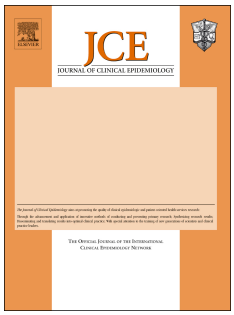


Reporting of randomised factorial trials was frequently inadequate

Brennan C. Kahan, Lecturer, Michael Tsui, Medical Student, Vipul Jairath, Associate Professor, Anna Mae Scott, Assistant Professor, Douglas G. Altman, Professor of Statistics in Medicine, Elaine Beller, Statistician, Diana Elbourne, Professor of Healthcare Evaluation

PII: S0895-4356(19)30430-5

DOI: https://doi.org/10.1016/j.jclinepi.2019.09.018

Reference: JCE 9979

To appear in: *Journal of Clinical Epidemiology*

Received Date: 14 May 2019

Revised Date: 23 August 2019

Accepted Date: 24 September 2019

Please cite this article as: Kahan BC, Tsui M, Jairath V, Scott AM, Altman DG, Beller E, Elbourne D, Reporting of randomised factorial trials was frequently inadequate, *Journal of Clinical Epidemiology* (2019), doi: https://doi.org/10.1016/j.jclinepi.2019.09.018.

Reporting of randomised factorial trials was frequently inadequate

Brennan C Kahan (Lecturer)[1], Michael Tsui (Medical Student)[2], Vipul Jairath (Associate Professor)[3, 4], Anna Mae Scott (Assistant Professor)[5], Douglas G Altman (Professor of Statistics in Medicine)[6], Elaine Beller (Statistician)[5], Diana Elbourne (Professor of Healthcare Evaluation)[7]

Affiliations:
[1] Pragmatic Clinical Trials Unit, Queen Mary University of London, London, UK
[2] Schulich School of Medicine and Dentistry, 1151 Richmond St, London, Ontario, Canada
[3] Department of Medicine, University of Western Ontario, London, Ontario, Canada
[4] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada
[5] Centre for Research in Evidence-Based Practice (CREBP), Bond University, Robina, QLD, Australia
[6] Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[7] Medical Statistics Department, London School of Hygiene & Tropical Medicine, London, UK

*Correspondence to: Brennan Kahan (b.kahan@qmul.ac.uk) (ORCID iD 0000-0001-9957-0844)

Abstract

Objective

Factorial designs can allow efficient evaluation of multiple treatments within a single trial. We evaluated the design, analysis, and reporting in a sample of factorial trials.

Study design and setting

Review of 2x2 factorial trials evaluating health-related interventions and outcomes in humans. Using MEDLINE, we identified articles published between January 2015 and March 2018. We randomly selected 100 articles for inclusion.

Results

Few trials (22%) provided a rationale for using a factorial design. Only 63 trials (63%) assessed the interaction for the primary outcome, and 39/63 (62%) made a further assessment for at least one secondary outcome. 12/63 trials (19%) identified a significant interaction for the primary outcome, and 16/39 trials (41%) for at least one secondary outcome. Inappropriate methods of analysis to protect against potential negative effects from interactions were common, with 18 trials (18%) choosing the analysis method based on a preliminary test for interaction, and 13% (n=10/75) of those conducting a factorial analysis including an interaction term in the model.

Conclusions

Reporting of factorial trials was often suboptimal, and assessment of interactions was poor. Investigators often used inappropriate methods of analysis to try to protect against adverse effects of interactions.

Key words: clinical trial, randomised trial, factorial trial, interaction

What is new

Key findings

- Interactions were common, with 12/63 trials (19%) identifying a significant interaction for the primary outcome, and 16/39 trials (41%) for at least one secondary outcome.
- Evaluation of interactions was often suboptimal, with only 63 trials (63%) assessing the interaction for the primary outcome, and 39/63 (62%) for at least one secondary outcome
- Few trials evaluated the size of the interaction or provided a measure of uncertainty

What this adds to what is known

- Investigators often used inappropriate methods of analysis to protect against potential interactions, with 18% of trials choosing the analysis method based on a preliminary test for interaction, and 13% (n=10/75) of those conducting a factorial analysis including an interaction term in the model

What is the implication, what should change now

- Improvement in the analysis and reporting of factorial trials is required to allow valid conclusions to be drawn when this design is used

Introduction

Factorial trials allow investigators to assess multiple interventions within a single trial without increasing the sample size, provided the treatments work independently [1-3]. For example, the ISIS-2 trial [4] allocated 17,187 patients with suspected acute myocardial infarction to streptokinase, aspirin, both streptokinase and aspirin, or double-placebo. Streptokinase and aspirin were thought to work through different mechanisms (with aspirin preventing clots, and streptokinase dissolving clots). The investigators conducted a 'factorial' (or 'at the margins') analysis, which compared all patients allocated to streptokinase (streptokinase alone + streptokinase and aspirin) vs. all those not allocated to streptokinase (aspirin alone + double-placebo), and a similar analysis was conducted for aspirin. They found that streptokinase and aspirin independently reduced 5-week vascular mortality. Assessing these interventions in separate trials of the same sample size likely would have been unfeasible.

The factorial analysis used in the ISIS-2 trial assumes that there is no interaction between treatments. In other words, the effect of aspirin is the same regardless of whether the patient also received streptokinase, and vice versa. When treatments do interact, results from a factorial analysis can be misleading [1-3, 5-10]. When treatments are known to interact, a 'multi-arm' (or 'inside-the-table') analysis is more appropriate (table 1). For example, in ISIS-2, this would involve separately comparing each of the three active treatment groups (streptokinase alone, aspirin alone, and streptokinase + aspirin) against double-placebo. This analysis is valid even when treatments interact, however it requires a larger overall sample size to achieve the same power as a factorial analysis.

When efficiency is the main aim, it is generally recommended that factorial trials are undertaken only when no interaction is expected (for instance, if treatments are expected to work through different mechanisms or target different endpoints). However, it is often difficult to rule out interactions entirely at the design stage, and so if a factorial analysis is used, some assessment of the interaction between treatments is warranted. This could involve assessing the magnitude of the interaction effect (along with a measure of uncertainty, such as a confidence interval), and conducting a multi-arm analysis as a sensitivity analysis to evaluate to what extent results from the factorial analysis may be affected by an interaction. However, previous research has shown that this is infrequently done [1, 2, 5]. Furthermore, investigators may sometimes use inappropriate analytical methods in an attempt to guard against the adverse consequences of an interaction, such as choosing the method of analysis based on results from an interaction test [6], or conducting a factorial analysis which includes a term for the interaction between treatments in the model (table 2).

We conducted a review of recently published factorial trials to evaluate whether the design, analysis, and reporting of trials was appropriate, and to identify the prevalence of reported interaction.

Methods

Data source and search methods

We searched Medline from inception to March 31st, 2018 (search performed April 2018) for published reports of randomised trials utilising a factorial design. The full search strategy is available in the supplementary file. One author then screened titles and abstracts to assess eligibility. Subsequently the full text was assessed for eligibility by two reviewers. The full eligibility criteria are available in the supplementary file. Briefly, articles were eligible if they were the main trial publication from a 2x2 factorial trial conducted in humans with health-related interventions and outcomes, published between January 2015 and March 2018 inclusive. We excluded trials with health economic outcomes only, laboratory or in vitro studies, dose response or dose finding studies, trials employing a crossover or N-of-1 design, educational interventions or interventions aimed at health practitioners rather than patients, studies where factors were not randomly allocated, or letters or commentaries/editorials.

After eligibility screening, we randomly selected 100 trials from the final set of eligible trials for data extraction. This was done by using a random number generator to sort eligible articles into a random order and selecting the first 100 articles. We specified this number (n=100) prior to conducting the search, as this sample size was likely to enable us to identify any major deficiencies around the design, analysis, or reporting.

Data extraction

All articles were extracted independently by two separate authors onto a pre-piloted data extraction form, and discrepancies were resolved by discussion. When discrepancies could not be resolved, a third author acted as arbitrator. During the data extraction period, we updated the data extraction form twice, either to clarify questions or to add additional options for certain questions. When necessary, we went back and updated previously extracted results to ensure they were consistent with the updated form. The final version of the data extraction form is available in the supplementary file.

We extracted information on the design, analysis, and reporting aspects specific to factorial trials. These included the rationale for utilising a factorial design, whether separate primary outcomes were specified for each factor, analysis approach used for the sample size calculation, randomisation approach, description of blinding, patient flow, method of statistical analysis, assessment of interactions, and reporting of descriptive statistics. Some data extraction items related to the primary outcome. We pre-specified the following strategy to identify a single primary outcome for extraction: (1) if only one outcome was listed as being primary, we used this; (2) if either no outcomes or multiple outcomes were listed as being primary, we used the outcome that was used in the sample size calculation; and (3) if no sample size calculation was performed, we used the first clinical outcome listed in the Objectives or Outcomes section of the article, or the first clinical outcome listed in the Results section of the article if there was no Objectives/Outcomes section.

We assessed additional online supplementary material only when this was specifically referred to within the main text as containing additional information on the item in question. We did not refer to other published papers (e.g. published protocols, secondary papers), as we wished to evaluate reporting of the paper as a standalone item.

We summarised results descriptively. All analyses were conducted using Stata v15.1.

Results

Our initial search found 3200 citations (figure 1). After removing duplicates and those not meeting the inclusion/exclusion criteria, 122 eligible articles remained. We randomly selected 100 of these for inclusion in this review. Two of the selected articles reported results for different factors from the same trial; we included these in the review as two separate articles, and evaluated reporting in each based on the standalone article. Design characteristics of trials are provided in table 3.

Quality of reporting

Quality of reporting is shown in table S1. Only 22% of trials (22/100) provided an explicit rationale for why they chose to use a factorial design. Only 25% of trials (25/100) reported the number of patients allocated to each factorial group, although 84% (84/100) did report the number allocated to each multi-arm group.

Over a quarter of trials (28%, 20/71) did not make clear whether their sample size calculation was based on a factorial or multi-arm analysis approach. The analysis approach was generally well reported, however of the 75 trials using a factorial analysis, 9 (12%) did not report whether they adjusted for the other factor in the analysis, and 21 (28%) did not report sufficient detail to allow readers to judge whether investigators included an interaction term in the analysis.

In the abstract, 24% of trials (24/100) did not report whether results were based on a factorial or multi-arm analysis, and 14% of trials (14/100) reported a treatment effect for only one of the two factors.

Assessment of interaction

The presence of interactions was poorly reported for the primary outcome (table 4); 37% of trials (37/100) made no mention of interactions, and most trials provided only a p-value or simply stated whether an interaction was present or not. Only 12% of trials (12/100) presented an estimate of the size of the interaction, with only 3% (3/100) providing a confidence interval for the estimate. Interactions for the primary outcome were also poorly reported in abstracts, with only 22% of trials (22/100) discussing interactions (11 trials provided a p-value, 11 stated whether an interaction was present). Of the trials which made an assessment of the interaction for the primary outcome, 39/63 (62%) further assessed the presence of interactions for secondary outcomes.

Of those trials reporting interactions, almost one fifth (n=12/63, 19%) found a statistically significant interaction for the primary outcome (at the 5% level), and 41% (n=16/39) found a significant interaction for at least one secondary outcome. In 33% of trials (n=13/39) there was an inconsistency between the primary and secondary outcomes in terms of whether a statistically significant interaction was observed (i.e. a significant interaction was observed for the primary outcome but not for all secondary outcomes, or vice versa).

Additional subgroup results are available in tables S4-S6 in the supplementary material.

Analysis of primary outcome

Most trials (51%, 51/100) used a factorial analysis as their primary analysis approach, and 23% (23/100) used a multi-arm analysis (table 5). Only 26% of trials (26/100) presented analysis results

from both a factorial and multi-arm analysis to allow comparison between the two approaches. Most trials presented descriptive statistics of the primary outcome by multi-arm groups only (38%, 38/100), with 19% (19/100) presenting only by factorial groups, and 27% (27/100) presenting for both multi-arm and factorial groups (table S3).

A number of trials used inappropriate methods of analysis to guard against the effects of an interaction; 18% of trials (18/100) chose their primary method of analysis on the basis of an interaction test, thereby introducing bias into the results. Furthermore, 13% of trials (10/75) which used a factorial analysis did so by including an interaction term in the model, thereby inadvertently reducing the sample size for each comparison by 50%, which may have led to an underpowered analysis.

Discussion

In this review of 100 articles reporting results from factorial trials, we found that reporting was often inadequate, making it difficult to ascertain the validity of results. For instance, few trials explicitly reported why they had chosen a factorial design, and in over a quarter of cases it was not clear which analysis approach the sample size calculation was based on. Most notably, assessment of interactions was poor. We found that 37% of trials (37/100) did not assess the presence of an interaction for their primary outcome. Those that did assess the interaction often did so poorly, by only presenting a p-value, or stating whether the interaction was significant. This approach is problematic, as trials are typically underpowered to detect clinically relevant interactions, and so this approach may falsely reassure investigators that results are robust. A preferable approach would be to present the size of the estimated interaction along with a measure of uncertainty, such as a confidence interval. Furthermore, we found that most trials that used a factorial analysis did not also present results from a multi-arm analysis to allow comparison between the two.

Many trials assessed the interaction for the primary outcome, but not for any secondary outcomes. This may be based on the mistaken belief that interactions amongst secondary outcomes are less important. However, this is incorrect, as interactions amongst secondary outcomes have the same implications for bias. Alternatively, investigators may believe that ruling out an interaction for the primary outcome means it is ruled out for the entire trial. We found this was generally not the case; for many trials, treatments were found to interact for some, but not all, outcomes. In practice, there may often be scientific reasons why treatments are more likely to interact for some outcomes but not others (e.g. to interact for efficacy outcomes but not safety outcomes, or vice versa).This highlights the need to assess interactions and perform appropriate sensitivity analyses for all outcomes, rather than solely for the primary outcome.

Statistically significant interactions were relatively common, with 19% of trials (12/63) reporting a significant interaction for their primary outcome, and 41% of trials (16/39) reporting a significant interaction for at least one secondary outcome. These results may be in part driven by multiplicity, as the more interactions that are reported, the more likely that some will be statistically significant just by chance. Furthermore, some of the reported interactions may have been statistical interactions based on the chosen analysis scale, rather than true biological interactions. For instance, an interaction may be observed on the risk difference scale even if treatment effects are constant on the risk ratio scale. However, these interactions still pose challenges in the analysis and interpretation of such trials, regardless of whether they are true biological interactions.

We found that investigators often used inappropriate methods to try to account for possible interactions. Almost a fifth of trials chose their method of analysis based on the results of a preliminary test for interaction. This figure may however be an underestimate, as it is possible that some trials used this approach without explicitly reporting it. This approach introduces bias into estimated treatment effects, and should not be used [6]. Furthermore, 13% of trials (10/75) included an interaction term in the model when implementing a factorial analysis. This approach inadvertently excludes 50% of the sample size from each comparison, and thus loses all the efficiency benefits from a factorial design. There are different approaches to dealing with potential interactions when a factorial design has been chosen for efficiency based on an a priori assumption of no interaction. Our preferred approach would be to evaluate the likely size of interaction through its estimated effect and 95% confidence interval, and, alongside the factorial analysis, to also include results from a multi-arm analysis (which does not depend on assumptions about no interaction) to allow comparison between the two analysis approaches. Sensitivity analyses which to evaluate to

what extent conclusions from a factorial analysis may be affected under different plausible assumptions about the size of the interaction could also be useful [11, 12].

The issues surrounding the design, analysis, and reporting often made it difficult to determine whether appropriate methods had been used, and whether results were at risk of bias. Guidance for reporting results from clinical trials are available in the CONSORT statement [13], with extensions available for several unique trial designs, such as cluster [14], non-inferiority [15], pragmatic [16], N-of-1 [17], pilot and feasibility [18], and within person trials [19]. Given the unique features of factorial trials, reporting guidelines for such designs are warranted.

Our results are similar to earlier evaluations of factorial trials, indicating that reporting has not improved over time [1, 2, 5]. McAlister *et al* [1] reviewed 44 trials with clinically important binary outcomes, with most trials (66%) being myocardial ischemia trials; Montgomery *et al* [2] reviewed 76 trials which evaluated complex interventions in community settings; and Freidlin and Korn [5] reviewed 30 oncology trials. The main area of difference was that McAlister *et al* found only 6% of interactions were statistically significant, which is substantially lower than the 19% we found for primary outcomes. This discrepancy may be in part due to the differences between samples, as we included trials regardless of outcome type or clinical area; it is possible that treatments used in myocardial ischemia trials or with clinically important binary outcomes are less likely to interact. Of note, our results are in line with those from Freidlin and Korn, who found 17% of trials in their review reported an interaction.

There were some limitations to our review. Only one author screened titles and abstracts for eligibility, and so some eligible articles may have erroneously been excluded. We only included trials published in journals indexed in Medline; trials published in other journals may be reported and conducted differently. It is likely that had we restricted our search to high impact-factor journals, our results might have shown better reporting. However, this is not guaranteed; some high impact-factor journals have policies of publishing separate articles for each factor (i.e. one article describing results for treatment A, and a separate article for treatment B), which typically leads to poor reporting of many aspects, notably around the interaction. Our review only included 2x2 factorial trials; higher order designs have additional issues. Finally, poor reporting in many articles made it difficult to assess whether appropriate methods had been used.

Conclusions

In this review of published 2x2 factorial trials, reporting around many aspects was inadequate, and often meant that it was impossible to assess whether main results were valid. Interactions were relatively common, and many trials did not appropriately assess whether interactions may have affected the validity of their results. Furthermore, many trials used inappropriate analysis methods to attempt to combat the effects of interactions.

Figure 1: Flow diagram of factorial trials

Table 1: Hypothetical 2x2 factorial trial

		Treatment B		*Margin*
		Yes	**No**	
Treatment A	**Yes**	Both A and B (multi-arm group: both A and B)	A alone (multi-arm group: A alone)	*All A* *(factorial group: treatment A)*
	No	B alone (multi-arm group: B alone)	Neither A nor B (multi-arm group: double-control)	*All non-A* *(factorial group: control A)*
Margin		*All B* *(factorial group: treatment B)*	*All non-B* *(factorial group: control B)*	

Table 2: Potential pitfalls in the analysis of factorial trials

Problem	**Explanation**
Inadequate assessment of interaction	Factorial analyses can be misleading when treatments interact; therefore, evaluation of the interaction is important to assess whether results from a factorial analysis are likely to be valid [1-3, 6]. Often interactions are assessed using a significance test, with a p-value > 0.05 indicating there is no interaction. This approach is problematic, as this test has very low power to detect true interactions, and so will usually give false reassurance. A preferable approach is to present the size of the interaction term with a measure of uncertainty (for instance, a 95% confidence interval).
Choosing the final analysis approach (factorial vs. multi-arm) based on a test for interaction (two-stage approach.	This approach involves performing a preliminary test for interaction. If the interaction is not statistically significant, a factorial analysis is used; if it is significant, a multi-arm analysis is used. This approach is not advisable, as it introduces bias into estimated treatment effects [6]. This bias occurs for two reasons. First, the preliminary test has low power to detect interactions, and so often leads to factorial analyses even for moderate or large interactions. Second, the interaction test is correlated with the size of treatment estimates from a multi-arm analysis, meaning that using these estimates only when the test is significant will lead to estimates that are too large [6].
Adjustment for the interaction term in the model when conducting a factorial analysis	It is unclear what the rationale behind this approach is, but it may be to try and maintain the efficiency of a factorial analysis whilst accounting for potential interactions. However, this approach is flawed, as it does not maintain efficiency. It is in fact identical to a multi-arm analysis using different parametrisation, and so involves inadvertently discarding 50% of the sample size from each comparison, leading to substantial losses in power and precision.

Table 3 – Design characteristics

	Trials (n=100)
Rationale for using factorial design	
Rationale stated for using factorial design	22 (22)
Rationale	
Efficiency	3
Assess presence of interaction	10
Both efficiency and interaction	3
Other	3
Unclear	3
Stated why no interaction expected (trials where efficiency was only aim)	0/3
Sample size	
Sample size calculation reported	
No*	29 (29)
Yes	71 (71)
Sample size calculation based on:	
Factorial analysis	30/71 (42)
Multi-arm analysis	21/71 (30)
Unclear	20/71 (28)
Interaction assumed in sample size calculation	
Yes	5/71 (7)
No	58/71 (82)
Unclear	8/71 (11)
Randomisation	
Randomisation approach	
Combination of factors	84 (84)
Separate factors	5 (5)
Unclear	11 (11)
Randomisation to different factors done all at once	
Yes	81 (81)
No	7 (7)
Unclear	12 (12)
All participants randomised to all factors	
Yes	88 (88)
No	8 (8)
Unclear	4 (4)
Blinding	
Description of blinding	
Described separately for each factor	32 (32)
Described generally, but not specific to each factor	48 (48)
Blinding not mentioned	19 (19)
Unclear	1 (1)

*No sample size calculation reported (n=23), article stated no sample size calculation performed (n=6)

Table 4 – Assessment of interaction

	Trials (n=100)
Interaction assessment in the main text	
Any interaction results presented for primary outcome?	
No	37 (37)
Yes	63 (63)
Reported effect size of interaction	12 (12)
Reported confidence interval for effect size of interaction	3 (3)
Reported p-value for interaction	44 (44)
Made statement that there was/was not an interaction present, but presented no other information	18 (18)
Any interaction results presented for any secondary outcomes?*	
No	61/63 (38)
Yes	39/63 (62)
Interaction assessment in the abstract	
Any interaction results presented for primary outcome?	
No	78 (78)
Yes	22 (22)
Reported effect size of interaction	0 (0)
Reported confidence interval for effect size of interaction	0 (0)
Reported p-value for interaction	11 (11)
Made statement that there was/was not an interaction present, but presented no other information	11 (11)
Statistical significance of interactions	
Interaction for primary outcome statistically significant at 5% level*	12/63 (19)
Interaction for at least one secondary outcome statistically significant at 5% level**	16/39 (41)
Discrepancy in statistical significance of interactions between primary and at least one secondary outcome (e.g. interaction for primary outcome significant, but at least one secondary outcome had non-significant interaction, or vice versa)**	13/39 (33)

*This data summary is limited to the subset of trials which presented interaction results for the primary outcome

**These data summaries are limited to the subset of trials which presented interaction results for both the primary and at least some secondary outcomes

Table 5 – Analysis approach

	Trials (n=100)
Analysis approach for primary outcome	
Factorial	51 (51)
Multi-arm	23 (23)
Chosen based on results from an interaction test (two-stage approach)	18 (18)
Both factorial and multi-arm analyses presented, unclear which is primary	5 (5)
Unclear	3 (3)
Analysis results presented from:	
Factorial analysis only	49 (49)
Multi-arm analysis only	22 (22)
Both sets	26 (26)
Unclear which set presented	3 (3)
Method of analysis reported in abstract	
Factorial	46 (46)
Multi-arm	25 (25)
Both methods	5 (5)
Unclear	24 (24)
Factorial analysis adjusted for other factor	
Yes	52/75 (69)
No	14/75 (19)
Unclear	9/75 (12)
Factorial analysis appropriately conducted (did not include interaction term)	
Yes	44/75 (59)
No	10/75 (13)
Unclear	21/75 (28)
Multi-arm analysis adjusted for multiple testing	
Yes	11/48 (23)
No	36/48 (75)
Unclear	1/48 (2)

Contributors

BCK, DAG, EB, and DE contributed to the concept, designed the study, and designed data extraction forms. EB performed the search and identified eligible trials. BCK, MT, VJ, AMS, EB, and DE extracted data. BCK analysed data and wrote the first draft of the manuscript, and is the guarantor. MT, VJ, AMS, EB, and DE contributed to the manuscript. DAG died in June 2018 and was therefore not able to contribute to the later stages of this review. The corresponding author attests that all other listed authors meet all authorship criteria and no others meeting the criteria have been omitted.

Competing interests

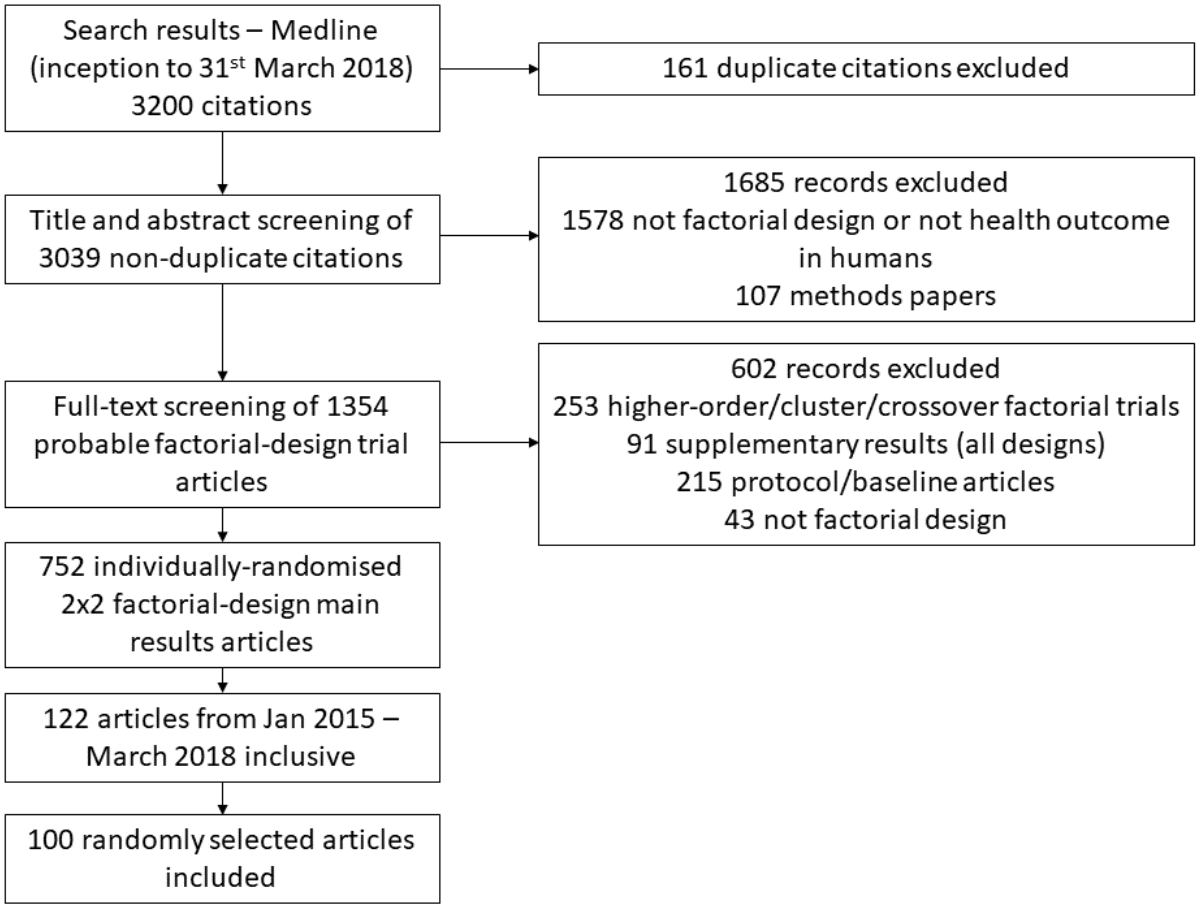
None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA : the journal of the American Medical Association. 2003;289(19):2545-53.
2. Montgomery AA, Astin MP, Peters TJ. Reporting of factorial trials of complex interventions in community settings: a systematic review. Trials. 2011;12:179.
3. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC medical research methodology. 2003;3:26.
4. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Lancet. 1988;2(8607):349-60.
5. Freidlin B, 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(9).
6. Kahan BC. Bias in randomised factorial trials. Statistics in medicine. 2013;32(26):4540-9.
7. Lubsen J, Pocock SJ. Factorial trials in cardiology: pros and cons. Eur Heart J. 1994;15(5):585-8.
8. Byth K, Gebski V. Factorial designs: a graphical aid for choosing study designs accounting for interaction. Clinical trials (London, England). 2004;1(3):315-25.
9. Foley RN. Analysis of randomized controlled clinical trials. Methods in molecular biology (Clifton, NJ). 2009;473:113-26.
10. Korn EL, Freidlin B. Non-factorial analyses of two-by-two factorial trial designs. Clinical trials (London, England). 2016.
11. Dakin H, Gray A. Economic evaluation of factorial randomised controlled trials: challenges, methods and recommendations. Statistics in medicine. 2017;36(18):2814-30.
12. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. BMC medical research methodology. 2014;14:11.
13. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS medicine. 2010;7(3):e1000251.
14. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.
15. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA : the journal of the American Medical Association. 2012;308(24):2594-604.
16. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337:a2390.
17. Vohra S, Shamseer L, Sampson M, Bukutu C, Schmid CH, Tate R, et al. CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. BMJ. 2015;350:h1738.
18. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ. 2016;355:i5239.
19. Pandis N, Chung B, Scherer RW, Elbourne D, Altman DG. CONSORT 2010 statement: extension checklist for reporting within person randomised trials. BMJ. 2017;357:j2835.


Search results – Medline (inception to 31st March 2018) 3200 citations
161 duplicate citations excluded
Title and abstract screening of 3039 non-duplicate citations
1685 records excluded
1578 not factorial design or not health outcome in humans
107 methods papers
Full-text screening of 1354 probable factorial-design trial articles
602 records excluded
253 higher-order/cluster/crossover factorial trials
91 supplementary results (all designs)
215 protocol/baseline articles
43 not factorial design
752 individually-randomised 2x2 factorial-design main results articles
122 articles from Jan 2015 – March 2018 inclusive
100 randomly selected articles included