

# Economic evaluation of factorial randomised controlled trials: challenges, methods and recommendations

Helen Dakin<sup>\*†</sup>  and Alastair Gray 

Increasing numbers of economic evaluations are conducted alongside randomised controlled trials. Such studies include factorial trials, which randomise patients to different levels of two or more factors and can therefore evaluate the effect of multiple treatments alone and in combination. Factorial trials can provide increased statistical power or assess interactions between treatments, but raise additional challenges for trial-based economic evaluations: interactions may occur more commonly for costs and quality-adjusted life-years (QALYs) than for clinical endpoints; economic endpoints raise challenges for transformation and regression analysis; and both factors must be considered simultaneously to assess which treatment combination represents best value for money. This article aims to examine issues associated with factorial trials that include assessment of costs and/or cost-effectiveness, describe the methods that can be used to analyse such studies and make recommendations for health economists, statisticians and trialists. A hypothetical worked example is used to illustrate the challenges and demonstrate ways in which economic evaluations of factorial trials may be conducted, and how these methods affect the results and conclusions. Ignoring interactions introduces bias that could result in adopting a treatment that does not make best use of healthcare resources, while considering all interactions avoids bias but reduces statistical power. We also introduce the concept of the opportunity cost of ignoring interactions as a measure of the bias introduced by not taking account of all interactions. We conclude by offering recommendations for planning, analysing and reporting economic evaluations based on factorial trials, taking increased analysis costs into account. © 2017 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

**Keywords:** factorial design; randomised controlled trial; guidelines; cost-utility analysis; trial-based economic evaluation

## 1. Introduction

Increasing numbers of economic evaluations are conducted alongside randomised controlled trials (RCTs) [1–3]. Such studies help inform decisions about which intervention provides best value for money.

There is an extensive literature on methods for trial-based economic evaluation [4,5], but currently little guidance on how such methods should be applied to different types of trial. In particular, no research has yet assessed the best methods for conducting economic evaluations alongside factorial trials,<sup>a</sup> which account for around 2–4% of RCTs [9,10].

Factorial RCTs evaluate  $\geq 2$  factors simultaneously, randomising patients to  $\geq 2$  levels of  $\geq 2$  factors, with different groups of subjects being randomly allocated to receive different combinations of levels for each factor. The factors under investigation may comprise the presence or absence of a drug or other

Nuffield Department of Population Health, University of Oxford, U.K.

<sup>\*</sup>Correspondence to: Helen Dakin, Health Economics Research Centre, Nuffield Department of Population Health, Old Road Campus, Headington, Oxford OX3 7LF, U.K.

<sup>†</sup>E-mail: helen.dakin@ndph.ox.ac.uk

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

<sup>a</sup>One of the authors (HD) searched Medline via PubMed for (factorial[Title]) AND ('economic evaluation' OR 'cost-effectiveness'), searched past papers presented to the Health Economists' Study Group (HESG, <http://www.hesg.org.uk>) for the term 'factorial' and searched the British Library Electronic Theses Online Service (ETHOS, <http://ethos.bl.uk>) for the terms 'factorial' and 'cost' in combination. Such searches were conducted on 7 October 2011, 7 August 2013 and 10 February 2016. These searches identified no previous methodological research on factorial trials, besides the programme of work conducted by the authors, applied examples (e.g. [6,7]) and one systematic review [8].

**Table I.** Illustration of a  $2 \times 2$  factorial trial evaluating the effect of two drugs (A and B).

		Factor A: presence/absence of Drug A	
		Level 0: Placebo for A	Level 1: Active drug A
Factor B: presence/absence of Drug B	Level 0: Placebo for B	0: Placebo for A + Placebo for B Mean outcome: $\mu_0$	a: Drug A + Placebo for B Mean outcome: $\mu_a$
	Level 1: Active drug B	b: Placebo A + Drug B Mean outcome: $\mu_b$	ab: Drug A + Drug B Mean outcome: $\mu_{ab}$

healthcare technology (as in Table I),  $\geq 2$  active treatments or any other aspect of dose or treatment administration for which outcomes can be usefully compared.

Factorial designs enable multiple questions to be addressed in the same study [11,12]: for example, we can assess the effectiveness and safety of A and B individually and in combination. In particular, factorial trials allow us to investigate interactions between treatments [13,14]. Interactions indicate that the effect of A differs depending on whether B is given, i.e. that the effect of using A and B in combination is not equal to the sum of the individual effects. Interactions may be super-additive or synergistic if the effect of A + B is greater than the effect of A and B individually, or sub-additive or antagonistic if the effect of A + B is less than the effect of A and B individually. Sub-additive interactions may be qualitative if the effect of one intervention changes sign, not just magnitude, depending on whether or not the other is given.

If there are no interactions, factorial designs provide greater power for the same sample size than three-arm studies evaluating the same treatments and (in theory) could provide the same information as two two-arm studies, but at substantially lower cost [13,15]. By contrast, three-arm studies do not offer this efficiency advantage and do not provide direct estimates of the magnitude of interactions. However, realising these benefits and obtaining consistent, unbiased and statistically efficient estimates of treatment effects depend on how results are analysed and whether or not treatments interact.

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

- Assume zero interaction: At-the-margins analysis (comparing outcomes for all patients receiving A with all those not receiving A) and regression without interaction term(s) are generally statistically efficient [12,13] but implicitly assume no interactions. They therefore produce biased estimates whenever the true interaction is not zero [13,15–17].<sup>b</sup>
- Assume important interactions exist: Inside-the-table analysis<sup>c</sup> (considering each combination of factors as a separate treatment) and regression analyses including interaction term(s) almost always have lower statistical power than at-the-margins analysis [17] but are always unbiased [16,17].

As a compromise between bias and inefficiency, a two-stage testing approach is commonly used [14,17], in which an initial test is done to assess whether statistically significant interactions exist, which determines the type of analysis done in the second step. Factorial designs are generally used either to efficiently evaluate multiple factors if it is believed that there is no interaction [13] or to investigate interactions that are believed to exist [13,22]. However, it is generally not possible to achieve both objectives simultaneously [22], because interactions reduce the efficiency gains possible with factorial trials [12,13] and bias the most efficient estimators [13,15–17], while large samples are needed to detect interactions [22,23]. Although the methods for statistical analysis of clinical endpoints are well established, the implications for economic evaluations have not been explored previously.

The next section discusses the four key challenges that factorial designs raise for researchers conducting economic evaluations alongside factorial RCTs, relating to interactions, data transformation, regression and defining the study question. Section 3 uses a hypothetical worked example to illustrate these challenges and demonstrate methods that can be used to analyse costs and quality-adjusted life-years (QALYs) alongside a factorial RCT. We also demonstrate the impact that different analytical

<sup>b</sup>Bias is a systematic tendency to over- or underestimate the parameter of interest. In this case, analyses excluding interactions systematically overestimate the effect of A compared with not-A in the absence of B by a bias equal to 50% of the true interaction term [17].

<sup>c</sup>Within this article, we use McAlister's term 'inside-the-table' [12] to describe the method of analysing a factorial trial in which each cell or factor-combination is analysed as a separate study arm, with no pooling, but in which regression analysis is not used. This term has been used by several recent factorial trials citing that review (e.g. [18,19]), although the earlier literature tends to refer only to analyses of this type as those estimating 'simple effects' [16,17,20,21] or conducting comparisons between cell means without reference to a specific term [21].

methods can have on the results and conclusions and introduce the concept of the *opportunity cost of ignoring interactions* as a measure of the bias associated with omitting interactions from the analysis. We conclude with some recommendations for health economists, statisticians and trialists when planning, analysing and reporting economic evaluations alongside factorial RCTs.

## 2. Challenges for economic evaluation of factorial trials

### 2.1. Challenge 1: Interactions

It has previously been recognised that interactions may arise due to non-compliance [13] or pharmacokinetic, biological or behavioural mechanisms [13,15,24]. Interactions may also be an artefact of the scale of analysis [13,14]: if two factors have a multiplicative effect, whereby the effect of A and B in combination is equal to the product (not the sum) of the individual effects of A and B, there will be an interaction on a natural scale, but not on a logarithmic scale.<sup>d</sup> Logistic regression is therefore often used to allow for treatments having a multiplicative effect on the incidence of clinical events. However, the incremental costs and QALYs associated with preventing clinical events must be interpreted on a natural scale (see point (5) below), where treatment effects will be non-additive (see Section 3.1). Many interventions also have non-additive effects on quality or length of life: particularly because diminishing marginal effects are built into many utility measures, such as EQ-5D [26] and the Health Utilities Index [27].

These mechanisms are likely to produce interactions for costs and QALYs even when there are no interactions for clinical endpoints. This has also been observed for interactions between treatment and subgroup/country, where incremental costs and QALYs often differ between subgroups or between countries even in trials that find treatment effects to be consistent across subgroups/countries for clinical endpoints [3]. Factorial designs are also often chosen when no interaction is expected for the primary clinical endpoint [12,13,28], although the likelihood of interactions for economic endpoints is rarely considered. This selection effect may reduce the number of factorial trials with interactions for clinical endpoints [12], but not the incidence of interactions for economic outcomes, further increasing the chance that there are interactions for economic endpoints in studies that have additive clinical effects.

As illustrated in Section 3.1, the combination of super-additive interactions for cost and sub-additive interactions for QALYs may also produce much larger sub-additive interactions in measures of cost-effectiveness, such as net monetary benefit ( $NMB = QALYs \cdot Rc - Cost$ , where  $Rc$  represents the ‘ceiling ratio’ indicating the amount a health system is willing or able to spend to gain one QALY). Indeed, all studies that observe non-zero interactions for costs or QALYs and find  $\geq 1$  treatment to be more costly and more effective than its comparator will observe qualitative interactions for NMB at some ceiling ratio.<sup>e</sup> These interactions are likely to introduce substantial bias into analyses that assume no interaction between interventions and could distort the conclusions.

### 2.2. Challenge 2: Difficulties with transformation

Because the strength of interactions depends on the scale of measurement, transformation is often used to eliminate or reduce interactions in clinical endpoints of factorial trials [13,14,28], as well as to normalise non-Gaussian distributions. For example, if treatment effects are multiplicative (i.e. interaction = 0 on a logarithmic scale), logistic regression or log-transformation are often used to analyse results and draw statistical inferences. However, analysis of transformed data raises additional problems for economic evaluation that are not raised by analysis of clinical endpoints:

<sup>d</sup>For example, if smoking increases the risk of dying of lung cancer by 36-fold, while a 100 Bq/m<sup>3</sup> increase in residential radon exposure increases the risk by 16% [25], there will be no interaction between smoking and radon on a logarithmic scale (e.g. in logistic regression), but raising radon levels will increase the number of lung cancer deaths from 0.42% [25] to 0.49% in non-smokers, but from 15.12% to 17.54% in smokers: an interaction of 2.35% (0.42% + 17.54%–0.49%–15.12%).

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

- 1 The correct transformation may not be known (or knowable) [3].
- 2 Transformation will not eliminate qualitative interactions [14,28]. Monotonic transformations (e.g. taking the logarithm, square-root or power of all values) can increase or decrease  $\mu_{ab} - \mu_b$  and  $\mu_a - \mu_0$  but will not cause them to have the same sign. In principle, non-monotonic transformations (e.g. trigonometric functions) could be used, but these are unlikely to reflect realistic data-generating mechanisms and would only eliminate qualitative interactions in very specific circumstances.
- 3 Total costs, total QALYs and NMB are the weighted sum of several components (e.g. the cost of drugs, clinical events and side-effects [29]), and treatments may have additive effects on some components (e.g. drug cost) and multiplicative effects on others (e.g. the cost of treating clinical events). Analysing total costs on a natural scale will therefore give an interaction arising from the multiplicative effect of events, while analysing total costs on a logarithmic scale will give an interaction arising from additive drug costs. In some cases, it may be possible to separate the components that are multiplicative from those that are additive so that we can, for example, analyse log-transformed event cost with one model and non-transformed drug costs with another. However, in many cases (e.g. where treatment has both additive and multiplicative effects on quality of life), it may not be possible to separate costs and QALYs into components amenable to different types of transformation.
- 4 Frequently, many trial participants have zero cost [3,30,31]. However, Box–Cox transformations (such as logarithms) can only be done on non-zero values [3,30]. Arbitrary constants are often added to all zero values to get around this problem, although the magnitude of such constants can affect results [3].
- 5 Resource allocation decisions must be based on incremental costs and QALYs measured on a *natural* scale [31]. Health gains from a finite healthcare budget are maximised by adopting those treatments with incremental cost-effectiveness ratios (ICERs) below a ceiling ratio that represents the opportunity cost of the healthcare activities that would be displaced by the new treatment [2,32,33]. Setting the ceiling ratio using a league table [2,33] necessarily requires *adding* the total cost of each treatment to the total already spent. The ICER must therefore be calculated as the *absolute* difference in cost divided by the *absolute* difference in effect. Decision-making on a log-scale, based on a measure of relative difference in cost divided by relative difference in effect, would fail to meet this objective. If analyses are conducted on a non-natural scale, it is therefore necessary to back-transform results so that conclusions can be based on inside-the-table means on a natural scale.
- 6 Transforming costs and QALYs prior to analysis means that coefficients, statistical inference and predictions are on the transformed scale and are not applicable to the natural scale [3,31,34]. Furthermore, estimates based on data subjected to non-linear transformations cannot be returned to a natural scale by simply inverting the transformation [30]. Instead, more complex back-transformation methods are needed, such as smearing estimators [3,31] or Taylor series approximation [31], which both require assumptions about the distribution or heteroskedasticity [3,31].

Issues 4–6 can be avoided using generalised linear models (GLM), which predict a function of the dependent variable [3,31]: for example, GLM with log-link could estimate linear effects of A and B on the natural logarithm of mean cost (in contrast to modelling the effect of each treatment on the mean of log-cost) [3]. As a result, GLM enables analysis of zero values [3], inferences that are valid on a natural scale [3] and simple back-transformation of the output [3,31].

### 2.3. Challenge 3: Difficulties with regression analysis

Regression analysis is generally the most convenient way to correctly analyse factorial trials and estimate the magnitude and statistical significance of interaction terms [15]. Regression also facilitates exploration of heterogeneity [31,35,36] and adjustments for between-centre effects/clustering [31,36] and imbalance in baseline characteristics [31,37]. However, economic evaluation of factorial trials raises several challenges for regression analysis.

First, more complex regression techniques (e.g. GLM or two-part models [3,31]) are frequently required to deal with skewed or kurtotic cost and QALY data, the frequently high proportion of zeros in cost data [3,30,31] and the unusual distribution of EQ-5D utilities [38,39].



Second, guidelines for economic evaluation recommend presenting uncertainty using cost-effectiveness acceptability curves (CEACs) and the value of information [4]. For two-arm trials, CEACs can be based on one-sided  $p$ -values from regression [35], while both the value of information and CEACs can be estimated from the mean incremental NMB and its standard error (SE) by assuming a normal distribution [40–42]. However, methods (other than Markov chain Monte Carlo, MCMC [43]) to estimate CEACs or value of information from regression-based cost-effectiveness analysis on  $>2$  alternatives are less well developed.

Third, the properties of ICERs make them unsuitable for use in regression analysis [2,44,45], particularly for studies evaluating  $>2$  interventions, where it is necessary to calculate ICERs relative to the next most effective non-dominated comparator [2]. Net monetary benefit is therefore frequently used in regression analyses as a measure of cost-effectiveness. Because the ceiling ratio ( $R_c$ ) representing the amount a health system is willing or able to pay to gain one QALY is unknown, researchers generally present results at multiple ceiling ratios [31,35,46]. However, conducting regression analyses separately at each ceiling ratio is problematic as the distribution of NMB, the covariates that affect it and the appropriate scale of analysis may vary with the ceiling ratio [35,46]. In factorial trials, the importance of interactions and the appropriate scale of analysis may also differ between costs and QALYs and therefore with ceiling ratio: for example, if there is a genuine interaction for QALYs, but treatments have additive effects on costs, omitting the interaction term may be appropriate at  $R_c = \text{£}0$  (which is equivalent to a regression on cost), but interactions will become increasingly important at higher ceiling ratios. Furthermore, if treatments have multiplicative effects on QALYs, but additive effects on cost, the appropriate functional form could be a linear model at  $R_c = \text{£}0$ , but GLM with log-link at  $R_c = \infty$  (which is equivalent to a regression on QALYs).

Conducting one regression analysis on costs and another on health outcomes avoids the need to replicate models at multiple ceiling ratios and enables model specification to differ between costs and effects [31,46,47]. However, such bivariate analyses must allow for correlations between costs and effects [36,47,48].

#### 2.4. Challenge 4: Framing the study question

For clinical endpoints of factorial trials, conclusions are often drawn independently for different factors. However, this approach may not be appropriate for economic evaluation, where the decision rules used to interpret results depend on whether treatments are considered to be independent or mutually exclusive.

When interventions are considered to be ‘independent’, the ICERs for these treatments (each calculated relative to their next best non-dominated alternative) can simply be compared side-by-side, or used in a league table [2,32,33], with all treatments having ICERs below the ceiling ratio being adopted simultaneously.

Conversely, given a set of ‘mutually exclusive’ alternatives, we must identify and exclude from consideration any treatments that are strongly or weakly dominated by others and calculate the ICER for each of the remaining options compared with its next best non-dominated alternative [2,32,33]. We can then adopt the single treatment that lies on the cost-effectiveness frontier (i.e. is not dominated) and has an ICER below our ceiling ratio.

Although there is ambiguity in the literature, many textbooks and reviews which describe decision rules specifically refer to non-additive effects as the defining feature that determines whether treatments should be considered mutually exclusive or independent [32,33,49–55]. This suggests that whenever treatments interact we should treat the combinations of treatments (e.g.  $0$ ,  $a$ ,  $b$  and  $ab$ ) as mutually exclusive alternatives, comparing the treatment combinations incrementally and selecting the single strategy with highest NMB. If there is no interaction, the interventions can be treated as independent alternatives and their ICERs compared separately against the ceiling ratio. It follows that whenever there is an interaction, economic evaluations of factorial trials should aim to identify which of the mutually exclusive treatment-combinations (e.g.  $0$ ,  $a$ ,  $b$  and  $ab$ ) is most cost-effective, rather than making independent decisions about each factor in isolation.

As well as making different assumptions about interactions, different methods for analysing factorial trials lend themselves to different decision rules for economic evaluation. At-the-margins analysis implicitly treats the factors as independent options, calculating separate ICERs for A versus not-A and for B versus not-B that are independently compared against the ceiling ratio and used to make separate decisions on the two treatments. However, even if decisions are made independently on each

factor, the decision to adopt both A and B in the same patients implicitly means that  $ab$  is recommended. By contrast, inside-the-table analysis treats the four cells as mutually exclusive options, estimating costs, QALYs and NMB for each of the four options and allowing us to identify the cost-effectiveness frontier and select the option that has highest NMB at our ceiling ratio.

Indeed, as shown in Supporting Information 1, treating the cells of a factorial trial as mutually exclusive alternatives and considering interactions between factors will always enable us to maximise the health gains from the budget regardless of whether treatments have additive effects. Analyses allowing for interactions give unbiased estimates of the difference between  $0$  and  $a$  [16,17] and therefore unbiased estimates of the expected NMB of each treatment. By implication, this means that allowing for interactions and considering the four cells as mutually exclusive alternatives must give the correct ranking of cells by NMB and therefore correctly identify the treatment with highest NMB. For treatments with perfectly additive effects on both costs and QALYs, inside-the-table analysis informing a joint decision between mutually exclusive alternatives and at-the-margins analysis informing independent decisions would give identical results and both approaches would maximise expected NMB.

By contrast, excluding interactions will bias NMB estimates whenever treatment effects are not genuinely additive [13,15–17]. Non-qualitative interactions in NMB will change the magnitude of differences between treatments, while qualitative interactions will change the ranking of treatments. Furthermore, qualitative interactions that change which treatment has highest expected NMB will change the conclusions about which treatment is best value for money, and ignoring such interactions will cause us to adopt a treatment that is not the best use of healthcare resources. If there are no qualitative interactions for NMB, ignoring interactions will give the same conclusions as allowing for interactions and making a joint decision between the four mutually exclusive alternatives. When there is evidence that interactions are negligible, it may therefore be appropriate to omit some or all interaction terms when analysing trial outcomes to maximise power and avoid the risk that a small, chance interaction drives the conclusions.

### 3. Methods for the worked example

#### 3.1. Data

To illustrate the four challenges identified above, a simple model was used to simulate the data that might be obtained from a  $2 \times 2$  factorial RCT of hypertension drugs A and B. Model assumptions and data inputs are reported in Supporting Information 2 and the data are provided [56]. In the absence of events and treatment, patients accrued no costs and an average of 25 QALYs. However, 0.23% of untreated patients experience clinical events, such as stroke, each year (odds = 0.30). Both drugs reduce the odds of events, with zero interaction between drugs on a log-odds scale. Each event increases healthcare costs and reduces QALYs by an amount independent of treatment. Treatment side-effects also reduce patients' quality of life. Microsoft Excel 2003 was used to generate 250 hypothetical patients within each treatment arm.

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

**Table II.** Group means and standard deviations (SD) for the worked example.

	$0$ : Placebo for A + placebo for B ( $n = 250$ )	$a$ : Drug A + placebo for B ( $n = 250$ )	$b$ : Drug B + placebo for A ( $n = 250$ )	$ab$ : Drug A + Drug B ( $n = 250$ )	Interaction*
Mean (SD) no. events per patient	7.2 (2.3)	5.9 (2.4)	5.3 (2.1)	4.5 (2.1)	0.5 (sub-additive)
Mean (SD) cost per patient	£87 804 (£33 508)	£98 324 (£32 408)	£109 109 (£28 851)	£125 015 (£29 958)	£5386 (super-additive)
Mean (SD) QALYs per patient	18.1 (6.6)	18.9 (5.9)	19.6 (6.0)	19.8 (5.6)	−0.7 (sub-additive)
Mean (SD) total	£455 010	£470 155	£479 504	£468 985	−£25 664
NMB per patient at £30 000/QALY ceiling ratio	(£211 746)	(£186 507)	(£188 881)	(£174 617)	(qualitative)

\*Interaction =  $\mu_{ab} - \mu_a - \mu_b + \mu_0$ , where  $\mu_k$  indicates the mean outcome in group  $k$ .

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

### 3.2. Analysis

The simulated trial data were analysed in Stata versions 11 and 12 (StataCorp LP, College Station, Texas) and WinBUGS (The BUGS Project, Cambridge, UK [57]) using at-the-margins, inside-the-table and regression analyses. Stata and WinBUGS code are provided in Supporting Information 2.

Ordinary least squares (OLS) and GLM were estimated on costs and QALYs separately, using the `suest` command in Stata [58] to allow for correlations between costs and QALYs in seemingly unrelated regression. An additional analysis used bootstrapping to generate an empirical distribution of costs and QALYs, by sampling with replacement 1000 times (stratified by treatment arm) and running OLS models predicting cost and QALYs with and without interactions on each replicate. Bootstrapping results were used to plot CEACs [59]; the expected value of perfect information (EVPI) [60] was also estimated (Supporting Information 2). Markov chain Monte Carlo was also used to evaluate the impact of using an informative prior on the interaction, following the methods used by Welton *et al.* [7] to allow for scepticism or prior beliefs about the magnitude of the interaction between drugs (Supporting Information 2).

### 3.3. Effect of analytical approach on economic evaluation results

The at-the-margins approach treats the factorial trial as though it were two overlapping two-arm RCTs [12] and effectively ignores the factorial design [61]. Results are typically presented separately for each factor (e.g. for A versus not-A and for B versus not-B), rather than for the individual cells (Table III) [24]. Factor means are generally calculated by pooling cells *a* and *ab* (and similarly cells *b* and *ab*) together and calculating the mean and SD across the pooled study groups, with treatment effects being calculated as the difference between factor means [12]. At-the-margins analysis therefore estimates the average treatment effect across the study population, weighted by the number of patients assigned to each treatment-combination.

In this example, at-the-margins analysis suggests that prescribing A (with/without B) costs £25 530/QALY gained versus not-A, and, independently suggests that B costs £20 189/QALY gained versus not-B (Table III). Based on a £30 000/QALY ceiling ratio, we might therefore conclude that both

**Table III.** At-the-margins results.

	Treatment A		Treatment B	
	Placebo ( <i>n</i> = 500)	Active drug A ( <i>n</i> = 500)	Placebo ( <i>n</i> = 500)	Active drug B ( <i>n</i> = 500)
Mean cost per patient (SD)	£98 456 (£33 005)*	£111 669 (£33 917)*	£93 064 (£33 348)*	£117 062 (£30 439)*
Difference in cost (SE)	£13 213 (£2116)*		£23 998 (£2019)*	
Mean QALYs per patient (SD)	18.9 (6.4)*	19.4 (5.8)*	18.5 (6.3)*	19.7 (5.8)*
Difference in QALYs (SE)	0.52 (0.38)		1.19 (0.38)*	
Mean total NMB per patient at Rc = £30 000 (SD)	£467 257 (£200 813)*	£469 570 (£180 480)*	£462 582 (£199 470)*	£474 245 (£181 783)*
Incremental NMB at Rc = £30 000 (SE)	£2313 (£12 075)		£11 662 (£12 069)	
Incremental cost/QALY	£25 530		£20 189	

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

A and B are cost-effective treatments with positive incremental NMB and recommend that both should be adopted (i.e. implicitly, we should use *ab*). However, this inference would be incorrect due to the qualitative interaction for NMB and the bias inherent within at-the-margins analysis.

Costs and QALYs were then analysed using seemingly unrelated regression. Ordinary least squares was used for simplicity although costs, QALYs and NMB were positively skewed ( $p \leq 0.001$ ) and variances were lower for patients receiving B; GLM is explored below. Regression without interaction terms replicated the mean at-the-margins treatment effects (Table IV), although SEs differed. However, prediction of group means and their SEs (Table IV, Supporting Information 2, Figure A1A) is easier after regression than at-the-margins analysis. These predictions can be used to consider the cost-effectiveness of the four cells as mutually exclusive options, draw conclusions about dominance and identify which of the four options maximises NMB, conditional on the model used. In this example, the predicted values from regression suggest that *a* is extendedly dominated by a combination of *b* and *0* (being more costly and less effective) and explicitly predict that *ab* is the optimal treatment at a £30 000/QALY ceiling ratio, costing £25 530/QALY gained compared with *b*. However, these predictions rely on the same assumptions as at-the-margins analysis and will be biased if interactions are present.

Repeating the regression analysis with an interaction term (Table V) demonstrates that there are in fact large interactions for costs (£5386) and QALYs (−0.68) and a qualitative interaction for NMB (−£25 664), which introduce substantial bias into at-the-margins estimates and those from regression without an interaction term. None of the interactions were statistically significant ( $p \geq 0.172$ ); given that most two-arm trials are not powered to detect significant differences in mean costs or NMB [4,62] and the variance around the interaction term is four-fold higher than that for main effects [17,23], this finding is likely to be common among economic evaluations of factorial trials.

Inside-the-table analysis treats each cell within the factorial design as a separate treatment [12] and estimates simple effects showing differences between individual cell means. For a  $2 \times 2$  factorial design, outcomes are therefore estimated and presented separately for each of the four treatment arms, based only on those patients randomised to that combination of treatments (Table VI). This allows the reader to see the effect of interactions directly [24] and avoids pooling cells. Inside-the-table analysis gives identical point estimates to regression with interaction terms (Table V), although the SEs differ due to heteroskedasticity. Both analyses allowing for interactions find that although *a* is cost-effective versus *0* (£12 297/QALY gained) and *b* is cost-effective versus *a* (£16 070/QALY gained), *ab* costs £88 573/QALY gained versus *b*. On that basis, *b* would maximise NMB at a £30 000/QALY ceiling ratio (not *ab* as at-the-margins analysis suggested). As discussed in Section 2 and Supporting

**Table IV.** Results of OLS regression without interaction term.

	Total cost/ patient	Total QALYs/ patient	NMB/ patient <sup>†</sup>	Cost/QALY		
				versus 0	versus <i>a</i>	versus <i>b</i>
Treatment effect for A (SE)	£13 213 (£1974)*	0.52 (0.38)	£2313 (£12 075)	—	—	—
Treatment effect for B (SE)	£23 998 (£1974)*	1.19 (0.38)*	£11 662 (£12 075)	—	—	—
Constant term (SE)	£86 457 (£1794)*	18.26 (0.35)*	£461 426 (£10 457)*	—	—	—
Predicted mean outcome (SE)						
0: Placebo for A + placebo for B ( <i>n</i> = 250)	£86 457 (£1794)*	18.3 (0.35)*	£461 426 (£10 457)*	—	—	—
<i>a</i> : Drug A + placebo for B ( <i>n</i> = 250)	£99 670 (£1753)*	18.8 (0.33)*	£463 739 (£10 457)*	£25 530	—	—
<i>b</i> : Drug B + placebo for A ( <i>n</i> = 250)	£110 456 (£1624)*	19.5 (0.33)*	£473 088 (£10 457)*	£20 189	£16 070	—
<i>ab</i> : Drug A + Drug B ( <i>n</i> = 250)	£123 669 (£1664)*	20.0 (0.31)*	£475 402 (£10 457)*	£21 809	£20 189	£25 530

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

<sup>†</sup>Based on a ceiling ratio of £30 000/QALY.



**Table V.** Results of OLS regression with an interaction term.

		Total cost/ patient	Total QALYs/ patient	NMB/ patient <sup>†</sup>	Cost/QALY		
					versus <i>0</i>	versus <i>a</i>	versus <i>b</i>
Treatment effect for A (SE)		£10 520 (£2944)*	0.86 (0.56)	£15 145 (£17 076)	—	—	—
Treatment effect for B (SE)		£21 305 (£2792)*	1.53 (0.57)*	£24 494 (£17 076)	—	—	—
Interaction (SE)		£5386 (£3945)	−0.68 (0.77)	−£25 664 (£24 149)	—	—	—
Constant term (SE)		£87 804 (£2116)*	18.09 (0.42)*	£455 010 (£12 074)*	—	—	—
Predicted mean outcome (SE)	<i>0</i> : Placebo for A + placebo for B ( <i>n</i> = 250)	£87 804 (£2116)*	18.1 (0.42)*	£455 010 (£12 074)*	—	—	—
	<i>a</i> : Drug A + placebo for B ( <i>n</i> = 250)	£98 324 (£2047)*	18.9 (0.37)*	£470 155 (£12 074)*	£12 297	—	—
	<i>b</i> : Drug B + placebo for A ( <i>n</i> = 250)	£109 109 (£1822)*	19.6 (0.38)*	£479 504 (£12 074)*	£13 956	£16 070	—
	<i>ab</i> : Drug A + Drug B ( <i>n</i> = 250)	£125 015 (£1892)*	19.8 (0.35)*	£468 985 (£12 074)*	£21 809	£31 375	£88 573

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

<sup>†</sup>Based on a ceiling ratio of £30 000/QALY.

Information 1, the correct conclusions are those from analyses allowing for interactions, which avoid bias from omitted interaction terms.

However, including the interaction term also substantially increased SEs around treatment effects and predicted costs and benefits for each group.<sup>f</sup> This affects statistical inference, CEACs and EVPI.

Uncertainty around at-the-margins analyses interpreting the trial as two independent questions (one on each factor) is sometimes displayed as pairwise CEACs (Supporting Information 2, Figure A2). However, the uncertainty around the entire decision between four mutually exclusive cells is more accurately shown by CEACs showing the proportion of bootstrap replicates where each of the four treatment-combinations had highest NMB. These CEACs for multiple comparisons differed markedly between analyses with (Figure 1A) and without (Figure 1B) interactions. In particular, the probability of any given cell being cost-effective was always closer to 25% (even-odds across the four treatments) when interactions were considered, except at ceiling ratios where the treatment maximising NMB differs between analyses; this appears to be a general finding.

Regression with interaction terms generally produced much higher EVPI estimates than analyses excluding interactions, except for the range of ceiling ratios at which the treatment having highest expected NMB differs between the two analyses (Supporting Information 2, Figure A3).

Repeating results using GLM with log-link and gamma family reduced the size of the interaction for cost compared with OLS: *ab* had 2% (£2833) higher costs than would be expected from the effect of A and B individually, versus £5386 in OLS. However, GLM for QALYs gave a −4% interaction, which is equivalent to −0.748 on a natural scale: somewhat larger than the −0.676 interaction in OLS. This highlights the difficulties associated with using transformation to eliminate interactions composite endpoints such as costs and QALYs: even when interactions arose due to multiplicative effects on the odds of clinical events.

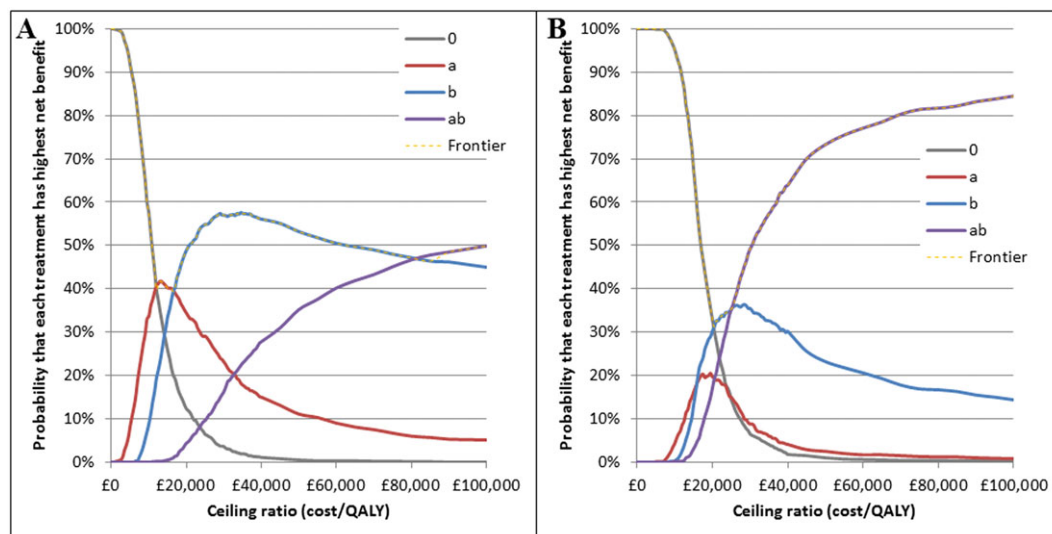
We also conducted seemingly unrelated OLS on NMB at different ceiling ratios. This showed that the interaction increased linearly with ceiling ratio, while its SE increased at a faster rate, such that the *p*-value on the interaction increased from 0.147 to 0.306 as the ceiling ratio increased from £5000 to £40 000/QALY.

<sup>f</sup>Although including interactions increases SEs, *p*-values may either increase or decrease, depending on the magnitude and direction of changes in the point estimate. For example, the two-sided *p*-value for the difference in QALYs between *a* and *0* was 0.128 with interactions and 0.177 without, because allowing for interactions also increased the difference between *a* and *0* from 0.518 (SE: 0.383) to 0.856 (SE: 0.563). By contrast, the two-sided *p*-value for the difference in QALYs between *b* and *0* was 0.007 with interactions and 0.002 without.

**Table VI.** Inside-the-table results.

	$\theta$ : Placebo for A + placebo for B ( $n = 250$ )	$a$ : Drug A + placebo for B ( $n = 250$ )	$b$ : Drug B + placebo for A ( $n = 250$ )	$ab$ : Drug A + Drug B ( $n = 250$ )
Mean cost per patient (SD)	£87 804 (£33 508)*	£98 324 (£32 408)*	£109 109 (£28 851)*	£125 015 (£29 958)*
Incremental cost per patient versus $\theta$ (SE)	N/A	£10 520 (£2948)*	£21 305 (£2797)*	£37 211 (£2843)*
Interaction: cost (SE)		£5386 (£3951)		
Mean QALYs per patient (SD)	18.1 (6.6)*	18.9 (5.9)*	19.6 (6.0)*	19.8 (5.6)*
Incremental total QALYs per patient versus $\theta$ (SE)	N/A	0.86 (0.56)	1.53 (0.57)*	1.71 (0.55)*
Interaction: QALYs (SE)		−0.68 (0.77)		
Mean total NMB per patient at $R_c = £30\,000$ (SD)	£455 010 (£211 746)*	£470 155 (£186 507)*	£479 504 (£188 881)*	£468 985 (£174 617)*
Incremental NMB versus $\theta$ at $R_c = £30\,000$ (SE)	N/A	£15 145 (£17 846)	£24 494 (£17 946)	£13 976 (£17 358)
Interaction: NMB at $R_c = £30\,000$ (SE)		−£25 664 (£24 149)		
Cost/QALY versus $\theta$	—	£12 297	£13 956	£21 809
Cost/QALY versus $a$	—	—	£16 070	£31 375
Cost/QALY versus $b$	—	—	—	£88 574

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



**Figure 1.** CEACs for multiple comparisons using **A** regression with interaction term and **B** regression without an interaction term. The lines for  $\theta$ ,  $a$ ,  $b$  and  $ab$  show the proportion of bootstrap replicates where each of the four treatment-combinations had highest NMB. The dotted line that generally follows the top-most curve shows the cost-effectiveness frontier, i.e. the probability that the treatment with highest expected NMB is cost-effective.

Including an informative prior on the interaction term within MCMC reduced the absolute magnitude of the interactions for both costs and QALYs and also changed the size of all other coefficients (Supporting Information 2, Table A2). This subjective prior also halved the size of SEs around the interaction term (as well as substantially reducing SEs for other coefficients), which may suggest that the type of prior proposed by Welton *et al.* [7] is highly informative and is given a large weight in the analysis.

### 3.4. Opportunity cost of ignoring interactions

The expected value of stratified decision-making has previously been proposed to quantify the benefits of allowing for heterogeneity in economic evaluation and making separate decisions for different subgroups [63,64]. The value of stratification (or the value of considering heterogeneity) is defined as the NMB gained from adopting treatment in only those subgroups for which it has highest expected NMB, minus the NMB of adopting treatment in the whole population [63,64]. The static value of stratification ( $\Delta_s NMB$ ) reflects the benefits of stratification under current information [63,64] (i.e. based on the trials that have been conducted to date and informed the analysis in question) and equals the expected NMB achieved by adopting the best treatment for each subgroup, minus the expected NMB from adopting the treatment that has highest NMB when averaged across all subgroups. Similarly, the dynamic value of stratification or heterogeneity reflects the NMB gained under perfect information (i.e. if sufficient evidence was collected to eliminate all uncertainty about which treatment had highest NMB, for example by conducting an infinitely large RCT). It equals the EVPI with stratification minus the EVPI without stratification.

Factorial RCTs are analogous to subgroup analyses but enable stronger inferences as patients are randomised to both factors. For factorial trials, stratification means that we consider the cells as separate, mutually exclusive treatments and adopt the treatment-combination that maximises expected NMB, rather than considering factors independently.

We propose the *opportunity cost of ignoring interactions* (OCII) as a measure (analogous to the value of stratification) of the NMB lost from assuming additive effects and making separate decisions on the two factors, rather than taking account of interactions and making a joint decision on both (or all) factors simultaneously.<sup>g</sup> The OCII under current information ( $OCII_{Current}$ ) is equal to the expected NMB for the treatment-combination that would be adopted if we took account of all interactions, minus the NMB for the treatment combination that would be adopted based on an analysis ignoring some or all interactions.  $OCII_{Current}$  equals zero whenever this pair of analyses give the same conclusion. To obtain a fair, consistent, unbiased estimate of the opportunity cost, both NMB estimates should be based on analyses allowing for interactions (i.e. inside-the-table).<sup>h</sup>

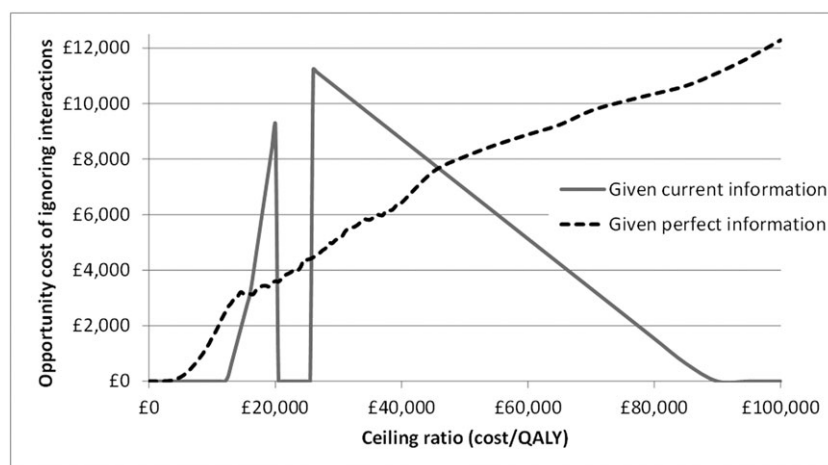
The OCII under perfect information ( $OCII_{Perfect}$ ) can be estimated by replicating the calculation of OCII across the posterior distribution of NMB for each treatment and averaging across the distribution. For the worked example described above, this was implemented by identifying the treatment that would be adopted based on inside-the-table and at-the-margins analyses for each bootstrap replicate and taking the difference between the inside-the-table NMB estimates for these two treatments.  $OCII_{Perfect}$  was then calculated by averaging these differences across all 1000 bootstraps.

Within this example, allowing for interactions and making a joint decision between the four cells would lead us to adopt the treatment that has highest expected NMB at the ceiling ratio of interest (£30 000/QALY): in this case *b*, which has an expected NMB of £479 504 per patient (Tables V, VI). If we conducted at-the-margins analysis and made independent decisions about A and B, we would instead adopt *ab*, which has an expected NMB of £468 985 (based on inside-the-table analysis). The  $OCII_{Current}$  (i.e. the value of allowing for interactions) therefore equals £10 518 (£479 504–£468 985) per patient at a £30 000/QALY ceiling ratio. This means that ignoring interactions and basing decisions on at-the-margins analysis would waste £10 518 of healthcare funds per patient treated, which is equivalent to losing 0.35 QALYs/patient. At a population level, spending £1 billion on *ab* will allow us to treat 7999 patients and accrue 158 381 QALYs, whereas allowing for interactions and adopting *b* will allow us to treat 9165 patients and accrue 179 825 QALYs: 21 443 more than we would achieve using the biased at-the-margins estimates and making separate decisions on A and B.  $OCII_{Current}$  was zero in three ceiling ratio ranges (£0–£12 000, £20 500–£25 500 and  $\geq$ £90 000; Figure 2), as analyses with and without interactions give the same treatment adoption decision at these ceiling ratios.

Bootstrapping also enables us to calculate the probability that ignoring interactions would give a misleading conclusion (i.e. adoption of a treatment that does not have highest NMB when we

<sup>g</sup>We define OCII in terms of NMB. However, it could equivalently be expressed in health units (e.g. QALYs) using net health benefit (NHB).

<sup>h</sup>If the OCII is calculated using NMB estimates from two different analyses (e.g. basing the NMB for the treatment combination that would be adopted based on at-the-margins analysis on the maximum NMB predicted in regression *without* interaction terms), the OCII may, misleadingly, appear to be zero (or even negative) if the bias inherent within at-the-margins analysis causes NMB for that treatment to be overestimated.



**Figure 2.** Opportunity cost of ignoring interactions given both current and perfect information.

allow for interactions), which is equal to the proportion of bootstrap replicates in which the OCII is  $>0$ . The probability is zero at ceiling ratios  $<£1000$  (because there is negligible chance that any treatment other than  $\theta$  has highest NMB regardless of whether interactions were considered) and peaks at £13 000/QALY (Supporting Information 2, Figure A4), just above the ICER for  $a$  versus  $\theta$ .

$OCII_{\text{Perfect}}$  takes account of the implications of adopting the wrong treatment, as well as the probability that the conclusions are sensitive to interactions; it therefore initially followed a similar trend to the probability but continued to rise with ceiling ratio as the value placed on maximising QALYs increases (Figure 2). Unlike  $OCII_{\text{Current}}$ , which equals zero when the point estimates of the analyses give the same conclusion,  $OCII_{\text{Perfect}}$  will not equal zero unless there is no uncertainty about which treatment has highest NMB.  $OCII_{\text{Perfect}}$  will generally be approximately equal to the difference in EVPI between the two analyses but will rarely be exactly equal because the EVPI estimates from at-the-margins analysis are biased due to interactions being ignored, while those from inside-the-table analysis may be overestimated due to inefficiency.

## 4. Discussion

In this paper, we have identified the challenges associated with economic evaluations conducted alongside factorial RCTs and demonstrated how such studies can be analysed with and without allowance for interactions, on a natural scale or using GLM and using a frequentist framework or a Bayesian framework with either non-informative or informative priors.

Within this hypothetical example, allowing for interactions had a relatively small effect on the predicted costs and QALYs for each treatment but changed the conclusions about which treatment was best value for money. Although hypothetical, this example illustrates the methods that can be used, the impact that interactions could have in real RCTs and the conflicting conclusions which different analytical methods may produce. As demonstrated above, the difference between analyses with and without interactions is due to the bias introduced by assuming no interaction; when ignoring interactions changes the conclusions, taking account of interactions in the analysis will achieve more efficient allocation of healthcare resources.

In this paper, we proposed the opportunity cost of ignoring interactions as a measure of the impact that interactions have on the conclusions and the value of avoiding bias by taking account of interactions could change the conclusions.  $OCII_{\text{Current}}$  demonstrates the extent to which the treatment adoption decision is affected by interactions at different ceiling ratios and the net loss to the healthcare system from ignoring interactions. Even if interactions do not change the conclusions at a £30 000/QALY threshold, there is likely to be at least one ceiling ratio range at which the conclusions do change.

$OCII_{\text{Perfect}}$  takes account of the implications of adopting the wrong treatment, as well as the probability that the conclusions are sensitive to interactions. Situations where  $OCII_{\text{Current}}=0$ , but  $OCII_{\text{Perfect}}>0$  are analogous to those with dynamic value of heterogeneity [64], suggesting that conducting further research could make the conclusions sensitive to interactions. However, whereas

$OCII_{Current}$  has a clear interpretation, it is less obvious how analysts and decision-makers should interpret  $OCII_{Perfect}$  or what the implications of this figure are for decision-making.

Both  $OCII_{Current}$  and  $OCII_{Perfect}$  are based on comparisons between two specific analyses; the analyses being compared need to be clearly stated in all applications. One limitation of these measures is that the inside-the-table analysis used to estimate NMB and obtain unbiased conclusions is not a 'gold standard' and could give the 'wrong' treatment adoption decision by chance.

Allowing for interactions increases SEs and the width of confidence intervals [17]. For economic evaluations, this means that allowing for interactions generally increases the EVPI, brings the probability that any given treatment-combination is cost-effective closer to 25% (even-odds across the four treatments) and decreases the height of the cost-effectiveness frontier, *except* at ceiling ratios where the treatment maximising expected NMB differs between analyses. It is unclear whether analyses excluding interactions underestimate the value of conducting further research, or whether including interactions overestimates the value of research. However, in situations where there are important interactions, it could be argued that the higher EVPI estimate from inside-the-table analysis more accurately reflects the value of collecting additional information. Nonetheless, given that analyses which take all interactions fully into account have less statistical power and may be more costly in time and resources, there may be value in excluding interactions that are considered unimportant based on pre-specified criteria, such as statistical significance, information criteria or magnitude.

## 5. Recommendations

Based on the challenges and the methods explored above, we have developed 14 recommendations for health economists, trialists and statisticians designing, analysing and reporting economic evaluations based on factorial trials. These recommendations represent a starting point for improving research practice and might later be developed into a formal set of consensus-based guidelines, similar to the CHEERS [65] or CONSORT [66] guidelines.

- 1 The aims and methods of economic evaluation are different from analyses of clinical trial endpoints, and it will frequently be inappropriate to replicate the analytical methods used for the primary clinical endpoint in the economic evaluation. In particular, economic evaluation focuses on estimation rather than hypothesis testing, and large super-additive or qualitative interactions are particularly likely to arise for costs and QALYs. Furthermore, the conclusions of economic evaluations must be drawn on a natural scale, unlike clinical endpoints, which are frequently interpreted on a logarithmic scale to eliminate interactions. Statisticians and health economists should discuss analytical methods while statistical analysis plans are being prepared to help minimise any unnecessary differences in approach.
- 2 The likelihood of interactions in each component of NMB (e.g. QALYs, intervention costs, hospitalisations, etc.) should be considered at the start of the study. Although it may not be appropriate to exclude interactions from the analysis solely on the basis of prior beliefs, *a priori* considerations may identify components of NMB that need to be analysed separately (see recommendation 7).
- 3 Any economic evaluation of factorial trials should follow best practice guidelines for trial-based economic evaluation [4,5,65] and take account of the distribution of costs and benefits, missing data, censoring and correlations between costs and benefits.
- 4 A clear decision rule, or combination of rules, determining the situations where interactions will be included in the base case analysis should be pre-specified to avoid data dredging and bias. For clinical endpoints, a two-stage testing approach is generally used, whereby interactions are only included in the main analysis if they are statistically significant in an initial test [16,17,21,67]; this approach increases statistical power but introduces a small amount of bias, which may be acceptable for analyses driven by statistical significance. By contrast, in economic evaluation, we need to choose between treatment combinations based on expected incremental costs and QALYs and statistical inference is often said to be irrelevant [60]. In economic evaluation, it may therefore be appropriate to include interactions unless proven to be negligible, rather than exclude them unless proven to be important, because it is preferable to conduct an inefficient analysis including all interactions than introduce bias that changes the conclusions. However, although unbiased, analyses allowing for interactions may nonetheless give the wrong



- answer by chance (especially for small samples) and would systematically overestimate the value of information and bias CEACs if there is genuinely no interaction.
- 5 Bayesian analyses attaching informative priors to interaction terms provide an alternative compromise between including and excluding interactions [7,43,68]. Although this has rarely been done for clinical endpoints, it shows particular promise for economic evaluation, which is frequently done in a Bayesian framework, or with Bayesian interpretations, and where consideration of external evidence alongside the current trial is recommended [69]. This can be done using MCMC [7], or Bayesian bootstrapping [70]. Informative priors could be based on prior evidence (e.g. previous factorial trials) or beliefs about the mechanisms by which interactions may arise; however, care should be taken to ensure that subjective priors are given a suitably low weight in the analysis. In this paper, we followed Welton *et al.* [7] in assuming that there is a 95% probability that the interaction for QALYs is sub-additive but non-qualitative. However, this prior appeared to be highly informative, substantially changing predicted costs and QALYs and halving the size of SEs around the interaction term; Welton *et al.* also found their trial results to be sensitive to the priors used for the interaction term [7,43].
  - 6 Regression analysis provides a convenient way to evaluate interactions and main effects, although inside-the-table analysis gives equivalent results in the absence of additional covariates. Regression analyses facilitate adjustment for baseline imbalance, which may be particularly important for factorial trials: especially for small trials and those with many treatment arms and/or imbalance in characteristics (e.g. baseline utility) that are incorporated within estimates of cost and QALYs. The model specification should be chosen carefully. If factors are thought to have multiplicative effects, GLM may be more appropriate than OLS on transformed data.
  - 7 If the likely importance of interactions or the appropriate scale of analysis differs between different types of cost or benefit (e.g. cost of drugs, side-effects or clinical events), it may be useful to analyse these components of NMB separately, while accounting for correlations between components.
  - 8 Multiple imputation models (if used) should include all treatment indicators that are considered in any analysis. In general, it will be necessary to include dummy variables for each factor and a full set of interaction terms as predictors of missing data. Omitting interaction terms from the imputation model could cause subsequent analyses to underestimate interactions.
  - 9 Any extrapolation of trial results should be conducted inside-the-table or allow for interaction terms unless interactions are shown to be negligible. When results are extrapolated using parametric survival models, researchers could estimate a single model with interaction terms and dummies for each factor, or (if the survival function differs between arms) estimate separate survival models for each cell (or several sets of cells). Similarly, event-based cost-effectiveness analyses [31,71] could allow for interactions within model(s) predicting the risk of events and/or analyse model output separately for each treatment combination.
  - 10 Abstracts and manuscripts describing economic evaluations based on factorial trials should state that the trial is factorial.
  - 11 It is more appropriate to report results for all factors in a single paper (rather than reporting each factor separately), unless there is clear evidence that treatment effects are additive or that results averaged over the levels for the other factor are meaningful. In particular, if it is necessary to allow for interactions in the analysis, it is also essential to interpret the results as a joint decision between all cells in the factorial design, which requires presentation of all factors simultaneously. Nonetheless, in some situations, it may be useful to present different factors to different audiences (e.g. surgeons and nutritionists), present early results for one factor before another reaches its primary endpoint or measure health outcomes using different units for different factors. In these situations, results for  $ab$  versus  $b$  and  $a$  versus  $0$  could be presented separately from the results of  $ab$  versus  $a$  and  $b$  versus  $0$ , in addition to a combined paper if journal editors allow.
  - 12 The results of a sensitivity analysis that includes all interactions (e.g. Tables V, VI) should always be reported whenever the base case analysis excludes any interactions, to evaluate the potential for bias resulting from assuming additive effects. Sensitivity analyses showing the results with different assumptions about interactions are also useful to demonstrate whether or not the analysis is sensitive to interactions. Estimating the opportunity cost of ignoring interactions can help evaluate the impact of excluding interactions. Presenting mean, disaggregated outcomes in each arm allowing for all interactions (e.g. Table II) also enables readers to see the magnitude and direction of interactions.

- 13 Researchers should present CEACs for multiple interventions (Figure 1), which show the probability that each cell of the factorial design has highest NMB and a cost-effectiveness frontier [59]. Pair-wise CEACs showing the probability that A is cost-effective compared with not-A (Supporting Information 2, Figure A2) are not sufficient for factorial trials unless there is clear evidence of additive effects. Similarly, the value of information should be evaluated for the joint decision between all mutually exclusive alternatives evaluated in the trial, not separately for each factor.
- 14 Conclusions should be based on an incremental comparison between ‘mutually exclusive’ treatment-combinations that considers the factors as interacting treatments (e.g. Tables V, VI), regardless of whether interactions are included in the analyses that estimate the mean cost and mean QALYs for each arm. Presenting the costs and QALYs for the full set of mutually exclusive alternatives and identifying the cost-effectiveness frontier reminds decision-makers of the importance of making a joint decision. In particular, presenting  $\theta$ ,  $a$ ,  $b$  and  $ab$  as mutually exclusive alternatives forces decision-makers to consider whether  $a$  is better than  $b$  or whether  $ab$  is appropriate, whereas considering A and B separately may lead to recommendations for both treatments without explicit guidance of whether they should be used together.

### 5.1. Funding sources

Alastair Gray was in receipt of a National Institute Health Research (NIHR) Senior Investigator Award (NF-SI-0509-10206) during the time this research was undertaken. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health.

### 5.2. Data sharing

The simulated data used in the analysis are available at [56], and the code used to generate the results is presented in Supporting Information 2.

## Acknowledgments

We would like to thank Richard Grieve, Oliver Rivero Arias, Iryna Schlackow and members of the Health Economists’ Study Group for their helpful comments on earlier versions of this manuscript.

## References

1. Doshi JA, Glick HA, Polsky D. Analyses of cost data in economic evaluations conducted alongside randomized controlled trials. *Value in Health* 2006; **9**(5):334–340.
2. Gray A, Clarke P, Wolstenholme J, Wordsworth S. In *Applied Methods of Cost-Effectiveness Analysis in Health Care*, Gray A, Briggs A (eds). Oxford University Press: Oxford, 2011.
3. Glick HA, Doshi JA, Sonnad SS, Polsky D. In *Economic Evaluation in Clinical Trials*, Gray A, Briggs A (eds). Oxford University Press: Oxford, 2007.
4. Ramsey SD, Willke RJ, Glick H, Reed SD, Augustovski F, Jonsson B, Briggs A, Sullivan SD. Cost-effectiveness analysis alongside clinical trials II—An ISPOR Good Research Practices Task Force report. *Value in Health* 2015; **18**(2):161–172.
5. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. *BMJ* 2011; **342**:d1548.
6. Oppong R, Jowett S, Nicholls E, Whitehurst DG, Hill S, Hammond A, Hay EM, Dziedzic K. Joint protection and hand exercises for hand osteoarthritis: an economic evaluation comparing methods for the analysis of factorial trials. *Rheumatology (Oxford)* 2015; **54**(5):876–883.
7. 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. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; **171**(Part 4):807–841.
8. Frempong SN, Goranitis I, Oppong R. Economic evaluation alongside factorial trials: a systematic review of empirical studies. *Expert Review of Pharmacoeconomics & Outcomes Research* 2015; **15**(5):801–811.
9. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ* 2010; **340**:c723.
10. Silagy CA, Jewell D. Review of 39 years of randomized controlled trials in the British Journal of General Practice. *The British Journal of General Practice* 1994; **44**(385):359–363.
11. Fisher RA. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 1926; **33**:503–513.
12. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA* 2003; **289**(19):2545–2553.

13. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. *Statistics in Medicine* 1989; **8**(2):161–171.
14. Armitage P, Berry G, Mathews JNS. *Statistical Methods in Medical Research* (4th edn). Blackwell Science Ltd.: Malden, MA, 2002.
15. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. *BMC Medical Research Methodology* 2003; **3**:26.
16. Hung HM, Chi GY, O'Neill RT. Efficacy evaluation for monotherapies in two-by-two factorial trials. *Biometrics* 1995; **51**(4):1483–1493.
17. Hung HM. Two-stage tests for studying monotherapy and combination therapy in two-by-two factorial trials. *Statistics in Medicine* 1993; **12**(7):645–660.
18. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Cochrane Database of Systematic Reviews* 2008; **2**:CD007176.
19. Kenyon S, Pike K, Jones DR, Brocklehurst P, Marlow N, Salt A, Taylor DJ. Childhood outcomes after prescription of antibiotics to pregnant women with spontaneous preterm labour: 7-year follow-up of the ORACLE II trial. *Lancet* 2008; **372**(9646):1319–1327.
20. Muller KE, Fetterman BA. *Regression and ANOVA: An Integrated Approach Using SAS Software*. Cary, NC: SAS Institute Inc, 2002.
21. Ng T. The Impact of a Preliminary Test for Interaction in a  $2 \times 2$  Factorial Trial. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*: Alexandria, VA, 1991; 220–227.
22. Piantadosi S. Chapter 19: Factorial trials. In *Clinical Trials: a Methodological Perspective*. Wiley Series in Probability and Statistics (Second edn). John Wiley & Sons Inc: Hoboken, NJ, 2005; 501–513.
23. Byar DP. Factorial and reciprocal control designs. *Statistics in Medicine* 1990; **9**(1–2):55–63.
24. Lubsen J, Pocock SJ. Factorial trials in cardiology: pros and cons. *European Heart Journal* 1994; **15**(5):585–588.
25. Gray A, Read S, McGale P, Darby S. Lung cancer deaths from indoor radon and the cost effectiveness and potential of policies to reduce them. *BMJ* 2009; **338**:a3110.
26. Dolan P. Modeling valuations for EuroQol health states. *Medical Care* 1997; **35**(11):1095–1108.
27. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care* 2002; **40**(2):113–128.
28. Byar DP, Piantadosi S. Factorial designs for randomized clinical trials. *Cancer Treatment Reports* 1985; **69**(10):1055–1063.
29. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *The New England Journal of Medicine* 1977; **296**(13):716–721.
30. Manning W. Dealing with skewed data on costs and expenditures. In *The Elgar Companion to Health Economics*, Jones A (ed). Edward Elgar: Cheltenham, UK, 2006; 439–446.
31. Willan AR, Briggs AH. In *Statistical Analysis of Cost-Effectiveness Data*, Senn S, Scott M, Bloomfield P, Barnett V (eds). Chichester: John Wiley & Sons Ltd, 2006.
32. Karlsson G, Johannesson M. The decision rules of cost-effectiveness analysis. *Pharmacoeconomics* 1996; **9**(2):113–120.
33. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. *Methods for the Economic Evaluation of Health Care Programmes* (3rd edn). Oxford University Press: New York, 2005.
34. Barber J, Thompson S. Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research & Policy* 2004; **9**(4):197–204.
35. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* 2002; **11**(5):415–430.
36. Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics* 2005; **14**(12):1217–1229.
37. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Economics* 2005; **14**(5):487–496.
38. Hernandez Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in Health* 2012; **15**(3):550–561.
39. Basu A, Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making* 2012; **32**(1):56–69.
40. Willan AR, Pinto EM. The value of information and optimal clinical trial design. *Statistics in Medicine* 2005; **24**(12):1791–1806.
41. Eckermann S, Willan AR. Expected value of information and decision making in HTA. *Health Economics* 2007; **16**(2):195–209.
42. Briggs A, Sculpher M, Claxton K. In *Decision Modelling for Health Economic Evaluation*, Gray A, Briggs A (eds). Oxford University Press: Oxford, 2006.
43. Brown J, Welton NJ, Bankhead C, Richards SH, Roberts L, Tydeman C, Peters TJ. A Bayesian approach to analysing the cost-effectiveness of two primary care interventions aimed at improving attendance for breast screening. *Health Economics* 2006; **15**(5):435–445.
44. Briggs A, Fenn P. Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics* 1998; **7**(8):723–740.
45. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Economics* 1997; **6**(4):327–340.
46. Briggs A. Statistical methods for cost-effectiveness analysis alongside clinical trials. In *The Elgar Companion to Health Economics*, Jones A (ed). Edward Elgar: Cheltenham, UK, 2006; 503–513.
47. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics* 2004; **13**(5):461–475.

48. Nixon RM, Wonderling D, Grieve RD. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Economics* 2010; **19**(3):316–333.
49. Johannesson M, O’Conor RM. Cost-utility analysis from a societal perspective. *Health Policy* 1997; **39**(3):241–253.
50. Johannesson M. Theory and Methods of Economic Evaluation of Health Care. Kluwer Academic Publishers: Dordrecht, The Netherlands, 1996.
51. Heshmat S. Chapter 9: Connecting value and cost: cost-effectiveness analysis. In *An Overview of Managerial Economics in the Health Care System*. Albany, NY: Delmar, 2001.
52. Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weinstein M. *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge University Press: Cambridge, UK, 2001.
53. Weinstein MC. Chapter 5: From cost-effectiveness ratios to resource allocation: where to draw the line? In *Valuing Health Care: Costs, Benefits and Effectiveness of Pharmaceuticals and Other Medical Technologies* (1st paperback edition edn), Sloan FA (ed). Cambridge University Press: Cambridge, 1996; 77–97.
54. Evans DB, Edejer TT, Adam T, Lim SS. Methods to assess the costs and health effects of interventions for improving health in developing countries. *BMJ* 2005; **331**(7525):1137–1140.
55. Tan-Torres Edejer T, Baltussen R, Adam T, Hutubessy R, Acharya A, Evans DB, Murray CJL. *Making Choices in Health: WHO Guide to Cost-effectiveness Analysis*. World Health Organization: Geneva, Switzerland, 2003.
56. Dakin HA, Gray A. Data accompanying “Economic evaluation of factorial randomised controlled trials: Challenges, methods and recommendations”, submitted to *Statistics in Medicine*. To be deposited on figshare, Wiley 2017, DOI to be confirmed on publication.
57. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
58. StataCorp. Suest – seemingly unrelated estimation. In *Stata Base Reference Manual: Release 11, Vol. 3, Q–Z*. StataCorp LP: College Station, TX, 2009; 1800–1818.
59. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Economics* 2001; **10**(8):779–787.
60. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics* 1999; **18**(3):341–364.
61. Couper DJ, Hosking JD, Cisler RA, Gastfriend DR, Kivlahan DR. Factorial designs in clinical trials: options for combination treatment studies. *Journal of Studies on Alcohol* 2005; **66**(SUPPL. 15):24–32.
62. Briggs A. Economic evaluation and clinical trials: size matters. *BMJ* 2000; **321**(7273):1362–1363.
63. Coyle D, Buxton MJ, O’Brien BJ. Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. *Health Economics* 2003; **12**(5):421–427.
64. Espinoza MA, Manca A, Claxton K, Sculpher MJ. The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making* 2014; **34**(8):951–964.
65. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)—explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value in Health* 2013; **16**(2):231–250.
66. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; **340**:c869.
67. Green S. Factorial designs with time to event endpoints. In *Handbook of Statistics in Clinical Oncology* (Second edn), Crowley J, Ankerst DP (eds). Chapman & Hall/CRC: Boca Raton, FL, 2006; 181–189.
68. Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. *Biometrics* 1997; **53**(2):456–464.
69. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? *Health Economics* 2006; **15**(7):677–687.
70. Sadatsafavi M, Marra C, Aaron S, Bryan S. Incorporating external evidence in trial-based cost-effectiveness analyses: the use of resampling methods. *Trials* 2014; **15**:201.
71. Mihaylova B, Briggs A, Armitage J, Parish S, Gray A, Collins R. Lifetime cost effectiveness of simvastatin in a range of risk groups and age groups derived from a randomised trial of 20,536 people. *BMJ* 2006; **333**(7579):1145.

## Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article.