## University of Bristol - Explore Bristol Research
### General rights

Debate

# Design, analysis and presentation of factorial randomised controlled trials

Alan A Montgomery*[1], Tim J Peters[1] and Paul Little[2]

Address: [1]Division of Primary Health Care, University of Bristol, Cotham House, Cotham Hill, Bristol BS6 6JL, UK and [2]Community Clinical Sciences Division (Primary Medical Care Group), Faculty of Medicine, Health and Biological Sciences, Southampton University, Aldermoor Health Centre, Southampton SO16 5ST, UK

Email: Alan A Montgomery* - alan.a.montgomery@bristol.ac.uk; Tim J Peters - tim.peters@bristol.ac.uk; Paul Little - p.little@soton.ac.uk

* Corresponding author

## Abstract

**Background:** The evaluation of more than one intervention in the same randomised controlled trial can be achieved using a parallel group design. However this requires increased sample size and can be inefficient, especially if there is also interest in considering combinations of the interventions. An alternative may be a factorial trial, where for two interventions participants are allocated to receive neither intervention, one or the other, or both. Factorial trials require special considerations, however, particularly at the design and analysis stages.

**Discussion:** Using a 2 × 2 factorial trial as an example, we present a number of issues that should be considered when planning a factorial trial. The main design issue is that of sample size. Factorial trials are most often powered to detect the main effects of interventions, since adequate power to detect plausible interactions requires greatly increased sample sizes. The main analytical issues relate to the investigation of main effects and the interaction between the interventions in appropriate regression models. Presentation of results should reflect the analytical strategy with an emphasis on the principal research questions. We also give an example of how baseline and follow-up data should be presented. Lastly, we discuss the implications of the design, analytical and presentational issues covered.

**Summary:** Difficulties in interpreting the results of factorial trials if an influential interaction is observed is the cost of the potential for efficient, simultaneous consideration of two or more interventions. Factorial trials can in principle be designed to have adequate power to detect realistic interactions, and in any case they are the only design that allows such effects to be investigated.

## Background

Randomised controlled trials provide the best quality evidence in medical research, [1] but they require a large commitment of time and effort, certainly from the investigators and often from participants. As a result, trials can be expensive. For these reasons, investigators may consider evaluating more than one intervention in the same study. For a controlled trial of two interventions, one could consider a parallel three-arm trial, or even a four-arm trial if two distinct control groups are required. An example is a comparison of mailed guidelines with and without an educational outreach visit from community pharmacists to improve prescribing in general practice.[2] If target differences for both interventions are identical, these would require increases in sample size of 50% and 100% respectively compared with a two-arm trial.

**Table 1: Sample sizes required for 90% power and 1% two-sided alpha: main effects. Intervention A target difference = 0.35 standard deviations (SDs), total sample size = 486 (243 allocated to Intervention A, 243 allocated to the relevant control). Intervention B target difference = 0.3 SDs, total sample size = 664 (332 allocated to Intervention B, 332 allocated to the relevant control). A total sample size of n = 664 participants yields 90% power to detect differences of 0.3 SDs for Intervention B and 97% power to detect differences of 0.35 SDs for Intervention A.**

|  |  | Intervention B | | |
|---|---|---|---|---|
|  |  | YES | NO | TOTAL |
| Intervention A | YES | 166 | 166 | 332 |
|  | NO | 166 | 166 | 332 |
|  | TOTAL | 332 | 332 | 664 |

To detect the same target differences with the same power and alpha in a three-arm parallel group trial would require 907 participants: 243 allocated to Intervention A, and 332 allocated to each of Intervention B and control.

Correspondingly, the analyses would involve only two thirds or half of the total sample size. Since the power to detect treatment differences depends on the number of participants in the groups being compared rather than the total number in the trial, this can represent a rather inefficient use of resources.

An alternative may be a factorial trial, where participants are allocated to receive neither intervention, one or the other, or both. An example of such a trial is an evaluation of two decision aids for newly diagnosed hypertensive patients – that is, individual decision analysis and an information video plus leaflet.[3] Other examples are a factorial trial of two interventions to improve attendance for breast screening,[4] and a factorial trial of two interventions to improve adherence to antidepressant drugs.[5]

Although their use to date may have been limited,[6] factorial trials have the potential to confer advantages over the standard parallel-groups design. First, they enable efficient simultaneous investigation of two interventions by including all participants in both analyses. Second, it is possible in a factorial trial to consider both the separate effects of each intervention and the benefits of receiving both interventions together. In order to realise these advantages, however, factorial trials require some special considerations, particularly at the design and analysis stages. Although these issues have been discussed previously, [7] factorial trials continue to be often inappropriately analysed and interpreted. The aim of this paper is to explore these issues in the context of an individually randomised 2 × 2 factorial trial, although in principle the methods generalise to trials of more than two interventions.

## Discussion
### Design considerations
The prime issue here is the sample size of the trial. The most common procedure is to perform a separate calculation based on target effect sizes for each of the interventions compared with their respective controls (Table 1). The trial sample size is then simply the larger of these, and the trial is said to be powered to detect the main effects of each intervention. However, this sample size is based on the crucial assumption that there is no interaction between the interventions – in other words, that the effect of intervention A does not differ depending on whether participants also receive intervention B. This will by no means always be a reasonable assumption, especially where interventions involve behavioural and/or organisational change.

If a trial is to have adequate power to detect an interaction, then the sample size will in general need to be increased. For example, to detect with the same power an interaction of the same magnitude as the main effects, a fourfold increase in sample size is required (Table 2).[8] With no increase in sample size, the interaction would need to be at least twice as large as the main effects to be detected with the same power;[8] this is very unlikely to be the case in practice. Smaller, more plausible interactions would require greatly increased sample sizes. If the interaction is of primary interest then it is essential that the trial is powered to detect a reasonable target interaction effect.

If the primary comparisons are the main effects then the approach in Table 1 is justifiable on grounds of efficiency. At the same time, it should be appreciated that the resultant precision for the interaction may be inadequate to exclude such an effect – that is, the confidence interval for the interaction will be relatively wide. In other words, the sample size will be insufficient to investigate the initial assumption that the interaction is unimportant. Virtually

**Table 2: Sample sizes required for 90% power and 1% two-sided alpha: interaction**

| Magnitude of effects (in units of standard deviations) | | |
| --- | --- | --- |
| Main effect | Interaction | Total sample size to detect interaction |
| 0.3 | 0.6 | 664 |
| 0.3 | 0.3 | 2656 |
| 0.3 | 0.15 | 10624 |

**Table 3: Descriptive statistics for the primary outcome (crude mean decisional conflict scores[3]) for the analysis of a 2 × 2 factorial trial**

| | | Video and Leaflet | | |
| --- | --- | --- | --- | --- |
| | | No | Yes | TOTAL |
| Decision analysis | No | 44 | 33 | 39 |
| | Yes | 28 | 27 | 28 |
| | TOTAL | 37 | 30 | 34 |

identical arguments apply to interactions for binary outcomes, although if logistic regression is used then the relative sizes of the interaction and main effects in Table 2 relate to the log odds scale.

### Analytical considerations

The second consideration is the analytical strategy, which should follow CONSORT guidelines.[9] In particular, the primary analyses should address the principal research questions. Table 3 presents the basic descriptive statistics for the analysis of an example 2 × 2 factorial trial.[3] To evaluate the decision analysis intervention, we compare patients who received both interventions plus those who received decision analysis only with patients who received video and leaflet only plus those who received neither intervention. In general, the correct analysis of such data requires the use of a multivariable regression model, especially if the numbers of subjects in each of the four combinations shown in Table 3 are unequal (in technical terms, if the design is 'unbalanced').

The approach in such models is essentially to obtain an average of the two differences (28–44) and (27–33), weighted according to the sample sizes. Regardless of the technical details, conceptually the primary analysis is a comparison of the *margins* of the 2 × 2 table. In the regression analyses, the effect of each intervention is adjusted for the other intervention as well as any necessary covariates, such as the outcome measure at baseline and stratification variables. In the context of a randomised trial with a continuous outcome, such adjustments are primarily to

improve precision, especially for individually randomised trials. [10,11]. For binary outcomes, a multivariable (logistic) regression analysis is required in order to obtain correct estimates of the effects and their standard errors.

In focussing on the average effect of each intervention, however, the above analysis assumes that the effect of each intervention is uninfluenced by the presence or absence of the other – that is, there is no *interaction* between them. Since factorial trials are rarely powered to detect interactions between the interventions, such effects are usually investigated as a secondary analysis. These are readily performed as extensions to the multivariable regression models described above, by simply introducing the appropriate interaction terms. However, the precision of the estimates of interaction is very likely to be too poor for large effects to be ruled out. In particular, a high p-value will most likely reflect low power and so cannot be taken as evidence for no interaction.

A special consideration for binary outcomes is the choice of regression method. Logistic regression is commonly used since, among other advantages, predicted proportions from this model are constrained to be in the allowable range (that is, between zero and one). [12] Logistic regression estimates odds ratios for the interventions and assumes that these effects operate multiplicatively on this scale. [13]

**Table 4: Presentation of the results of the primary analyses in a 2 × 2 factorial trial[3]**

|  | **Decision Analysis** n = 100 | **no Decision Analysis** n = 112 | **Video/leaflet** n = 104 | **no Video/leaflet** n = 108 |
|---|---|---|---|---|
| Total Decisional Conflict, mean (SD) | 27.6 (12.1) | 38.9 (18.3) | 30.3 (13.4) | 36.8 (18.8) |
| Adjusted difference[1,2] (95% CI) | -9.4 (-13.0 to -5.8)[3] | | -4.2 (-7.8 to -0.6)[3] | |
| p value | < 0.001 | | 0.021 | |

[1] Adjusted for age, sex, decisional conflict at baseline, factorial design and general practice (all comparisons are for the intervention compared with its respective control) [2] Negative differences represent a favourable outcome for the relevant intervention [3] The interaction between the interventions was investigated as a secondary analysis and was found to be significant (interaction coefficient = 12 (95% CI 5 to 19), p = 0.001)

### Presentation of results from a factorial trial

Regarding the results obtained at the main trial follow-up, the primary analysis relating to the margins of the 2 × 2 table should give estimates (such as a difference, odds ratio or risk ratio) and 95% confidence intervals comparing those individuals allocated to receive an intervention with those allocated to not receive it. The number of such comparisons will be equal to the number of interventions investigated in the trial. A common misunderstanding is that the outcome measures should be analysed and presented separately for each of the four factorial cells, but to do so would fail to realise the full efficiency and purpose of the factorial design. Even in trials powered for main effects, a test and confidence interval for the interaction should be provided. An indication of the imprecision of the results for the interaction is especially important given the above concerns about the adequacy of the sample size to investigate such effects. Table 4 demonstrates how the results of the primary analyses in our example trial were presented. [3]

In addition to the primary comparative statistics noted above, it is also advisable to present descriptive statistics for outcome measures at follow-up within each of the factorial 'cells' in the trial (four in the case of a 2 × 2 design). These can either be tabulated or included in the text of the paper along with the regression coefficient and 95% confidence interval for the interaction term. This allows interpretation of the magnitude of any antagonism or synergism between the interventions, and would of course be essential if the interaction was the primary effect of interest. In our example, there was a significant antagonistic interaction, such that there was no added benefit from a second intervention (Tables 3 and 4).

The most appropriate presentation of baseline data depends on the original primary research question and the results obtained. If an interaction is either posited or observed, then descriptive baseline data for the four cells is more helpful; otherwise, the margins are more relevant to the issue of baseline comparability and correspond to the primary analysis. With more than two interventions the marginal approach increasingly becomes the only feasible option.

### Implications

Factorial designs provide an efficient method of evaluating more than one intervention in the absence of interactions. This raises the question, however, of the degree of certainty one might have in advance that there is no interaction between the interventions. Although Bayesian methods might be helpful here in that they formalise such prior information/beliefs, in practice there will be much uncertainty, and so the issue is rather one of a judgement as to how influential any likely interaction might be in the context of the trial. In particular, if the direction of the effect of intervention A is different for the levels of intervention B (a 'qualitative' interaction) then a factorial trial would be appropriate if this interaction was of key interest, in which case the trial should be powered to detect the interaction. If there is likely to be only a minor difference in magnitude in the effect of intervention A across the levels of intervention B (that is, a small 'quantitative' interaction) then a factorial trial powered to detect the main effects is more appropriate.[3] In any case, the practical question of how to present the intervention effects in the presence of a sizeable interaction remains. If the interaction is qualitative then the main effects will almost certainly be misleading and the cell means and interaction effect together with separate estimates and confidence intervals for the relevant subgroups will be the only option. [14] For quantitative interactions such as in our example, the main effects will over-estimate the effect for some individuals and under-estimate it for others. Whilst the interaction and the cell means must still be presented, the main effects may nonetheless be a reasonable representation of the intervention effects, both separately and combined.

A factorial trial would be unsuitable for interventions that could not be used in conjunction with one another, such as two different minor surgical procedures for a dermatological problem. For interventions such as those in Table 3, though, factorial trials are an especially useful option if the principal interest is in comparing each intervention

with its respective control and also in considering if there is any suggestion of an interaction between them. Indeed, an appropriately powered factorial trial is the only design that allows such effects to be investigated. Conversely, factorial designs would be contra-indicated if primary interest was in the direct comparison of the two interventions applied individually – for example, decision analysis alone versus video/leaflet alone.

The decision as to the suitability of the factorial design must therefore take a number of issues into account – in particular, the nature of the interventions, the setting of the study including the participants, the comparisons of interest and the outcome measure. For instance, interactions may be considered to be more likely with behavioural interventions, when as in our example the benefits may be achieved with either intervention and there is relatively little additional benefit from receiving a second intervention.[3] In terms of the outcome measure, a consideration for binary variables beyond the issues covered in this paper is the choice of the statistical model employed – that is, whether the effects of the interventions are presumed to work additively in a linear model for proportions, or multiplicatively as in a linear logistic model. [12] Since the presence or absence of interactions for a binary outcome depends on the statistical model employed, choice of the latter is an important issue.

## Summary
Difficulties in interpreting the results of factorial trials if an influential interaction is observed should be recognised as the cost of the potential for efficient, simultaneous consideration of two or more interventions. As described in this paper, factorial trials can in principle be designed to have adequate power to detect realistic interactions, but this has major implications for the sample size. On the other hand, unlike parallel groups trials a factorial design does enable investigation of interactions in the analysis, albeit with limited power. Researchers should be aware of such issues when using factorial designs.

## Competing interests
None declared.

## Authors' contributions
AM and TP drafted the paper, with input from PL. All authors read and approved the final manuscript.

## References
1.  Grimes DA, Schulz KF: **An overview of clinical research: the lay of the land.** *Lancet* 2002, **359**:57-61.
2.  Watson M, Gunnell D, Peters T, Brookes S, Sharp D: **Guidelines and educational outreach visits from community pharmacists to improve prescribing in general practice: a randomised controlled trial.** *Journal of Health Services Research and Policy* 2001, **6**:207-213.
3.  Montgomery AA, Fahey T, Peters TJ: **A factorial randomised controlled trial of decision analysis and an information video plus leaflet for newly diagnosed hypertensive patients.** *Br J Gen Pract* 2003, **53**:446-453.
4.  Bankhead C, Richards SH, Peters TJ, Sharp DJ, Hobbs FDR, Brown J *et al.*: **Improving attendance for breast screening among recent non-attenders: a randomised controlled trial of two interventions in primary care.** *Journal of Medical Screening* 2001, **8**:99-105.
5.  Peveler R, George C, Kinmonth A-L, Campbell M, Thompson C: **Effect of antidepressant drug counselling and information leaflets on adherence to drug treatment in primary care: randomised controlled trial.** *BMJ* 1999, **319**:612-615.
6.  Sheikh A, Smeeth L, Ashcroft R: **Randomised controlled trials in primary care: scope and application.** *Br J Gen Pract* 2002, **52**:746-751.
7.  Ottenbacher KJ: **Interpretation of interaction in factorial analysis of variance design.** *Statistics in Medicine* 1991, **10**:1565-1571.
8.  Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G: **Subgroup analyses in randomised controlled trials quantifying the risks of false-positives and false-negatives.** *Health Technol Assess* 2001, **5(33)**:.
9.  Moher D, Schulz KF, Altman DG: **The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials.** *Lancet* 2001, **357**:1191-1194.
10. Senn S: *Statistical issues in drug development Chichester: John Wiley & Sons Ltd*; 1997.
11. Armitage P, Berry G, Matthews JNS: *Statistical methods in medical research* Fourthth edition. *Oxford: Blackwell Science Ltd*; 2002.
12. Collett D: *Modelling binary data* Secondth edition. *Boca Raton, Florida: Chapman & Hall/CRC*; 2003.
13. Kirkwood BR, Sterne JAC: *Medical statistics* Secondth edition. *Oxford: Blackwell Science Ltd*; 2003.
14. Janosky JE: **Interpretation of interaction in factorial analysis of variance design – letter.** *Statistics in Medicine* 1992, **11**:1403.

## Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/26/prepub