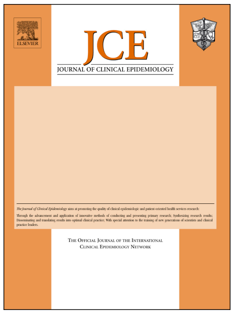

The split-plot design was useful for evaluating complex, multi-level interventions but there is need for improvement in its design and report

Beatriz Goulão, Graeme MacLennan, Craig Ramsay

PII: S0895-4356(17)30401-8
DOI: 10.1016/j.jclinepi.2017.10.019
Reference: JCE 9515

To appear in: *Journal of Clinical Epidemiology*

Received Date: 14 April 2017
Revised Date: 20 September 2017
Accepted Date: 30 October 2017

The split-plot design was useful for evaluating complex, multi-level interventions but there is need for improvement in its design and report

Beatriz Goulão[a], Graeme MacLennan[a], Craig Ramsay[a]

[a] Health Services Research Unit, University of Aberdeen (address below)

Corresponding author:

Beatriz Goulão

Health Services Research Unit

3rd Floor

University of Aberdeen

Health Sciences Building

Foresterhill

Aberdeen

AB25 2ZD

UK

Email: beatriz.goulao@abdn.ac.uk

Tel: +44 1224 438086

Fax: +44 1224 438165

Abstract

Objective: To describe the sample size calculation, analysis and reporting of split-plot randomised controlled trials (RCT) in healthcare (trials that use two units of randomisation: one at a cluster-level and one at a level lower than the cluster).

Study design and setting: We carried out a comprehensive search in the EMBASE database from 1946 to 2016. Healthcare trials with a split-plot design in human subjects were included. Three authors screened and assessed the studies and data were extracted on methodology and reporting standards based upon CONSORT.

Results: 18 split-plot studies were included, with authors using nine different designations to describe them. Units of randomisation were unclear in nine abstracts. Explicit rationale for choosing the design was not givenTen studies presented a sample size calculation accounting for clustering; the analyses were coherent with that. Flow of participant diagrams were presented but incomplete in 14 articles.

Conclusion: Split-plot designs can be useful complex designs, but challenging to report. Researchers need to clearly describe the rationale, sample size calculation and participant flow. We provide a suggested CONSORT style participant flow diagram to aid reporting. There is need for more research regarding sample size calculation for split-plots.

Word count: 193/200

1. Introduction

Randomised controlled trials (RCT) usually allocate individual participants to each of the interventions under evaluation (referred to here as a 'standard RCT') [1]. The standard RCT may be replaced with a factorial RCT when researchers are interested in testing the effects of two or more interventions on the same outcome. In a full factorial RCT every participant is randomised to receive (or not) every intervention. [2] Factorial RCTs can be statistically efficient compared with standard RCTs, assuming no interaction effects between the interventions.[3]

Some interventions are not suitable for evaluation using a standard RCT and there is a need to randomise groups of individuals (cluster randomised trial, C-RCT), for example: wards, hospitals, or communities. Cluster randomised trials are usually used either because it's not feasible to randomise the individual or to avoid intervention contamination (the unintentional spill-over of intervention effects from one treatment group to another) [4].

The factorial RCT can be adapted to involve randomising clusters of individuals. Factorial C-RCTs have been increasingly promoted due to their potential efficiency. [2,5] If the interventions target the same group of individuals (i.e. cluster) or must for ethical or administrative reasons be randomised at the cluster level, a factorial cluster RCT might have all interventions randomised at the same level. However, if the interventions target different levels (e.g. schools and teachers; hospitals and doctors; work sites and managers) a design which randomises one of the interventions at the cluster level and the other at the participant level could be useful. There are also practical issues that might dictate the need for two units of randomisation: for example, the costs of adding clusters might be superior to the costs of adding participants within clusters. [6] This type of design is called a *split-plot* (S-P) design. As an example, the IQuaD trial randomised dental practices to provide either routine or personalised oral hygiene advice to all of their dental patients. Each dental patient within a practice was also randomly allocated to receive either none, 12-monthly or 6-monthly scale and polishes.[7]

Split-plot designs have been used in the healthcare literature and can be a useful design when evaluating two treatments at different levels, however to our knowledge there has been no systematic evaluation of the methodology used in the reporting and analysis of S-P designs in healthcare research. Doing so can identify potential limitations in the methodology of these studies, contributing for their improvement, as well as highlight the existence and purpose of this design. The aim of this study was to review the use of S-P designs conducted in healthcare with a special focus on relevant CONSORT items, [4,8] the sample size calculation and statistical methods used in the analysis.

2. Methods

The study protocol is available from the authors.

2.1. Data sources

We searched studies published from January 1946 to March 2016 and listed on the EMBASE database. We did not screen abstracts that had already been screened and identified by an existing review of factorial C-RCTs that identified seven split-plot designs from 1946 to 2012.[5] The search strategy was designed to identify factorial and cluster designs and RCTs, and was limited to English, Portuguese and Spanish language studies. The search strategy is shown in **Appendix A**.

2.2. Eligibility criteria

We included primary publications of RCTs conducted in human subjects. We excluded secondary analyses, economic evaluation, pilot studies and protocols. We defined an S-P as a design with two units of randomisation: one at the cluster-level and another within the cluster; if the second randomisation was within the intervention cluster only, we called this a partial S-P; when outcome collection was at a lower level than the randomisations we called this a multi-stratum S-P; we defined a nested split-plot as a split-plot nested within a larger trial. **Appendix B** includes diagrams of the different types of split-plot design.

2.3. Study selection

BG conducted the searches. We quality checked 20 randomly selected abstracts. Agreement between three researchers (BG, GM and CR) was classified as the need to assess full text or not. If agreement between reviewers was greater than 80%, BG would proceed with the retrieval of full texts. BG assessed full texts for inclusion in the final review. GM assessed 10% of the same full-texts. More than 80% of agreement was needed to proceed. A third reviewer (CR) arbitrated on disagreement. The data were extracted by BG and included: author, journal, year of publication, title, goal of the study, design, primary outcomes, sample size calculation, cluster and participant-level intervention, number of clusters and participants, rationale of the design, statistical analysis, report of diagram for flow of participants, and units of randomisation identified in the abstract. To quality assure data entry, we randomly selected 10% of included studies for double data extraction. Where there was disagreement a consensus about the appropriate approaches was sought and reached through discussion and those approaches were applied to the remaining studies. If the disagreement rate was higher than 10%, we would consider all the studies.

2.4. Guidance on the report of S-P designs

Using the cluster [4] and standard [8] CONSORT statements we identified, through a consensus discussion between the authors, topics from the checklists that deserved special consideration when reporting an S-P design. We identified the following: title and abstract, introduction (rationale), sample size, statistical methods, results (participant flow). We provide comment and guidance on the report of these specific topics.

2.5. Case study: IQuaD

To introduce the S-P design in a more detailed way, we will describe IQuaD as a case study [7]. IQuaD was a multi-level factorial cluster randomised trial, multi-centre with blinded outcome evaluation based in dental primary care in Scotland and North East of England. Clinicians recruited 1,860 adult patients. Dental practices were cluster randomised to provide routine oral hygiene advice or personalised. To test the effects of periodontal instrumentation (scale and polish, SP) each individual patient participant was randomised to one of three groups: no SP, 6-monthly SP (current practice) and 12-monthly SP. The primary objectives of IQuaD were to test personalised OHA versus routine OHA and no SP versus 6-monhtly SP. It was assumed there wouldn't be a "substantive interaction between the SP interventions and the personalised OHA". IQuaD had two primary outcomes: gingival inflammation / bleeding on probing at the gingival margin at 3 year follow-up and oral hygiene self-efficacy at 3 year follow-up.

3. Results

3.1. Literature search and study selection

The agreement for retrieval of full-texts was 90%. After obtaining the full-texts, agreement on whether those were split-plot designs was 100%.

From 8,245 abstracts found, 154 full texts were assessed and from those 136 were excluded (Figure 2). The most common reasons for exclusion were cluster randomised trials with no factorial element (n=63) and cluster factorial designs with one randomisation unit (n=47).

>> insert Figure 1 – PRISMA diagram<<

3.2. Characteristics of included studies

Eighteen studies were included in the review: eleven were traditional S-P designs, four were partial, two were multi-stratum and one was a nested S-P design.

The included studies were published between 2002 and 2016 (Table S1, Supplemental material). Researchers used nine different designations to describe the S-P design. Nine studies referred clearly two different randomisation units in the abstract. The remaining studies were unclear, using expressions such as "subtrial" [9], "embedded" [10] , "patient-level" and "community-level" intervention [11,12]. Article titles had incomplete information about the design: four didn't mention randomisation; four stated "cluster randomised"; no study mentioned the factorial design.

At the cluster-level, the units of randomisation were using health related units (n=8), geographical clusters (n=4), health professionals (n=4), nursing homes (n=1) and schools (n=1). The number of clusters included varied from 6 to 91. At the participant-level, the majority of the studies used individuals as the unit of randomisation (n=11).

Many interventions at the cluster-level were education related, for example: osteoporosis' workshops, [12] school education interventions [9] and training in shoulder disorders. [13] At the participant-level, the types of interventions included tailored letters [14] or e-mails [12], timing of intervention [15] or physiotherapy sessions.[16] Two studies evaluated a drug intervention at the participant level. [17,18]

Some studies did not state a reason for choosing the design (n=7), others stated that there was an interest in assessing the interaction between interventions (n=5) or to avoid contamination by using cluster randomisation (n=4). No explicit justification is given by any of the trials to randomise two units, but one study hypothesized that interventions addressing multiple levels of a structure (such as individual, community, policy) might be more effective at changing behaviour than those focusing on a single level. [12]

Half of the studies included (n=9) had a primary aim related to implementation of guideline interventions (such as promoting colorectal cancer screening in primary care), followed by health (n=5) and public health interventions (n=2).

3.3. ***Sample size calculation and analysis***

Twelve studies presented a sample size calculation (two as a *post hoc* power calculation) (Table 2). In all of them, except two, an intraclass correlation or inflation factor was used to account for the clustered nature of the data. From those, four studies were unclear about the comparison level used to calculate the sample size[10,15,19,20], one study used a cluster level comparison only[21], one study used a participant level comparison only[22], two studies presented both comparison levels[9,23] and two studies used a clinical meaningful difference to calculate their sample size[16,18]. Of the two studies that did not use an inflation factor, one used a participant level comparison[17] and the other was unclear about the comparison level[24]. No study based its sample size calculation on the interaction between the two interventions. All studies but one [17] ignored the interaction between the interventions in the sample size calculation.

The analyses used in S-P designs were in general consistent with the sample size calculation. Hierarchical models and generalised estimating equations were the most commonly used approaches to account for clustering. Five studies used separate models to analyse the cluster and individual components of the studies - three studies included the intervention at the cluster-level in a multilevel model and analysed the participant-level with a simpler model without adjustment for clustering. [11,16,23,25], the other studies used a cluster-level analysis. [21,25]

The interaction result was reported in most of the papers (9 out of the 14 eligible studies).

[1] Partial split-plots do not have a factorial design and therefore cannot present an interaction result

>>insert Table 2 – Sample size calculation, analysis and interaction considerations<<

3.4. *Flow of participants' diagram*

The majority of the trials had a flow of participants' diagram (15/18) that focused on both the cluster and participant-level interventions (13/15) but all studies, except one, had incomplete information. The description of who or what was randomised and how many units were randomised were the most common information reported in the diagrams, whereas the number of clusters or participants lost to follow-up were the most commonly missed (Table 3).

>> insert Table 3 Quality of information presented regarding the flow of participants<<

3.5. CONSORT diagram suggestion

We developed a diagram that incorporates the CONSORT suggestions for C-RCT [4] and individually RCTs. The diagram presents information about the number of clusters assessed and randomised, the average cluster size and a measure of variation, and then how many participants were assessed, randomised and analysed within the cluster interventions **(Appendix C)**.

4. Discussion

This review is, to our knowledge, the first review about methodological issues and statistical analysis focused exclusively in S-P designs. We included eighteen studies and found different design variations. We identified key items of the CONSORT statement to focus on when using S-P designs and limitations in their design and report. Sample size calculations were one of the most challenging issues in the design of S-Ps: they were either omitted or based on the cluster randomisation element of the design, even though the evidence to do that is unclear. Finally, we provided a flow of participants' diagram template to report S-P designs in a clear way.

Variations of S-P designs had common features: the randomisation of two different entities, their sample size calculation and analysis, as well as their report. This suggests similar rationales for choosing the design, even though that was not made explicit by researchers. Logistic and design issues, such as the inability to randomise all the participants in the clusters (because, for example, participants were not willing to participate or didn't comply with the eligibility criteria), were found across designs. Each type of S-P design presents different challenges: partial S-P designs do not have a factorial element and multi-stratum S-P designs have more than two levels of information to report (the randomisation units, as well as the outcome collection level).

Reporting of CONSORT items was suboptimal, particularly those that require special consideration in S-P designs. We recommend researchers using this design to: use a designation for the S-P trial that helps other researchers understand the design – for example “multi-level factorial cluster randomised trial” when describing their studies; clearly present the randomisation units and the different interventions used in the abstract; and make their rationale explicit for choosing this design.

Estimating the sample size and statistical power for a study is an essential part of its design with implications on its statistical precision [6]. Calculating a sample size for S-P designs is challenging due to treatment at multiple levels (cluster and participant) and correlation at multiple levels (within-cluster and within-participant). There is no closed formula for that purpose and statistical simulation may be the best technique [6] However, none of the studies included a simulation-based sample size calculation, which could be due to the technical knowledge needed to do so. Six studies presented no sample size calculation and this is a similar prevalence when compared with cluster RCTs [26].

To calculate a sample size for an S-P design, the factorial and cluster randomised elements need to be considered using formulae available[27] and an inflation factor should be used to account for the clustered nature of the data. However, there is need for more research regarding the estimation of sample sizes in S-P designs and space for improvement in terms of its current reporting. When using a sample size calculation in S-P designs, researchers based it on the cluster-level randomisation target treatment difference. This approach is reasonable if the target treatment difference is assumed to be the same or bigger for the participant-level intervention. Such an approach will lead to the participant trial component having more power than the C-RCT component of the study. However, this was not made explicit in most of the trials sample size rationale. The primary interest of the trial will also have implications for its sample size calculation: an S-P design can be chosen when the interest is to determine whether there is an interaction or not between interventions, but it could equally be chosen when it is clear there is no interaction between interventions. We recommend that the target difference for each intervention in the S-P design is presented, unless its sample size is based on a clinical meaningful difference [28]. If there is an interest in the interaction between interventions, that should be made explicit as well as the expected consequences for the sample size of the trial. Besides helping assessing researchers’ assumptions, it would reduce the risk of overvaluing a statistically significant result that is clinically meaningless.

We considered the analysis used in the S-P designs reported here to be mostly adequate and coherent with the available sample size calculations. Since this design is challenging to implement and report, this could reflect the fact that more experience researchers or methodologists are opting to use it.

S-P designs should present results regarding interaction between interventions, when applicable, and make their a priori hypothesis explicit to facilitate assessment of whether a factorial design is an appropriate one. [5,29] One of the reasons to use a traditional S-P design is an interest in the interaction between two interventions (comparable to other types of factorial designs [5 29]), but only one study accounted for that in its design.

We recommend researchers using an S-P design to use a complete and understandable flow of participants' diagram, such as the one we present in **Appendix C**. Flow of participants' diagrams were particularly challenging to understand in the included studies, possibly due to the complexity of the design. The CONSORT diagram is a tool to aid the understanding of a trial, but there needs to be a trade-off between the amount of information provided and its clarity – there is no gain in presenting a lot of information in a diagram, if it is harder for the reader to follow.

S-P designs were challenging to identify and assess: out of the 154 full texts assessed, 18 ended up in our review which reflects the variety of terms used to describe it and the lack of information about randomisation units in the abstract. Even after identifying a study as an S-P design, abstracting its elements was testing and had to be resolved by discussion between researchers. The difficulty identifying S-P designs leads to the unavoidable limitation of possibly overlooking studies. However our process was submitted to a quality check and an agreement was achieved between the three reviewers to ensure a high-level standard of the screening process.

5. Conclusion

S-P designs are potentially efficient and widely applicable designs that help answer complex, multi-level research questions. In this review, we found several limitations in their report and design, including challenges for calculating their sample sizes. We recommend researchers using this design to comply with the CONSORT guidelines, giving special consideration to the key items of rationale, sample size, statistical methods and flow of participants' diagram.

References

1. Sedgwick P. Clinical trials: units of randomisation. *BMJ*. 2014;3297(May):1-2. doi:10.1136/bmj.g3297.

2. Crespi CM. Improved Designs for Cluster Randomized Trials. *Annu Rev Public Health*. 2016;37(1):1-16. doi:10.1146/annurev-publhealth-032315-021702.

3. Montgomery AA, Peters TJ, Little P, et al. Design, analysis and presentation of factorial randomised controlled trials. *BMC Med Res Methodol*. 2003;3(1):26. doi:10.1186/1471-2288-3-26.

4. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ*. 2012;345(sep04 1):e5661-e5661. doi:10.1136/bmj.e5661.

5. Mdege ND, Brabyn S, Hewitt C, Richardson R, Torgerson DJ. The 2 x 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: A systematic review. *J Clin Epidemiol*. 2014;67(10):1083-1092. doi:10.1016/j.jclinepi.2014.06.004.

6. Arnold BF, Hogan DR, Colford JMJ, Hubbard AE. Simulation methods to estimate design power: an overview for applied research. *BMC Med Res Methodol*. 2011;11:94. doi:10.1186/1471-2288-11-94.

7. Clarkson JE, Ramsay CR, Averley P, et al. IQuaD dental trial; improving the quality of dentistry: a multicentre randomised controlled trial comparing oral hygiene advice and periodontal instrumentation for the prevention and management of periodontal disease in dentate adults attending dental pri. *BMC Oral Health*. 2013;13:58. doi:10.1186/1472-6831-13-58.

8. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Int J Surg*. 2012;10(1):28-55. doi:10.1016/j.ijsu.2011.10.001.

9. Rosen L, Manor O, Engelhard D, et al. Can a handwashing intervention make a difference? Results from a randomized controlled trial in Jerusalem preschools. *Prev Med (Baltim)*. 2006;42(1):27-32. doi:10.1016/j.ypmed.2005.09.012.

10. Middleton JF, McKinley RK, Gillies CL. Effect of patient completed agenda forms and doctors' education about the agenda on the outcome of consultations: randomised controlled trial. *BMJ*. 2006;332(7552):1238-1242. doi:10.1136/bmj.38841.444861.7C.

11. Scholes D, Grothaus L, McClure J, et al. A randomized trial of strategies to increase chlamydia screening in young women. *Prev Med (Baltim)*. 2006;43(4):343-350. doi:10.1016/j.ypmed.2006.04.019.

12. Blalock SJ, DeVellis B, Patterson C, Campbell M, Orenstein D, Dooley MA. Blalock 2002.pdf. *Am J Heal Promot*. 2002;16(3):146-156.

13. Watson E, Clements A, Lucassen A, Yudkin P, Mackay J, Austoker J. Education improves general practitioner (GP) management of familial breast/ovarian cancer: findings from a cluster randomised controlled trial. *J Med Genet*. 2002;39(42):779-781. doi:10.1136/jmg.39.10.779.

14. Ling BS, Solano RE, M TJ, et al. Physicians Encouraging Colorectal Screening: A Randomized Controlled Trial of Enhance Office and Patient Management on Compliance With Colorectal Cancer Screening. *Arch Intern Med*. 2009;169(1):47-55. doi:10.1001/archinternmed.2008.519.

15. Weymiller A, Montori V, Jones L, et al. Helping Patients With Type 2 Diabetes Mellitus Make Treatment Decisions. *Arch Intern Med*. 2007;167:1076-1082.

16. Lamb SE, Gates S, Williams MA, et al. Emergency department treatments and physiotherapy for acute whiplash: A pragmatic, two-step, randomised controlled trial. *Lancet*. 2013;381(9866):546-556. doi:10.1016/S0140-6736(12)61304-X.

17. Riemersma-van der Lek RF, Swaab DF, Twisk J, Hol EM, Hoogendijk WJ, Van Someren EJ. Effect of Bright Light and Melatonin on Cognitive and Noncognitive Function in Elderly Residents of Group Care Facilities. *JAMA*. 2008;299(22):2642. doi:10.1001/jama.299.22.2642.

18. Watson J, Helliwell P, Morton V, et al. Shoulder acute pain in primary healthcare: Is retraining effective for GP principals? SAPPHIRE - A randomized controlled trial. *Rheumatology*. 2008;47(12):1795-1802. doi:10.1093/rheumatology/ken360.

19. Daucourt V, Saillour-Glenisson F, Michel P, Jutand M-A, Abouelfath A. A Multicenter Cluster Randomized Controlled Trial of Strategies to Improve Thyroid Function Testing. *Med Care*. 2003;41(3):432-441.

20. Grunfeld E, Manca D, Moineddin R, et al. Improving chronic disease prevention and screening in primary care : results of the BETTER pragmatic cluster randomized controlled trial. 2013:1-12.

21. Luby SP, Agboatwalla M, Painter J, Altaf A, Billhimer WL, Hoekstra RM. Effect of intensive handwashing promotion on childhood diarrhea in high-risk communities in Pakistan: a randomized controlled trial. *JAMA*. 2004;291(21):2547-2554. doi:10.1001/jama.291.21.2547.

22. Richards DA, Lovell K, Gilbody S, et al. Collaborative care for depression in UK primary care: a randomized controlled trial. *Phychological Med*. 2008;38:279-287. doi:10.1017/S0033291707001365.

23. Jordan R, Adab P, Sitch A, et al. Targeted case finding for chronic obstructive pulmonary disease versus routine practice in primary care (TargetCOPD): a cluster-randomised trial. *Lancet Respir Med*. 2016;4(9):720-730.

24. McNulty C, Thomas M, Bowen J, et al. Interactive workshops increase chlamydia testing in primary care — a controlled study. *Fam Pract*. 2008;25(June):279-286. doi:10.1093/fampra/cmn032.

25. Warming S, Ebbehøj NE, Wiese N, Larsen LH, Duckert J, Tønnesen H. Little effect of transfer technique instruction and physical fitness training in reducing low back pain among nurses: a cluster randomised intervention study. *Ergonomics*. 2008;51(10):1530-1548. doi:10.1080/00140130802238606.

26. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: A review. *J Clin Epidemiol*. 2015;68(6):716-723. doi:10.1016/j.jclinepi.2014.10.006.

27. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold Publishers; 2000.

28. Cook JA, Hislop J, Altman DG, et al. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials*. 2015;16(1):12. doi:10.1186/s13063-014-0526-8.

29. McAlister F a, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA*. 2003;289(19):2545-2553. doi:10.1001/jama.289.19.2545.

Sample size calculation - n	
Reported	12
Not reported	6
Type of sample size calculation - n	
Reported to use inflation factor	10
Cluster level comparison presented	1
Participant level comparison presented	1
Both comparison levels presented	2
Used clinical meaningful difference	2
Comparison level unclear	4
Did not report an inflation factor	2
Participant level comparison presented	1
Comparison level unclear	1
Analysis – n	.
Hierarchical model	12
Generalised estimating equations	4
Aggregated analysis	2
Interaction report – n	.
Yes	9
No	3
Unclear	2
Not applicable[1]	4

[1] Partial split-plots do not have a factorial design and therefore cannot present an interaction result

Table 2 – Sample size calculation, analysis and interaction considerations

CONSORT diagram items	Number of articles included in which the information indicated is present in the diagram (N=18)
Diagram presented	15
Cluster-randomisation level	
Number of clusters assessed for eligibility	11
Number of clusters randomised	13
Average cluster size or equivalent (mean/median)	3
Range of cluster size or equivalent (variance, standard deviation)	3
Number of clusters lost to follow-up	3
Number of cluster analysed	3
Participant-randomisation level	
Number of participants assessed for eligibility	8
Number of participants randomised	13
Number of participants lost to follow-up	12
Number of participants analysed	14

Table 3 Quality of information presented regarding the flow of participants

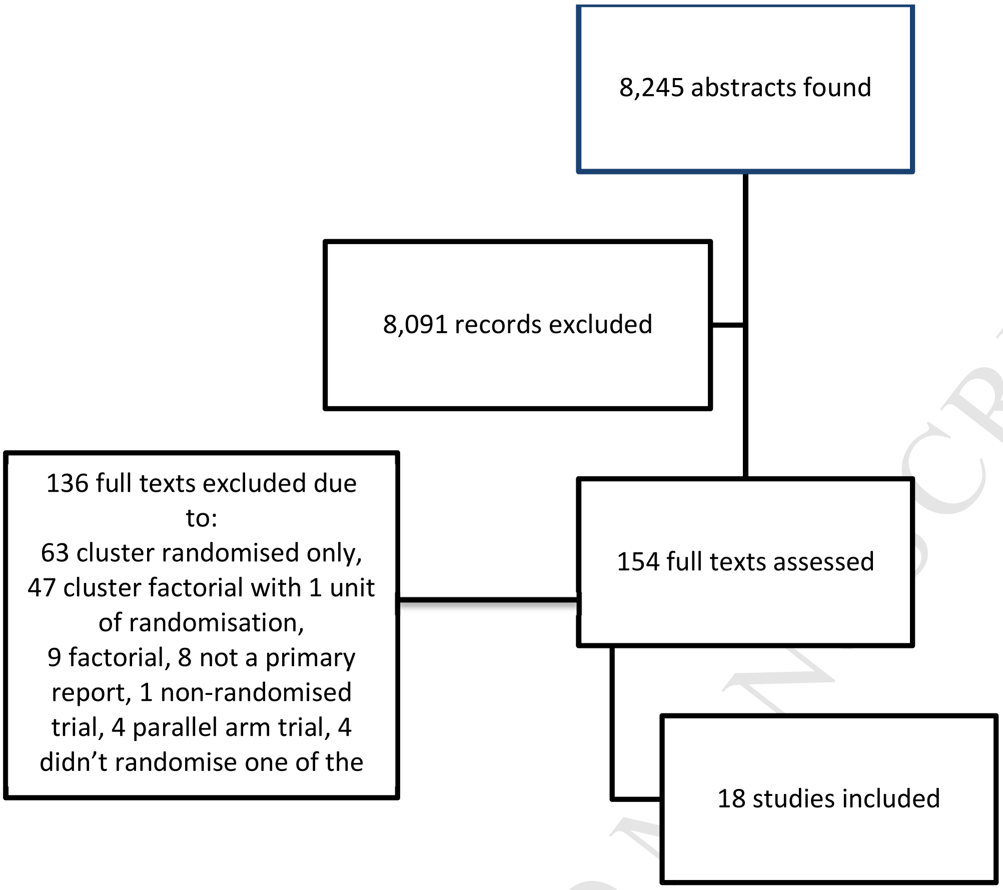
8,245 abstracts found
8,091 records excluded
136 full texts excluded due to: 63 cluster randomised only, 47 cluster factorial with 1 unit of randomisation, 9 factorial, 8 not a primary report, 1 non-randomised trial, 4 parallel arm trial, 4 didn’t randomise one of the
154 full texts assessed
18 studies included

Figure 1 – PRISMA diagram

What is new?

Split-plot designs are used mainly for education and implementation of guidelines' research and there are several variations of the design: we classified these as traditional, partial, multi-stratum and nested split-plot designs.

We identified key items of the CONSORT statement to focus on when using split-plot designs, such as the rationale, sample size, statistical methods and flow of participants' diagram and limitations in their design and report. We provide guidance on how to report these key items, as well as suggested a flow of participants' diagram to help visualise the complex structure of the design and its flow of participants'.

Sample size calculations for the split-plot design are challenging due to its complex structure and need to take into account different levels of correlation. There is need for more research on how to calculate sample sizes for split-plot designs, but included trials could improve on the report and clarity of their calculations. Potential interaction between interventions is one of the reasons to use this design, however all but one trial ignored interaction at the design stage.

Conflicts of interest / Funding Source declarations

Conflicts of interest: none.

No funding was provided for this systematic review. No funding source had any role in design, data collection, interpretation or reporting of the trial or these analyses. Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Government Health and Social Care Directorates.