When should factorial designs be used for late-phase randomised controlled trials?

Ian R White[1*], Alexander J Szubert[1*], Babak Choodari-Oskooei[1], A Sarah Walker[1], Mahesh K B Parmar[1]

[1] MRC Clinical Trials Unit at UCL, London, UK.

* Equal contribution

Running head (38 chars, limit 50): When should factorial designs be used?

Length: abstract (404 words, limit 425), text (4013, 4000), figures (1, + 0 supplementary figures, limit 6 tables or figures), Tables (2, + 1 supplementary tables), 51 references (no limit)

Corresponding author: Ian R White, MRC Clinical Trials Unit at UCL, 2nd Floor, 90 High Holborn, London WC1V 6LJ, UK. Tel: +44 20 7670 4715. ian.white@ucl.ac.uk

Funding: The authors disclose receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the UK Medical Research Council [grant numbers MC_UU_12023/29 and MC_UU_00004/09].

ABSTRACT

Background: A 2-by-2 factorial design evaluates two interventions (A vs. control and B vs. control) by randomising to control, A-only, B-only or both A and B together. Extended factorial designs are also possible (e.g. 3-by-3 or 2-by-2-by-2). Factorial designs often require fewer resources and participants than alternative randomised controlled trials (RCTs), but they are not widely used. We identified several issues that investigators considering this design need to address, before they use it in a late-phase setting.

Methods: We surveyed journal articles published in 2000-2022 relating to designing factorial randomised controlled trials (RCTs). We identified issues to consider based on these and our personal experience.

Results: We identified clinical, practical, statistical and external issues that make factorial RCTs more desirable. Clinical issues are: 1) interventions can be easily co-administered; 2) risk of safety issues from co-administration above individual risks of the separate interventions is low; 3) safety or efficacy data are wanted on the combination intervention; 4) potential for interaction (e.g. effect of A differing when B administered) is low; 5) it is important to compare interventions with other interventions balanced, rather than allowing randomised interventions to affect choice of other interventions; 6) eligibility criteria for different interventions are similar. Practical issues are: 7) recruitment is not harmed by testing many interventions; 8) each intervention and associated toxicities is unlikely to reduce either adherence to the other intervention or overall follow-up; 9) blinding is easy to implement or not required. Statistical issues are: 10) a suitable scale of analysis can be identified; 11) adjustment for multiplicity is not required; 12) early stopping for efficacy or lack of benefit can be done effectively. External issues are: 13) adequate funding is available; 14) the trial is not intended for licensing purposes. An overarching issue (15) is that factorial design should give a lower sample size requirement than alternative designs. Across designs with varying non-adherence, retention, intervention effects and interaction effects, 2-by-2 factorial designs require lower sample size than a three-arm alternative when one intervention effect is reduced by no more than 24-48% in the presence of the other intervention compared with in the absence of the other intervention.

Conclusions: Factorial designs are not widely used and should be considered more often using our issues to consider. Low potential for at most small to modest interaction is key, for example where the interventions have different mechanisms of action or target different aspects of the disease being studied.

Keywords: randomised controlled trial, clinical trial, design, factorial

BACKGROUND

In a factorial design for a randomised controlled trial (RCT), two or more randomised comparisons are carried out independently in the same sample of patients[1]. This design has the potential to address multiple questions in an efficient way. For example, a 2-by-2 factorial design compares each of two interventions (A, B) to control by randomising participants to control, A-only, B-only or both A and B; we call each of these four randomised combinations an arm. The analysis then compares all those randomised to A (with or without B) to all those not randomised to A (with or without B), and similarly for B. Compared to two 2-arm RCTs (A vs. control and B vs. control) or a three-arm RCT (A vs. B. vs. control), a 2-by-2 factorial RCT comparing A to control, and B to control, often requires fewer participants and hence resources.

Factorial designs are generally used in late-phase trials. For example, the TRISST trial considered whether men with successfully treated testicular cancer could safely have a lighter surveillance schedule[2]. Men were randomised to 3 scans or the standard 7 scans over 5 years, and also to MRI scans or the standard CT scans. Separate analyses addressed the non-inferiority of 3 scans versus 7 and MRI versus CT scans.

Extended factorial designs are possible and allow more questions to be addressed simultaneously. The TRACT trial was a 3-by-2-by-2 factorial in which children admitted to hospital with severe anaemia were randomised to (i) liberal blood transfusion, conservative transfusion or no transfusion, (ii) post-discharge multi-vitamin supplementation or routine care, and (iii) post-discharge cotrimoxazole prophylaxis or no prophylaxis[3].

We use the term “factorial randomisation” to mean a randomisation to all possible intervention combinations, “factorial analysis” to mean an analysis comparing all those randomised to a intervention (alone or with other interventions) with all those randomised to the corresponding control (alone or with other interventions), and “factorial design” to mean a trial with both factorial randomisation and factorial analysis. A “partial factorial design” is a factorial design where some participants do not participate in some randomisations or are randomised between a reduced set of interventions: for example, children with severity signs in the TRACT trial were not eligible for randomisation to no transfusion[3]. We consider factorial designs aiming to evaluate each intervention separately; larger sample sizes are needed to assess, for example, whether the combined intervention is more effective than either individual intervention[4, 5].

Factorial designs are sensitive to interactions between interventions, an issue we discuss below. Because of this, factorial randomisation is sometimes used without factorial analysis. This is appropriate where the aim is to explore interactions between interventions[1, 6] and/or an interaction between interventions is a real possibility. It is also appropriate where the aim is to compare each intervention combination to control. An example is the STAMPEDE trial in

oncology, where the initial aim was to compare each of zoledronic acid, celecoxib, docetaxel, zoledronic acid plus docetaxel and celecoxib plus docetaxel with control[7]. Factorial analysis was avoided because "there is no good evidence to support the notion that the intervention effects would only be additive at the patient level and we cannot rule out other forms of interaction: synergy, antagonism and a ceiling effect"[8]. Finally, factorial randomisation without factorial analysis can be used where the aim is to determine which intervention combination is the best[9-11].

Factorial trials have started to be used more, particularly in the field of infections where the different interventions are typically in different "domains" of intervention[12, 13]. However, we believe that factorial designs may be an under-used tool for addressing multiple questions at little extra difficulty or cost. For example, we searched a registry of RCTs, clinicaltrials.gov, to identify all phase III/IV interventional studies with randomised allocation first posted to the registry from 1 January 2015-31 July 2022 (search done 2 Aug 2022). Each clinicaltrials.gov record including the word "factorial" was reviewed by AS/IW to determine whether it truly represented a factorial design. Reasons why trials might be reported as factorial but determined as non-factorial included an A/B/A+B design with no control group, and a 2-way randomisation across two subgroups where the subgroup variable was not randomised. Of 22403 phase III/IV interventional studies with randomised allocation, 206 (0.9%) were factorial. Factorial trials constituted a smaller fraction (0.14%, 9/6212) of trials wholly-funded by industry than of other trials (1.22%, 197/16191, $p<0.001$ for difference).

We therefore aimed to compile a checklist of issues to consider when choosing between a factorial design and other designs, based on narrative review of literature relating to design of factorial RCTs.

METHODS

Narrative review of design of factorial RCTs

The full text of articles published in the following methodological journals between 1 January 2000-31 July 2022 was searched (5 April 2019, 11 August 2022) for "factorial": Biometrics, BMC Medical Research Methodology, Clinical Trials, Journal of the Royal Statistical Society (all series), Pharmaceutical Statistics, Statistics in Medicine, Trials and Statistical Methods in Medical Research. PubMed was also searched (26 April 2019, 11 Aug 2022) with the same date ranges to identify citations with "factorial" in the title (articles relating specifically to designing factorial RCTs were considered likely to include this). Titles, relevant abstracts and then the full text of relevant articles were reviewed by AS/IW to identify articles relating to designing factorial RCTs, including less recent key references in relevant articles. Articles

relating to phase II and cluster-randomised RCTs were excluded, as, in general, were articles relating to the protocol or results of a specific RCT. Issues to consider were identified based on these and the authors' personal experience (including [2, 3, 14-16] and ongoing trials).

Sample size required for factorial vs. other designs

A key potential advantage of a factorial design is its ability to answer multiple questions with very little, if any, increase in total size. However, this depends on a number of issues including the assumption of no substantial interaction. In order to assess the gains in efficiency in practice, we compared the sample size required to show superiority of one intervention (A) over control using factorial and other designs under various scenarios.

Specifically, we computed the required sample sizes to give 90% power for showing superiority of A compared with control in a 2-by-2 factorial design, assuming a 5% two-sided significance level. The four arms are described as 0, A, B and AB. We assumed values for each arm for: the expected outcome with perfect adherence (taking A to be effective, allowing B to be effective or ineffective, and varying the interaction between A and B); the expected proportion of non-adherers, assumed for sample size purposes to have the same outcome mean as the control arm; and the expected proportion of missing outcome data, assumed to be excluded from the analysis. We allowed unequal allocation ratios. We calculated the relative efficiency of evaluating A vs. control in a factorial design compared with a corresponding three-arm trial (A vs. B vs. control) as the inverse ratio of sample sizes required to achieve the same power. Sample size formulae are given in the supplemental material.

Initially, we assumed that the outcome is quantitative and that B is ineffective. The actual sample size and outcome variance did not affect the relative efficiency results. We considered the following scenarios:

1. Base case: all arms have 10% missing data and 10% non-adherence among observed data; equal allocation.

2. Perfect case: no missing data, perfect adherence, equal allocation.

3. Double missing with A: like base case, but arms A and AB have 20% missing data.

4. Double missing with B: like base case, but arms B and AB have 20% missing data.

5. Double non-adherence with A: like base case, but arms A and AB have 20% non-adherence.

6. Double non-adherence with B: like base case, but arms B and AB have 20% non-

adherence.

7. Double controls: like base case, but twice as many participants are allocated to the control arm as to each other arm. (i.e. 4:2:2:1 for factorial and 4:2:2 for three-arm).

We then repeated this set of calculations three times. In set 2, the outcome was again quantitative, but intervention B was effective. This gave the same results as for set 1, except that scenario 6 lost efficiency (results not shown). In sets 3 and 4, the outcome was binary, with proportion 40% in the control arm expected to reduce to 24% in the A arm with perfect adherence (risk ratio = 0.6). In set 3, B was ineffective, and in set 4 B was as effective as A.

We measured the interaction as a percentage of the main effect of A and varied it from 0% to 100%. For binary outcome, this was done on the risk difference scale (other scales are possible).

RESULTS

27 articles relating to designing factorial RCTs were identified including 14 in methodological journals[1, 5, 9, 17-30] and 13 in clinical journals[4, 31-43], of which some were overview articles. Based on these articles and the authors' personal experience, issues to consider when planning a factorial RCT were identified.

Clinical issues

Co-administration. Most factorial RCTs require co-administration of interventions. Factorial RCTs are more desirable if co-administration is easy. Difficulty of co-administration could be a reason to reject a factorial design out of hand.

Safety. Factorial RCTs are more desirable if the risk of safety issues from co-administration, above the individual risks of the separate interventions, is low.

Combination intervention. Factorial RCTs are more desirable if initial or additional safety or clinical data are wanted on the combination intervention. Safety data could be valuable to explore the risk of drug interactions, but sample sizes in a factorial design might not be adequate for a full safety evaluation of the combination intervention. Clinical data could include pharmacokinetic or pharmacodynamic data.

Interaction (effect modification). Factorial RCTs are more desirable if the potential for interaction is low: that is, if the effect of one intervention is unlikely to differ substantially when another intervention is also administered. It is sometimes stated that the factorial analysis rests on a no-interaction assumption: for example, ICH E9 Statistical Principles for Clinical Trials

recommends that evidence that there is likely to be no interaction is established in advance using prior information and data[6]. However, it is more useful to accept that interactions may occur and consider their plausibility, potential size and impact.

Interactions are less likely if the different interventions target different underlying domains of intervention[33]: for example in the TRACT trial above, blood transfusion, anti-bacterial prophylaxis and nutritional support are unlikely to interact[3]. Similarly the REALITY trial assessed the impact of interventions to reduce HIV viral load faster, reduce risks of co-infections and improve nutritional status on mortality in those with advanced HIV starting treatment[16].

Factorial designs usually have little power to detect interaction, but the suspicion of interaction has negative consequences for the interpretation of the trial. This is because the factorial analysis estimates the average intervention effect in a population randomly assigned to the other interventions, but in the presence of interaction this is rarely the estimand of clinical interest. Instead, clinical interest usually lies in the average intervention effect in a population either *receiving* or *not receiving* the other interventions; for example, if the other intervention B is found to be effective, then clinical interest is more likely to be in a population *receiving* intervention B[28].

Interaction can also reduce the power of a factorial design if interventions are less effective in the presence of other interventions[44]: we explore this below, and revisit the interaction issue in the discussion.

Health economists have suggested that interaction may be more likely to occur for costs and quality-adjusted life years, and may therefore have larger impact on health economic evaluations[25, 26].

Balancing other interventions. In situations where there is no “standard of care” for the alternative intervention(s), these may be used or not used dependent on physician preference, and hence, in an open-label trial, may be unbalanced across other comparisons. A factorial randomisation forces the alternative intervention(s) to be balanced across groups of the intervention of interest and may be more desirable if the estimand of interest is a pure intervention comparison.

Eligibility criteria. Factorial RCTs are more desirable if eligibility criteria for the different comparisons are similar. If eligibility criteria are different, then requiring all participants to be eligible for all comparisons would reduce the pool of potential participants; an alternative is a partial factorial design, where participants are included if they are eligible for any one or more comparisons, and are randomised in all eligible comparisons [30].

Practical issues

Recruitment. Factorial RCTs are less desirable if they compromise recruitment. For example, if there is a clinician or patient preference against one particular intervention, then including that intervention in a factorial design may adversely affect recruitment. On the other hand, addressing a number of different clinical questions could improve recruitment by increasing enthusiasm of both patients and investigators for a factorial trial.

Adherence. Factorial RCTs are less desirable when one or more of the interventions (or any toxicities associated with it) is likely to reduce adherence to other interventions, since this will reduce the effectiveness of the other intervention and reduce power.

Retention. Factorial RCTs are less desirable when one or more of the interventions is likely to reduce retention in the study, since this will increase the amount of missing data for all comparisons and hence reduce power and increase concerns about bias.

Where one intervention may adversely affect recruitment, adherence or retention, it may be preferable to explore that intervention in a separate trial where these complex issues can be specifically addressed without impacting on other comparisons.

Blinding. Factorial RCTs are more desirable if blinding is easy to implement (or not required). In a 2-arm RCT, it is often feasible to manufacture a placebo with similar physical properties (including size, colour, taste and smell) to the active medicine. In a 2-by-2 factorial RCT, one option is to use A+B, A+placebo B, B+placebo A and placebo A+placebo B ("double-dummy"). However, this might be undesirable because the increased number of pills makes non-adherence more likely. To minimise pill burden, it may be preferable to use combination pills containing A+B, A-only, B-only or placebo. However, this might not be feasible, for example if interventions are given at different frequencies (e.g. A once daily in the morning vs. B twice daily). It might also be undesirable because participants cannot stop just one of the interventions.

Statistical issues

Scale of analysis. Factorial RCTs are more desirable if a suitable scale can be identified on which interaction is unlikely. For example, for binary outcomes, lack of interaction on the odds ratio scale is not the same as lack of interaction on the risk ratio scale[45]. Interactions tend to be less common on the odds ratio and risk ratio scales than on the risk difference scale[46]. Once a suitable scale has been identified, the analysis (and in particular assessment of interactions) must be performed on that scale.

Multiplicity. Factorial RCTs are less desirable if adjustment for multiplicity is required. A standard 2-arm RCT would typically be designed with a two-sided false-positive (type I) error rate of 5%, and two independent 2-arm RCTs would have an overall false-positive error rate of nearly 10%. A factorial RCT may be viewed as two 2-arm RCTs, and therefore an error rate of 5% per comparison may be considered acceptable, or alternatively, control of the overall error rate may be considered necessary. If different scientific questions are being addressed, particularly with different primary endpoints, then adjustment for multiplicity should not be required[47] but has been the topic of debate[4, 21, 22, 31].

Stopping. Factorial RCTs are less desirable if adequate ways cannot be found to stop comparisons early, either for efficacy or for lack of benefit, without damaging other ongoing comparisons. We recommend using standard statistical stopping guidelines that control the pairwise type 1 error rate for each comparison using factorial analysis[30], recognising that statistical guidelines are only part of the decision to stop a trial arm early.

External issues

Funders. Factorial RCTs are more desirable if adequate funding is available. Even if they are an efficient way to answer multiple questions, factorial designs may still be more expensive than a single two-arm trial, and funders may conservatively prefer the latter, depending on the scientific importance of the additional questions. However, funders increasingly prefer platform trials for their increased efficiency[48], and factorial trials share with platform trials an ability to test many interventions within a single protocol.

Regulators. Factorial RCTs are not typically used for licensing trials, where focus is usually instead on pre-specification of other interventions, making the estimand more clear-cut.

Other issues

The literature review raised the concern that complex factorial designs with small sample sizes can lead to larger covariate imbalances across arms than simpler designs, and therefore minimisation procedures may be preferable to other randomisation schemes[29]. This point requires attention in some small trials, but does not affect the desirability of a factorial design.

Overarching issue: sample size

The issues described above are summarised in Table 1, together with the overarching issue of sample size. The main rationale for factorial RCTs is that they can have a lower sample size than alternative designs, including two independent 2-arm RCTs or 3-arm RCTs. However,

sample size requirement may be increased by a number of the issues considered above, in particular interaction, adherence and retention[19, 44].

Sample size in factorial RCTs may also need inflation to allow for multiple interventions being effective[49]. With a survival-type endpoint, the power depends on the number of events (e.g. deaths). For a factorial RCT with the same time-to-event outcome for each comparison (or correlated time-to-event outcomes, e.g. overall and progression-free survival), one intervention being effective decreases the number of events available for the other comparisons. The sample size might need to be inflated to allow for this. For a binary endpoint, whether sample size inflation is necessary depends on the proportions hypothesised under the null and alternative hypotheses and on the anticipated effect of the combined intervention.

We illustrate how to take account of interaction, adherence, retention and efficacy of other interventions when comparing the sample size required for a factorial design with alternative designs. Figure 1 shows the relative efficiency of the factorial design compared with the three-arm design for estimating the effect of intervention A. The panels represent outcome type (quantitative or binary) and whether B is effective. The panel for quantitative outcome and B ineffective is not shown, as the results are the same as for quantitative outcome and B effective, except that “double non-adherence with A” is the same as the base case. Results for “Double missing with A” are not shown, as they were the same as for the base case. Results for “Perfect” and “Double missing with B” are not shown in some panels where their results were the same as for the base case. The supplemental material gives the numerical results.

Comparing scenarios within panels of Figure 1, we see that missing data with A affects factorial and three-arm designs equally, while missing data with B and non-adherence with B only affect the factorial design. Non-adherence with A affects both designs equally if B is ineffective, but affects the factorial design more if B is effective (because AB non-adherers lose both intervention effects). Doubling controls moves all relative efficiencies towards 100%.

Comparing panels in Figure 1, the main finding is that effectiveness of B on a binary outcome improves efficiency of the factorial design for low interactions. This is because low interactions make the risks smaller and hence the variances smaller; this change in variance does not occur with a quantitative outcome.

In the base case for a quantitative outcome, the factorial design is superior when the interaction is less than 37% of the effect of A in the absence of B (i.e. the A vs. 0 effect). In the other scenarios, this figure of 37% ranges from 24% to 48% (Table 2).

CONCLUSIONS

Many factorial RCTs have been conducted, but they still account for less than 1% of all RCTs, suggesting that there may be scope for their greater use given increased emphasis on efficiency in trial design. We have therefore proposed a number of issues to consider when deciding between factorial and other designs.

Factorial RCTs should be considered as alternatives to multi-arm RCTs, and as extensions to standard 2-arm RCTs, particularly when there may be an opportunity to address additional management questions. A particular strength of factorial designs is assessing multiple different domains of intervention for one condition, where interaction is *a priori* much less likely, and additional scientific questions can be addressed in many cases for free, or nearly for free. The TRACT and REALITY trials are good examples of this kind of approach, assessing multiple different underlying mechanisms to improve outcomes from severe anaemia and HIV[3, 16]. Other examples are the large platform trials REMAP-CAP and SNAP, which have partial factorial interventions in multiple different domains of treatment for pneumonia in ICU (e.g. antibiotic choice, duration, adjunctive macrolides, corticosteroids) and Staphylococcus aureus bacteraemia[12, 13].

Of the issues discussed, interaction is often seen as the largest problem with factorial designs, with consequences for the trial's power and interpretability. At trial design stage, investigators should assess the possible magnitude and likelihood of an interaction. Often, this relies on clinical judgement, since prior data can seldom reliably inform interaction size. We suggest that the impact of plausible interactions is best explored by their consequences for the trial's power, since this can be precisely quantified using the methods described here. If important interaction is plausible then alternative designs should be considered including (in the 2x2 setting) a 2-arm design (ignore one intervention), a 3-arm design (omit the combined intervention), a 4-arm design (using a non-factorial analysis)[9] or variants on the factorial analysis [5, 50].

A finding of unexpected interaction complicates the interpretation and may damage the credibility of a factorial trial. In some cases this will represent genuine complexity which would not have been discovered with a simpler design; in other cases it may be a chance finding. If a *quantitative interaction* is observed, where the estimated effects of A with and without B both suggest benefit from A, but of different magnitude, then in practice the impact on inference may be small. If a *qualitative interaction* is observed[51], where there is some statistical evidence that A has benefit if given without B, but not with B, or even more importantly that it is harmful with B, then this is important knowledge which would never have been uncovered without a factorial design. One of the most relevant factors to consider may be whether the other intervention, B, is already being used in clinical practice. The best approach is to set out a clear statistical

analysis plan which specifies (1) when a factorial analysis will be abandoned: if interactions are a priori unlikely then it may be appropriate to pre-specify a strong statistical significance level (e.g. $p<0.01$) here; and (2) what alternative analysis will be used, taking account of the clinical setting, if a factorial analysis is abandoned.

We showed that a factorial design has greater power than a 3-arm design when the interaction is less than about 30-40% of the main effect. This needs to be assessed at trial design stage. Assessing it from trial results is usually unhelpful (because sampling variation is large) and should not be used to argue retrospectively that a factorial design was inappropriate. Future research could explore empirical evidence about interaction sizes in finished trials, using meta-analysis methods to remove sampling variation.

Limitations of our study include that our issues to consider are derived from our literature review and our experience and not from a Delphi survey. Our sample size comparisons used a broad but by no means exhaustive set of scenarios; slightly better results for the three-arm design could have been achieved by using unequal allocation ratios.

In summary, the issues described should be considered when designing factorial RCTs. This applies both for the primary endpoint and also for secondary endpoints including patient-reported outcomes and cost-effectiveness[25, 26].

DECLARATION OF CONFLICTING INTERESTS

The Authors declare that there is no conflict of interest.

ACKNOWLEDGEMENTS

We thank Brennan Kahan for helpful comments on an earlier version of this paper.

REFERENCES

1. Montgomery AA, Peters TJ and Little P. Design, analysis and presentation of factorial randomised controlled trials. *BMC medical research methodology* 2003; 3: 26. 2003/11/25. DOI: 10.1186/1471-2288-3-26.

2. Joffe JK, Cafferty FH, Murphy L, et al. Imaging Modality and Frequency in Surveillance of Stage I Seminoma Testicular Cancer: Results From a Randomized, Phase III, Noninferiority Trial (TRISST). *Journal of Clinical Oncology* 2022. DOI: 10.1200/JCO.21.01199.

3. Mpoya A, Kiguli S, Olupot-Olupot P, et al. Transfusion and Treatment of severe anaemia in African children (TRACT): A study protocol for a randomised controlled trial. *Trials* 2015; 16: 1-15. DOI: 10.1186/S13063-015-1112-4/TABLES/3.

4. Freidlin B and Korn EL. Two-by-Two Factorial Cancer Treatment Trials: Is Sufficient Attention Being Paid to Possible Interactions? *Journal of the National Cancer Institute* 2017; 109 2017/09/28. DOI: 10.1093/jnci/djx146.

5. Korn EL and Freidlin B. Non-factorial analyses of two-by-two factorial trial designs. *Clinical trials (London, England)* 2016; 13: 651-659. 2016/07/22. DOI: 10.1177/1740774516659472.

6. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *Statistical Principles for Clinical Trials*. 1998. Geneva.

7. James ND, Sydes MR, Clarke NW, et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *The Lancet* 2016; 387: 1163-1177.

8. Sydes MR, Parmar MKB, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: The MRC STAMPEDE trial. *Trials* 2009; 10: 39-39. DOI: 10.1186/1745-6215-10-39.

9. Jaki T and Vasileiou D. Factorial versus multi-arm multi-stage designs for clinical trials with multiple treatments. *Statistics in medicine* 2017; 36: 563-580. 2016/11/03. DOI: 10.1002/sim.7159.

10. Luna J, Jaynes J, Xu HQ, et al. Orthogonal array composite designs for drug combination experiments with applications for tuberculosis. *Statistics in medicine* 2022; 41: 3380-3397.

DOI: 10.1002/sim.9423.

11. Saha S, Brannath W and Bornkam B. Testing multiple dose combinations in clinical trials. *Statistical methods in medical research* 2020; 29: 1799-1817. DOI: 10.1177/0962280219871969.

12. Tong SYC, Mora J, Bowen AC, et al. The Staphylococcus aureus Network Adaptive Platform Trial Protocol: New Tools for an Old Foe. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2022; 75: 2027-2034. DOI: 10.1093/cid/ciac476.

13. Angus DC, Berry S, Lewis RJ, et al. The REMAP-CAP (Randomized Embedded Multifactorial Adaptive Platform for Community-acquired Pneumonia) Study. Rationale and Design. *Ann Am Thorac Soc* 2020; 17: 879-891. DOI: 10.1513/AnnalsATS.202003-192SD.

14. Bielicki JA, Stohr W, Barratt S, et al. Effect of Amoxicillin Dose and Treatment Duration on the Need for Antibiotic Re-treatment in Children With Community-Acquired Pneumonia: The CAP-IT Randomized Clinical Trial. *JAMA* 2021; 326: 1713-1724. DOI: 10.1001/jama.2021.17843.

15. McCabe L, White IR, Chau NVV, et al. The design and statistical aspects of VIETNARMS: a strategic post-licensing trial of multiple oral direct-acting antiviral hepatitis C treatment strategies in Vietnam. *Trials* 2020; 21: 413. 20200518. DOI: 10.1186/s13063-020-04350-x.

16. Hakim J, Musiime V, Szubert AJ, et al. Enhanced Prophylaxis plus Antiretroviral Therapy for Advanced HIV Infection in Africa. *The New England journal of medicine* 2017; 377: 233-245. DOI: 10.1056/NEJMoa1615822.

17. Montgomery AA, Astin MP and Peters TJ. Reporting of factorial trials of complex interventions in community settings: a systematic review. *Trials* 2011; 12: 179. 2011/07/21. DOI: 10.1186/1745-6215-12-179.

18. Julious SA. Seven useful designs. *Pharmaceutical Statistics* 2012; 11: 24-31.

19. Byth K and Gebski V. Factorial designs: a graphical aid for choosing study designs accounting for interaction. *Clinical trials (London, England)* 2004; 1: 315-325. 2005/11/11. DOI: 10.1191/1740774504cn026oa.

20. Kahan BC. Bias in randomised factorial trials. *Statistics in medicine* 2013; 32: 4540-4549. 2013/06/05. DOI: 10.1002/sim.5869.

21. Lin DY, Gong J, Gallo P, et al. Simultaneous inference on treatment effects in survival

studies with factorial designs. *Biometrics* 2016; 72: 1078-1085. 2016/03/19. DOI: 10.1111/biom.12507.

22. Snapinn S. Some remaining challenges regarding multiple endpoints in clinical trials. *Statistics in medicine* 2017; 36: 4441-4445.

23. Bowers M, Stanton L and Thursz M. Design, method and application of stopping rules in a phase III 2x2 factorial randomised controlled trial. *Trials* 2015; 16: P207.

24. Merrill PD and McClure LA. Dichotomizing partial compliance and increased participant burden in factorial designs: the performance of four noncompliance methods. *Trials* 2015; 16: 523. 2015/11/18. DOI: 10.1186/s13063-015-1044-z.

25. Dakin H and Gray A. Economic evaluation of factorial randomised controlled trials: challenges, methods and recommendations. *Statistics in medicine* 2017; 36: 2814-2830. 2017/05/05. DOI: 10.1002/sim.7322.

26. Dakin HA, Gray AM, MacLennan GS, et al. Partial factorial trials: comparing methods for statistical analysis and economic evaluation. *Trials* 2018; 19.

27. Wolbers M, Heemskerk D, Chau TT, et al. Sample size requirements for separating out the effects of combination treatments: randomised controlled trials of combination therapy vs. standard treatment compared to factorial designs for patients with tuberculous meningitis. *Trials* 2011; 12: 26. 2011/02/04. DOI: 10.1186/1745-6215-12-26.

28. Kahan BC, Morris TP, Goulao B, et al. Estimands for factorial trials. *Statistics in medicine* 2022. DOI: 10.1002/sim.9510.

29. Kuhn J, Sheldrick RC, Broder-Fingert S, et al. Simulation and minimization: technical advances for factorial experiments designed to optimize clinical interventions. *BMC medical research methodology* 2019; 19. DOI: 10.1186/s12874-019-0883-9.

30. White IR, Choodari-Oskooei B, Sydes MR, et al. Combining factorial and multi-arm multi-stage platform designs to evaluate multiple interventions efficiently. *Clinical Trials*. DOI: 10.1177/17407745221093577.

31. Green S, Liu PY and O'Sullivan J. Factorial design considerations. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2002; 20: 3424-3430. 2002/08/15. DOI: 10.1200/jco.2002.03.003.

32. Bria E, Di Maio M, Nistico C, et al. Factorial design for randomized clinical trials. *Annals of oncology : official journal of the European Society for Medical Oncology* 2006; 17: 1607-

1608. 2006/06/13. DOI: 10.1093/annonc/mdl106.

33. Iwamoto FM and Lassman AB. Factorial clinical trials: a new approach to phase II neuro-oncology studies. *Neuro-oncology* 2015; 17: 174-176. 2014/12/03. DOI: 10.1093/neuonc/nou314.

34. Whelan DB, Dainty K and Chahal J. Efficient designs: factorial randomized trials. *The Journal of bone and joint surgery American volume* 2012; 94 Suppl 1: 34-38. 2012/08/01. DOI: 10.2106/jbjs.l.00243.

35. Korttila K and Apfel CC. Factorial design provides evidence to guide practice of anaesthesia. *Acta anaesthesiologica Scandinavica* 2005; 49: 927-929. 2005/07/28. DOI: 10.1111/j.1399-6576.2005.00622.x.

36. Collins LM, Dziak JJ, Kugler KC, et al. Factorial experiments: efficient tools for evaluation of intervention components. *American journal of preventive medicine* 2014; 47: 498-504. 2014/08/06. DOI: 10.1016/j.amepre.2014.06.021.

37. Baker TB, Smith SS, Bolt DM, et al. Implementing Clinical Research Using Factorial Designs: A Primer. *Behavior therapy* 2017; 48: 567-580. 2017/06/05. DOI: 10.1016/j.beth.2016.12.005.

38. Pandis N, Walsh T, Polychronopoulou A, et al. Factorial designs: an overview with applications to orthodontic clinical trials. *European journal of orthodontics* 2014; 36: 314-320. 2013/07/26. DOI: 10.1093/ejo/cjt054.

39. Pandis N. Factorial trial. *American journal of orthodontics and dentofacial orthopedics : official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics* 2012; 142: 147-148. 2012/07/04. DOI: 10.1016/j.ajodo.2012.03.020.

40. Esposito M and Nieri M. Editorial: Randomised controlled trials of factorial design alias on how to speed up research on effectiveness of interventions without compromising its validity. *European journal of oral implantology* 2016; 9: 3-4. 2016/03/30.

41. Bangdiwala SI. Factorial experimental designs. *International journal of injury control and safety promotion* 2016; 23: 110-111. 2016/01/13. DOI: 10.1080/17457300.2016.1132583.

42. Krishnan P. When and how to use factorial design in nursing research. *Nurse researcher* 2021; 29: 26-31. 2020/12/04. DOI: 10.7748/nr.2020.e1757.

43. Kahan BC, Tsui M, Jairath V, et al. Reporting of randomized factorial trials was frequently

inadequate. *Journal of clinical epidemiology* 2020; 117: 52-59. 2019/10/05. DOI: 10.1016/j.jclinepi.2019.09.018.

44. Brittain E and Wittes J. Factorial designs in clinical trials: The effects of non-compliance and subadditivity. *Statistics in medicine* 1989; 8: 161-171. DOI: 10.1002/sim.4780080204.

45. White IR and Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? *BMC medical research methodology* 2005; 5: 15-15. DOI: 10.1186/1471-2288-5-15.

46. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in medicine* 2002; 21: 1575-1600. DOI: 10.1002/sim.1188.

47. Molloy SF, White IR, Nunn AJ, et al. Multiplicity adjustments in parallel-group multi-arm trials sharing a control group: Clear guidance is needed. *Contemporary clinical trials* 2021; 113: 106656-106656. DOI: 10.1016/j.cct.2021.106656.

48. Parmar MK, Sydes MR, Cafferty FH, et al. Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. *Clinical trials (London, England)* 2017; 14: 451-461. 20170822. DOI: 10.1177/1740774517725697.

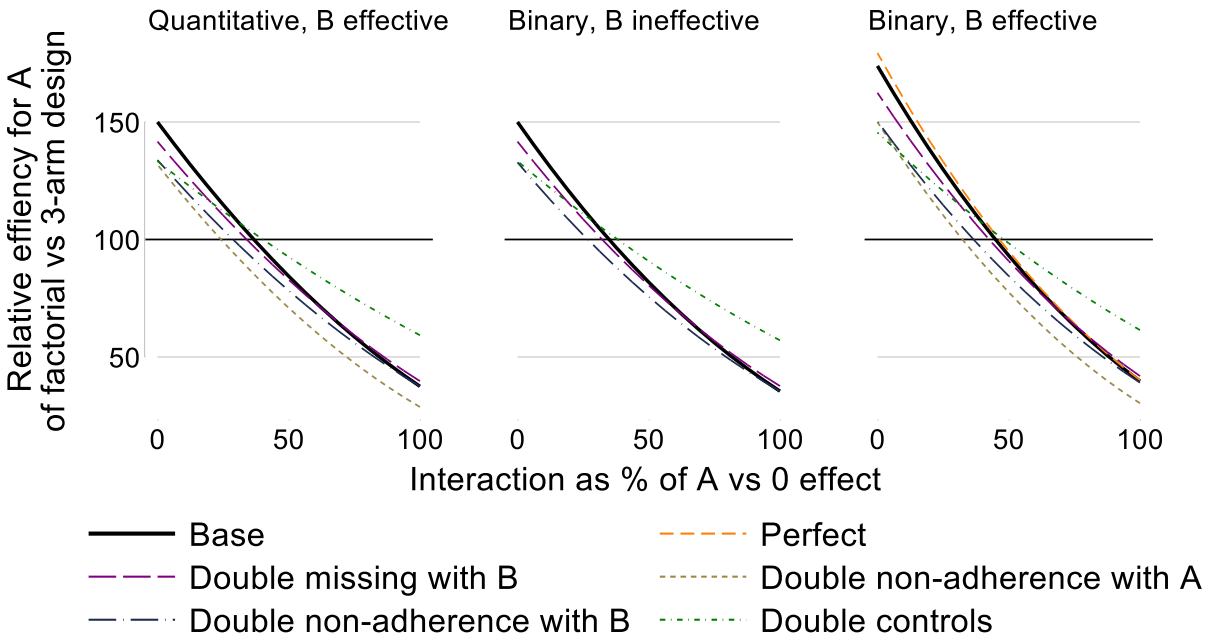
49. Babiker A and Walker AS. Statistical Issues Emerging from Clinical Trials in HIV Infection. In: Geller NL (ed) *Advances in Clinical Trial Biostatistics*. CRC Press, 2003.

50. Leifer ES, Troendle JF, Kolecki A, et al. Joint testing of overall and simple effects for the two-by-two factorial trial design. *Clinical trials (London, England)* 2021; 18: 521-528. 20210818. DOI: 10.1177/17407745211014493.

51. Gail M and Simon R. Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics* 1985; 41. DOI: 10.2307/2530862.

FIGURES

Figure 1. Relative efficiency of factorial vs. three-arm design graphed against interaction, by outcome type (quantitative or binary) and whether B is effective.



TABLES

Table 1: Key issues to consider around factorial designs

Issue to consider	A factorial design is more desirable if
Clinical issues	
Co-administration	Interventions can be easily co-administered.
Safety	Risk of safety issues from co-administration, above individual risks of the separate interventions, is low.
Combination intervention	Safety or clinical data are wanted on the combination intervention.
Interaction (effect modification)	Effect of each intervention is unlikely to be substantially different in the presence of the other intervention(s).
Balancing other interventions	Other interventions are likely to be unbalanced if not randomised
Eligibility criteria	Eligibility criteria for all comparisons are similar.
Practical issues	
Recruitment	Recruitment is not harmed by including many interventions.
Adherence	Each intervention and the toxicities associated with it is unlikely to reduce adherence to (and hence effectiveness of) the other intervention.
Retention	Each intervention is unlikely to reduce overall follow-up.
Blinding	Blinding is easy to implement or not required.
Statistical issues	
Scale of analysis	A suitable scale of analysis can be identified on which interaction is unlikely.
Multiplicity	Adjustment for multiplicity is not required.
Stopping	Adequate ways can be found to stop comparisons early for efficacy or for lack of benefit.
External issues	
Funders	Funding is adequate for the complexity of the design.
Regulators	The trial is not intended for licensing purposes.
Overarching issue	
Sample size	Factorial design gives a lower sample size requirement than alternative designs.

Table 2. Critical value of interaction (expressed as % of A vs. 0 effect on difference scale) above which factorial design becomes less efficient than three-arm design.

Outcome	Quantitative		Binary	
Nature of intervention B	Ineffective	Effective	Ineffective	Effective
Base case	37	37	35	45
Perfect case	37	37	35	46
Double missing with A	37	37	35	45
Double missing with B	34	34	32	43
Double non-adherence with A	37	24	35	32
Double non-adherence with B	29	29	27	37
Double controls	40	40	38	48