engl J Med. 2006; 355: 1432-44.
22. British Medical Association. British National Formulary 62 (http://bnf.org/bnf/index.htm). London: Pharmaceutical Press, September 2011.
23. 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 (in press). 2012.
24. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011; 30: 377-99.
25. Curtis L. Unit Costs of Health and Social Care 2011. Pages 91, 129-150, 177-213. Canterbury, UK: PSSRU Personal Social Services Research Unit, 2011.
26. Department of Health. National Schedule of Reference Costs 2010-11 for NHS Trusts and PCTs combined. London, United Kingdom: Department of Health, November 2011.
27. Abangma G, Dakin H, Wordsworth S. Micro-costing, gross-costing or HRGs: Does the choice of cost approach matter in clinical trials? Podium presentation at the International Health Economics Association (iHEA) 8th World Congress, Toronto, Canada, 10th-13th July 2011. 2011.
28. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. J Health Econ. 1999; 18: 341-64.
29. Briggs A. Economic evaluation and clinical trials: size matters. Bmj. 2000; 321: 1362-3.
30. Willan AR, Briggs AH. Statistical analysis of cost-effectiveness data. Chichester: John Wiley & Sons Ltd, 2006.
31. 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: 327-40.
32. 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: 487-96.
33. Briggs A, Clark T, Wolstenholme J, et al. Missing... presumed at random: cost-analysis of incomplete data. Health Econ. 2003; 12: 377-92.
34. National Institute for Health and Clinical Excellence. Social value judgements: Principles for the development of NICE guidance. Second Edition. (http://www.nice.org.uk/media/C18/30/SVJ2PUBLICATION2008.pdf). 2008.
35. Couper DJ, Hosking JD, Cisler RA, et al. Factorial designs in clinical trials: Options for combination treatment studies. Journal of Studies on Alcohol. 2005; 66: 24-32.
36. Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. Biometrics. 1997; 53: 456-64.
37. Welton NJ, Ades AE, Caldwell DM, et al. 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: 807-41.