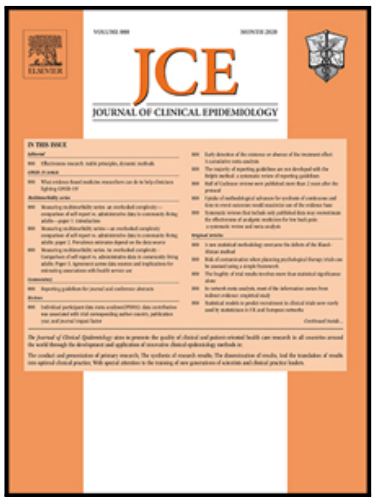


Cluster randomised trials of individual-level interventions were at high risk of bias

Christina Easter , Jennifer A. Thompson , Sandra Eldridge , Monica Taljaard , Karla Hemming

PII: S0895-4356(21)00199-2
DOI: https://doi.org/10.1016/j.jclinepi.2021.06.021
Reference: JCE 10546

To appear in: *Journal of Clinical Epidemiology*

Accepted date: 22 June 2021

Please cite this article as: Christina Easter , Jennifer A. Thompson , Sandra Eldridge , Monica Taljaard , Karla Hemming , Cluster randomised trials of individual-level interventions were at high risk of bias, *Journal of Clinical Epidemiology* (2021), doi: https://doi.org/10.1016/j.jclinepi.2021.06.021

Highlights

- Due to the risks of identification and recruitment bias, opting for a cluster design when individual randomisation would be feasible needs a strong justification. Concerns around contamination are unlikely to be acceptable justifications; although estimation of indirect effects might be.
- When cluster randomisation is adopted, we recommend that authors provide a clear justification for the choice of cluster randomisation and clearly outline strategies to mitigate increased risks of bias. This should include identification and recruitment by someone blind to the treatment allocation and minimal or objective individual-level eligibility criteria.
- Other good conduct procedures which are routinely implemented in individually randomised trials should be followed. These include implementation of the randomisation using an accepted method of allocation concealment, for example, by using an independent statistician to generate the allocation sequence; blind outcome assessment when outcomes are subjective; and clear pre-specification (in a protocol or trial registration site) of the primary outcome including primary assessment time and method of primary analysis.
- All these aspects should be clearly reported as per CONSORT guidelines. To ensure particular clarity around identification and recruitment, authors should also provide a timeline-cluster diagram.

Cluster randomised trials of individual-level interventions were at high risk of bias

Christina Easter[1] Jennifer A. Thompson[2] Sandra Eldridge[3] Monica Taljaard[4] and Karla Hemming[5]*

1. Institute of Applied Health Research, University of Birmingham, Birmingham, UK. c.l.easter@bham.ac.uk;

2. Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. Jennifer.thompson@lshtm.ac.uk;

3. Centre for Clinical Trials and Methodology, Queen Mary University of London, London. s.eldridge@qmul.ac.uk;

4. Clinical Epidemiology Program, Ottawa Hospital Research Institute, 1053 Carling Avenue, Ottawa, Ontario, Canada; and School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada. mtaljaard@ohri.ca.

5. Institute of Applied Health Research, University of Birmingham, Birmingham, UK. k.hemming@bham.ac.uk;

** Corresponding author:*

Acknowledgements

Acknowledgements are given to Stuart Nicholls (SN, snicholls@ohri.ca) Kelly Carroll (KC, kecarroll@ohri.ca) and Austin R Horn (ARH, ahorn5@uwo.ca) for undertaking search to identify studies; and to Caroline Kristunas (c.a.kristunas@bham.ac.uk) and James Martin (j.martin@bham.ac.uk) for helping with the data abstraction.

Author contributions

KH led the development of the project and wrote the first draft of the paper. MT led the search process and led the identification of studies for inclusion. CE designed and developed the data abstraction tools and conducted the statistical analysis. MT, SE and JT provided important oversight to the project. All authors helped develop the data abstraction tools, provided critical insight, contributed to the data abstraction exercise, and commented on the draft paper.

Funding

Kar[illegible] funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union (MR/R010161/1). Christina Easter is funded by the NIHR (SRF-2017-10-002). This research was also partly funded by the UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative.

Abstract

Objectives To describe the prevalence of risks of bias in cluster-randomised trials of individual-level interventions, according to the Cochrane Risk of Bias tool.

Study design and setting Review undertaken in duplicate of a random sample of 40 primary reports of cluster-randomised trials of individual-level interventions.

Results The most common reported reasons for adopting cluster randomisation were the need to avoid contamination (17, 42.5%) and practical considerations (14, 35%). Of the 40 trials all but one was assessed as being at risk of bias. A majority (27, 67.5%) were assessed as at risk due to the timing of identification and recruitment of participants; many (21, 52.5%) due to an apparent lack of adequate allocation concealment; and many due to selectively reported results (22, 55%), arising from a mixture of reasons including lack of documentation of primary outcome. Other risks mostly occurred infrequently.

Conclusion Many cluster-randomised trials evaluating individual-level interventions appear to be at risk of bias, mostly due to identification and recruitment biases. We recommend that investigators carefully consider the need for cluster randomisation; follow recommended procedures to mitigate risks of identification and recruitment bias; and adhere to good reporting practices including clear documentation of primary outcome and allocation concealment methods.

Key words: Cluster randomised trials; risk of bias; individual-level interventions; selection bias

Running title: Risk of bias in cluster randomised trials of individual-level interventions

Intr

In individually randomised trials, patients are randomly allocated to different interventions, henceforth referred to as treatment or control conditions. Rather than randomising individual patients, cluster-randomised trials randomise entire clusters (such as wards, schools or social groups) to treatment or control conditions [Murray 1998; Eldridge 2012; Turner 2017a; Turner 2017b]. Cluster-randomised trials can be used to evaluate different types of interventions, sometimes delivered at the level of the entire cluster (cluster-level interventions), sometimes delivered at the level of the health care professionals (professional-level intervention), sometimes delivered directly to individual patients (individual-level intervention) and sometimes a mixture [Edwards 1999, Eldridge 2005]. Cluster-level and professional-level intervention necessarily require cluster randomisation.

Cluster-randomised designs are known to be at increased risk of bias compared to the individually randomised design [Puffer 2003; Hahn 2005; Brierley 2012; Froud 2012, Diaz-Ordaz 2013, Eldridge 2008; Yang 2017; Bolzern 2018]. These risks of bias often challenge the strength of the evidence generated from this design and downgrade the quality of evidence that they contribute to systematic reviews [Leyrat 2019]. Risks of bias in randomised trials have been carefully described in the Cochrane systematic review Risk of Bias tool (RoB2.0) [Higgens 2016] and an adaptation of the main guidance has been developed for cluster trials. Recruitment and identification biases are a unique source of bias under cluster randomisation, with trials being particularly vulnerable to this bias when it is necessary to identify or recruit individuals into the study after randomisation [Eldridge 2009; Bolzern 2018]. For example, to evaluate a pharmacological intervention without blinding and with randomisation at the level of a village, if recruitment occurs after randomisation then the decision to participate (or not) might be affected by knowledge that they will receive the active intervention (or not). Such beliefs can affect outcomes, and therefore may bias the study's estimates of the between-group effect. Recommendations suggest that to avoid or reduce these risks, trials adopt broad eligibility criteria at the level of the individual and, if participants cannot be identified and recruited prior to randomisation, identification and recruitment of participants is by someone who is blind to the cluster allocation [Hahn 2005; Eldridge 2009; Giraudeau 2009].

Whilst there may be good reasons for adopting cluster randomisation including to avoid contamination (e.g., individuals in the control condition being exposed to interventions) and for logistical simplicity (e.g., to simplify the fieldwork by having only one type of intervention in a particular cluster or geographical area) [Torgerson 2001], individual-level interventions could, in theory, be evaluated with an individually randomised trial. Whilst other reviews have documented risks of bias in cluster trials more generally, none have documented risks of bias in cluster trials of individual-level intervention where individual randomisation would in theory be feasible. Here, we report the results of a review of the risks of bias in contemporary primary reports of cluster-randomised trials of individual-level interventions. Our objectives were to (i) identify the prevalence of key risks of bias in cluster-randomised trials of individual-level interventions; (ii) to describe prevalence of design features associated with increased risks of bias and (iii) formulate design recommendations to avoid such risks. We also describe the reliability of the two independent assessments of risk of bias.

Methods

Scope of review

We used a convenience sample of trials identified in a previously published review of cluster trials of individual-level interventions published in the interval from 2007 to 2016 [Taljaard 2020]. In brief, the review included primary reports of cluster-randomised trials of individual-level therapeutic interventions conducted in Canada, USA, European Union, Australia, and Low- or Middle-Income Country (LMIC) and published in English. Individual-level interventions were defined as any intervention that is aimed solely at the individual; thus, we excluded evaluations of cluster-level or professional-level interventions and evaluations where these types of intervention were included alongside an individual-level intervention. Therapeutic interventions were defined broadly as medicinal, clinical or surgical based interventions (see Taljaard 2020 for a full definition). Full text articles were screened in a random sequence until a sample size of 40 was achieved.

Justification for scope

We used an existing database of primary reports of individual-level cluster-randomised trials for logistical reasons: screening and review of a very large number of citations from the general medical literature to isolate primary

rep

using this existing sampling frame allowed us to obtain a random sample of such trials. Including individual-level interventions only, whilst narrowing scope of generalisability, allows us to meet our objective of evaluating risk of bias in situations where a theoretical alternative is the individually randomised design. Focusing on therapeutic interventions targets our finding to the evaluation of interventions intended to bring about improvements in health.

Data abstraction process

Data were abstracted from the full trial reported. We additionally searched the full trial reports to identify any reference to study protocols or statistical analysis plans (which sometimes included additional study information such as patient information and consent forms) and searched for trial registration documentation for each included study by using any trial registration reported in the text, or using google searches to identify any registration. All data was abstracted by one reviewer (CE) and independently and in duplicate by a second randomly allocated reviewer (KH, CK, JT or JM). After both assessments were completed, disagreements were identified, and a consensus (henceforth referred to as the joint assessment) reached by discussion. Where necessary, a third reviewer was consulted to reach agreement (KH or JT). The data capture was electronic (using RedCap). Study reports were randomly sorted before data abstraction.

Data abstracted on general characteristics of trials

We abstracted the following trial characteristics: publication year; country of conduct; type of cluster; rationale for cluster design; trial design (parallel, factorial, cross-over, stepped-wedge); number of clusters randomised; average (realised) cluster size. We also extracted whether a trial protocol, statistical analysis plan or trial registration were available because in the absence of such documentation, it is impossible to determine whether the primary outcome was pre-specified. We extracted the number of eligibility criteria at the participant level as more eligibility criteria increases the likelihood of differential inclusion [Giraudeau 2009]. We also classified each trial based on whether it was reported that an independent person conducted the randomisation as this is an indicator of concealment of the randomisation process. Additionally, we extracted our assessment of whether the outcome was subjective or objective.

Data abstracted on risk of bias

For each study report, reviewers were provided with a detailed risk of bias assessment form (Supplementary Material 1). This risk of bias assessment aimed to assess the risk of bias for each of the five domains in the RoB2.0 tool (Table 1). These domains are (i) bias arising from the (a) randomisation process and (b) the timing of identification and recruitment of participants in relation to the timing of the randomisation; ii) bias due to deviations from the intended intervention; iii) bias due to missing outcome data; iv) bias due to the measurement of the outcome; and v) bias due to the selection of the reported result.

In the RoB2.0 tool, under an extension for cluster trials (accessed May 2019; dated 20th October 2016), these risks are identified by a series of *signalling questions* with an extensive set of *elaborations* providing extensive detail about how to answer the signal questions [Higgins 2016]. To avoid having to refer back to the extensive elaborations, we mapped the *signalling questions* from RoB2.0 and their associated set of *elaborations* onto a set of data abstraction *items* (Supplementary Material 2). As an illustrative example, domain 1a is "Bias arising from the randomisation process" and one of the three signalling questions for this domain is "Was the allocation sequence random?" and the associated question on our mapped data abstraction item was "How was the randomisation of clusters to allocated treatment(s) conducted? (Tick all that apply)". Following the reasoning outlined in the elaboration of RoB2.0, trial reports which were identified as using one of the random methods of allocation defined in the explanatory material were then classified as using a random allocation method. Another associated signalling question is "Was the allocation sequence concealed until clusters were enrolled and assigned to interventions?" and the associated data abstraction items were "Who conducted the randomisation?" and "How was the randomisation allocation of clusters concealed?" Again, following the elaboration outlined in RoB2.0, study reports which reported the randomisation to be by someone independent, or using a trials unit, or using some other acceptable concealment mechanism, such as internet-based randomisation or sealed envelopes, were classified as having a concealed allocation process.

Fro

independent assessment stage also included the option "unclear" but this option was not retained at the joint assessment; we did not use the classification of "probably yes" or "probably no"). We followed the RoB2.0 mapping from these signalling questions to risks of bias assessment for each domain to classify each trial under each domain as "low risk of bias", "some concerns" or "high risk of bias" (again at the independent assessment stage the option "unclear" was also included). Of note, this means that no trials were assessed as at unclear risk as this is no longer a domain in the RoB2.0 tool (any assessments of "no information" are mapped to the relevant category following the RoB2.0 mapping). Finally, again following RoB2.0 we created an overall study assessment of risk of bias: a study is judged at high risk of bias if it is assessed at high risk in at least one domain or some concerns for multiple domains; low risk of bias if it is assessed as low risk in all domains; and some concerns otherwise. A small number of assumptions were made along the way and these are noted in the table footnotes and in the results section below.

Statistical analysis

We describe the assessment of risk of bias (based on the consensus / joint agreement) for all domains and signalling questions, using simple descriptive statistics (numbers and percentages). We also describe the reliability of the independent assessments (not the final joint assessment), by computing the percentage agreement (including raw percentage agreement and the Gwet's AC value [Gwet 2014; Wongpakaran 2013]) between the two independent assessments for each broad domain and for each of the signalling questions. Reliability was computed across a non-ordinal four-point scale for both risk of bias (high risk of bias / some concerns /low risk of bias / unclear); and across signalling questions ("yes", "no", "no information", "unclear"). Gwet's AC statistic was unweighted due to the non-ordinal categories for the signalling questions but weighted for the risk of bias (with the penalization set to thirds: low penalization set to 2/3 for high-some concerns, low-some concerns and anything-unclear; and high penalization set to 1/3 for high-low concerns).

Results

Study characteristics

Full information on the random sample selection can be found elsewhere [Taljaard 2020], in brief the search identified 10,014 potential studies (after removal of duplicates), of which 3,097 were not excluded at the abstract screen. Of these 1,190 underwent a full text screen until 40 were identified as meeting the eligibility criteria. A description of the 40 trials is provided in Table 2. The trials were conducted between 2007 and 2016 and covered a range of settings including LMICs (21, 52.5%), Canada / USA (7, 17.5%) and Europe (11, 27.5%) amongst others; the most common reported reason for adopting cluster randomisation was avoiding contamination (17, 42.5%) and practical reasons (14, 35%), and 10% (10 trials) did not report the rationale for cluster design. The most common form of cluster was a residential area (15, 37.5%) or hospital / nursing home / clinic (15, 37.5%); the median number of clusters included in each study was 24 (inter-quartile range, IQR: 12-49.5); the median cluster size was 114 (IQR: 35-456); and most designs were parallel (28, 70%). Only a minority of trials had an accessible protocol paper or statistical analysis plan (16, 40%), although most were registered on a trial registration site (33, 82%). A sizeable minority (6, 15%) had no documentation available to verify any pre-specification, for example of the primary outcome. Few used an independent statistician to implement the randomisation (11, 27.5%). The majority had more than three eligibility criteria at the level of the individual (24, 60%). Most studies (30, 75%) were assessed to have objective primary outcome.

Broad assessment of risk of bias

Overall, all but three of the trials were assessed as at high risk of bias and only one was assessed at low risk of bias (Table 3, Figure 1). Most trials were assessed as high risk on one (9, 22.5%) or two (14, 35%) domains; with a smaller number being assessed as risk on up to 4 (6, 15%) or 5 (2.5%) domains. Breaking down these assessments into finer categories (Supplementary Tables 1a to 5) helps identify the design features associated with these risks of bias. We next consider each domain separately.

Domain 1a bias arising from the randomisation process: Around half of the trials (21, 52.5%) were assessed as being at high risk of bias due to the randomisation process. Whilst all were assessed to use a random method to allocate clusters to treatment conditions, many (21, 52.5%) were assessed as not having concealed the allocations (i.e., not clearly reporting randomisation by someone independent, or using a trials unit, or not using some acceptable

con

report any cluster-level characteristics to allow any assessment of balance of the randomisation process.

Domain 1b bias arising from identification or recruitment of participants within clusters: A large majority of the trials (27, 67.5%) were assessed as at risk of bias due to the timing of identification and recruitment of participants. Most trials (36, 90%) were assessed as identifying or recruiting participants after randomisation and most (27, 67.5%) were assessed to include participants in such a way that selection could have been affected by knowledge of the intervention. As shown in Supplementary Table 6, this is because many trials both recruited participants post randomisation and those recruiting participants were not reported to be blind to the intervention. In some trials (15, 37%), we identified baseline imbalances that suggest differential identification or recruitment of individual participants between arms.

Domain 2 bias due to deviations from intended interventions: Most trials (34, 85%) were at low risk of bias due to deviations from the intended interventions. However, in a large number of trials, we deemed that participants were aware that they were in a trial (27, 67.5%) and aware of their assigned intervention (20, 50%), as did trial personnel (34, 85%). Despite this, only a minority of trials (8, 20%) were assessed as showing evidence of deviations from the intended intervention beyond what would be expected in usual practice; and in only a few trials (6, 15%) were these deviations from intended intervention unbalanced between groups and assessed as likely to have affected the outcome (Supplementary Table 7). Here we assumed that a deviation of the intended intervention occurred if more than 10% of the participants were reported not to have received the intended intervention condition. In all trials, most clusters and participants were reported to be analysed according to randomisation (i.e., by intention to treat).

Domain 3 bias due to missing outcome data: Most trials (33, 82.5%) were assessed as at low risk of bias due to missing outcome data, mostly because missing data arose infrequently: only in a small number of trials (9, 22.5%) was the outcome data unavailable for more than 10% of participants. In a small number of cases (4, 10%) outcome data were deemed to be differential across treatment arms.

Domain 4 bias in measurement of the outcome: Most of the trials were assessed as being at low risk of bias due to measurement of the outcome (31, 77.5%), although some (9, 22.5%) were assessed as being at high risk of bias. Whilst in almost all trials (36, 90%), outcome assessors were aware the trial was taking place and in many (26, 65%) they were aware of the intervention received by the participant, because most outcomes were assessed as objective (30, 75%, Table 2) this lack of blinding was assessed as inconsequential (for outcome assessment).

Domain 5 bias in selection of the reported result: A large proportion of the trials (22, 55%) were assessed as at high risk of bias in the selection of the reported result, and this arose due to multiple reasons. For a sizeable number of trials (14, 35%) the primary outcome was not clearly defined, either because the outcome itself was not clearly defined (7, 17.5%) in any of the trial registration database, study protocol, or methods section of the main trial report, or, because the primary assessment time was not clearly defined (9, 22.5%). For a few trials it was not stated if the primary analysis would be adjusted or unadjusted for covariates (6, 15%). Almost all trials reported the scale the primary outcome would be measured on, and how any binary variables would be categorised, but some were assessed as not having a plan for how they would handle missing data despite having missing data (9, 22.5%).

Reliability of independent assessments

The raw percentage agreement between the independent assessments were calculated for each signalling question, domain and overall risk of bias for each paper (Table 3 and Supplementary Table 8). For the overall assessment of each study the agreement was high (Gwet's AC: 0.92 95% CI: 0.85,0.99), but this varied across the different domains: agreement was 0.46 (95% CI: 0.20,0.72) for domain 1a (randomisation process); 0.59 (95% CI: 0.37,0.81) for domain 1b (identification and recruitment process); 0.85 (95% CI: 0.74,0.96) for domain 2 (deviations from intended interventions); 0.77 (95% CI: 0.62,0.92) for domain 3 (missing outcome data); 0.79 (95% CI: 0.64,0.95) for domain 4 (measurement of the outcome) and 0.44 (95% CI: 0.19,0.70) for domain 5 (selection of reported result).

Particular signalling questions which had strikingly low reliability included whether the allocation was concealed from the clusters at randomisation (0.41, 95% CI: 0.19,0.62); whether the selection of individual participants was likely affected by knowledge of the intervention (0.56, 95% CI: 0.36, 0.76); whether there were baseline imbalances across individual-level characteristics (0.53, 95% CI: 0.33,0.73); whether participants were aware of their assigned intervention (0.53, 95% CI: 0.33,0.74); whether proportions of missing data were similar across interventions (0.59,

95% [illegible] analysis (0.52, 95% CI: 0.30,0.74).

Discussion

Summary of findings

In our review of a random sample of 40 cluster-randomised trials of individual-level interventions, we found that all but one was at risk of bias. Trials were at risk of bias across multiple domains, but a prominent source was identification and recruitment bias. We found that the vast majority of cluster-randomised trials of individual-level interventions identify or recruit research participants after randomisation of clusters to treatment conditions and fail to report use of any strategies to prevent identification and recruitment bias. In many it was deemed possible that selection of individual participants could be affected by knowledge of the intervention; with some showing evidence of baseline imbalance on individual-level characteristics across treatment arms.

We identified other possible risks of bias not necessarily specific to the use of cluster randomisation. For example, many trials were assessed as not implementing randomisation in a way that is clearly concealed. This is something which is easily correctable by use of an independent statistician or other acceptable concealed randomisation method. Other risks of bias included a failure to clearly specify or document the primary outcome or primary assessment time: a small minority of trials neither publish a protocol paper (or statistical analysis plan) nor pre-register the trial on a trial registration database. In these trials, there is no possible way to verify any pre-specified primary outcome and these trials will be at risk of selective reporting. Related to this, many trials were assessed as not clearly documenting other features of their outcomes (such as primary assessment time) and analysis plan. Some studies were assessed as being at risk of bias due to measurement of the outcome; this might be surmountable in some trials by using blind outcomes assessors when outcomes are subjective.

The one trial identified as low risk of bias was a trial of skin cleansing wipe in new-born babies with a placebo control [Tielsch 2007]. The placebo control helps minimize risk of bias in most domains: for example, despite the use of post-randomisation identification and recruitment, there is no risk of identification and recruitment bias because the placebo control ensures recruitment is blind to the intervention condition. Furthermore, the outcome assessment is blinded (and in this trial also happened to be objective, namely mortality).

Limitations

We used a convenience sample of trials identified in another review. This means we have assessed risk of bias in a relatively small sample of 40 trials over an extended period of time between 2007 and 2016. Both reporting and conduct might have improved in recent years with the use of the CONSORT statement extension for cluster randomised trials [Campbell 2012], but most evaluations of reporting and conduct suggest that improvements are minimal at best [Cook 2021]. Moreover, these trials are a true random sample of cluster-randomised trials of individual-level interventions across all journals, which should mean these results are representative of other cluster-randomised trials of similar types of interventions. We opted to use this sample as identifying a true random sample of cluster trials of individual-level interventions is very labour intensive and beyond our scope. Rather than taking a random sample, as much less labour-intensive search strategy would have been to focus on specific journals, such as high impact journals, but this tends to underestimate the scale of any problem.

Our assessment of bias, by following RoB2.0, assesses in part theoretical risk as well as manifestations of actual risk such as imbalance across trial arms [Higgins 2016]. We also used an earlier version of this tool (downloaded in May 2019, dated 20th October 2016) and there have subsequently been several minor revisions (March 2021). Assessment of risk of bias in both randomised and non-randomised studies is important, and despite availability of multiple tools, can be difficult. Others have shown that the reliability of assessments based on reviewing trial reports might be low for assessments which involve subjectivity [Losilla 2018; Hemming 2019; Minozzi 2019]; and our results are consistent with these findings: independent assessments showed low reliability for questions which involve some subjectivity (e.g., whether there was any imbalance) and were generally lower than those that might be considered more objective (e.g., was the study randomised).

assessments of reliability should not be considered an assessment of reliability of the RoB2 tool. To assess the reliability of the RoB2 tool it is necessary to assess the reliability of the joint assessments and to this end it would be necessary to repeat the two independent assessments and their discussion, so as to obtain two joint assessments. The reliability of the joint assessment is expected to be higher than the reliability of the independent assessments as the joint consensus involved extensive discussion process to reconciliate individual assessments. We therefore do not suggest that our assessment is an assessment of the reliability of the RoB2 tool, despite others having suggested reliability between two independent measures can assess the reliability of RoB2 [Minozzi 2020]. Nonetheless domains or signalling questions with low agreement might be indicative of domains or signalling questions which are less clearly amenable to an assessment of bias than those with higher agreement, and this might be translate more generally when others are using the RoB2 tool to assess risk of bias within the context of a review. Low reliability might either reflect poor reporting of the relevant items in the primary paper or the requirement to make a subjective assessment and in both cases, it might be necessary for reviewers to make assumptions.

By necessity we made assumptions. For example, not all trials clearly reported whether participants were actively recruited into the study, here we assumed that any mention of "consent" equated to active recruitment. In many trials it was difficult to identify if recruitment occurred post randomisation. Again, here we made assumptions, for example, in an acute setting such as the intensive care unit, we assumed patient accrual had to occur post randomisation; or when the recruitment period was reported to last a considerable duration, such as more than a year. Most trials did not clearly specify if participant recruitment was blind to the treatment allocation, and we assumed it was not blind unless specifically mentioned. Conversely, for those trials without any active patient recruitment, we assumed any knowledge of the intervention would not influence selection of identification of participants for inclusion, even though in practice these biases can arise in cluster trials without direct recruitment. We also made an arbitrary decision that a deviation from the intended intervention had occurred when more than 10% of the participants were reported not to have received their intended intervention condition, or that the authors had reported significant concerns around deviations. The issue of deviation of intended treatments is nuanced for pragmatic trials where the objective is to evaluate the effect of the offer of treatment not necessarily the effect of adherence to the treatment – meaning that this lack of adherence might not be important from a pragmatic perspective.

Research in context

Knowledge of treatment condition at the time of patient recruitment is known to be a risk factor for differential identification and recruitment of participants across treatment arms [Bolzern 2018; Giraudeau 2009; Hahn 2005; Yang 2017], unless recruitment and identification are conducted by someone blind to the treatment allocation or the inclusion criteria are broad [Brierley 2012; Eldridge 2009; Giraudeau 2009]. Methodological reviews have identified that many cluster trials are at risk of these identification and recruitment biases because they recruit participants with knowledge of allocated treatment and this often manifests in baseline imbalances [Puffer 2003; Brierley 2012; Bolzern 2018]. These assessments of risk have taken varying forms and it is difficult to compare across reviews. For example, in a review of recent randomised trials, cluster trials were reported to be more likely to have a significant baseline imbalance on age, whereas individually randomised trials were not [Bolzern 2018]. Others have assessed about 40% of cluster trials to be at risk of these types of biases [Puffer 2003; Brierley 2012; Diaz-Ordaz 2013]; and sometimes this has been reported to be somewhat lower despite including many trials with post randomisation recruitment [Eldridge 2008, Froud 2012]. Thus, the prevalence of risks of bias due to identification and recruitment reported here is higher than in previous reviews. This is likely explained by the fact that we focused on cluster-randomised trials of individual-level interventions, whereas other reviews have included cluster-level interventions where patient recruitment is less common or may more likely to occur prior to randomisation.

We also identified that many trials did not report using an allocation method that was clearly concealed. This information was assessed on the basis of whether the randomisation was conducted by someone independent, how the randomisation was implemented and whether the clusters were all recruited before randomisation. This finding is consistent with findings in individual randomised trials which have also been identified at risk of bias due to implementation of the randomisation process [Kahan 2015]. We also identified evidence of lack of clear specification of the primary outcome, primary assessment time and primary analysis method, again similar to that identified in

indi

might represent lack of good reporting practices. Whist we did not directly assess quality of reporting, despite the existence of specific reporting guidelines for cluster trials [Campbell 2012], we identified many elements were not well reported. However, lack of awareness of reporting may reflect a lack of awareness around conduct too. Timeline diagrams provide one method of improving reporting of the elements around timing and blinding status of identification and recruitment of participants [Caille 2016].

Finally, we identified that the most common reasons for adopting cluster randomisation were due to either a concern over contamination or for practical reasons; and this echoes what others have found [Taljaard 2017]. In a comparison between a novel treatment and usual care any bias due to contamination will attenuate the true treatment effect [Torgerson 2001, Moerbeek 2005, Hemming 2021]. Yet, in the very specific setting of cluster randomised trials of individual-level interventions with post randomization recruitment without blinding, we have identified a high risk of bias due to the differential recruitment across treatment arms. Individually randomised trials, by their nature of not having to recruit post randomisation, would not be at risk of this bias. Biases due to identification and recruitment bias operate in an unpredictable direction. Thus, concerns over contamination is unlikely to be an acceptable justification for using cluster randomisation in most evaluations of individual-level interventions with unblinded recruitment. Selecting a cluster randomised trial with knowledge that it will be at high risk of bias and without taking steps to mitigate these risks should be considered a poor use of resource at best and at worst unethical [CIOMS 2016]. On the other hand, where interest lies in total effects of individual-level interventions (both direct and indirect benefits), so when contamination a positive feature of implementation, then cluster randomisation might be the only design of choice [Hox 2014].

Recommendations

1. Due to the risks of identification and recruitment bias, opting for a cluster design when individual randomisation would be feasible needs a strong justification. Concerns around contamination are unlikely to be acceptable justifications; although estimation of indirect effects might be.
2. When cluster randomisation is adopted, we recommend that authors provide a clear justification for the choice of cluster randomisation and clearly outline strategies to mitigate increased risks of bias. This should include identification and recruitment by someone blind to the treatment allocation and minimal or objective individual-level eligibility criteria.
3. Other good conduct procedures which are routinely implemented in individually randomised trials should be followed. These include implementation of the randomisation using an accepted method of allocation concealment, for example, by using an independent statistician to generate the allocation sequence; blind outcome assessment when outcomes are subjective; and clear pre-specification (in a protocol or trial registration site) of the primary outcome including primary assessment time and method of primary analysis.
4. All these aspects should be clearly reported as per CONSORT guidelines. To ensure particular clarity around identification and recruitment, authors should also provide a timeline-cluster diagram.

Tab

Domain	Description
Domain 1a: Bias arising from the randomisation process	Randomisation refers to the process of allocating clusters to arms. Biases can arise if this allocation is not random or is not adhered to (at the level of the cluster).
Domain 1b: Bias arising from identification or recruitment of participants within clusters	When identification and recruitment of participants occurs with knowledge of the treatment allocation this can lead to differential recruitment and identification between treatment conditions.
Domain 2: Bias due to deviations from intended interventions	Trials which intend to measure the effect of offering treatment in everyday practice are unlikely to be conducted with blinding of the participant to allocated treatment. Deviations from the intended intervention can occur if those in the control condition receive the intervention condition (or vice versa). This is sometimes referred to as contamination or performance bias.

for cluster trials

Dor	Where the missingness is differential across treatment conditions, this can cause bias. Missingness can be differential across treatment conditions even when the proportion missingness is similar across conditions (for example when missingness is dependent on prognostic factors).
Domain 4: Bias in measurement of the outcome	Trials in which the treatment status is known by those assessing outcomes might be at risk of bias because of (subconscious) assessments of outcomes being preferential in one treatment condition. Outcomes which are objective (e.g., mortality) will be at reduced risk of this bias. This is sometimes referred to as outcome assessment bias.
Domain 5: Bias in selection of the reported result	Trials which do not pre-specify the primary outcome, along with primary assessment time, or clear method of analysis (including factors for adjustment) are at risk of selecting positive outcomes at the time of reporting.

.

Tab

Characteristic	n (%)
Publication year	
2007-2010	9 (22.5)
2011-2013	20 (50.0)
2014-2016	11 (27.5)
Country of study conduct	
Canada and/or USA	7 (17.5)
United Kingdom and/or EU	11 (27.5)
Australia	1 (2.5)
LMICs	21 (52.5)
Type of cluster	
Residential areas	15 (37.5)
Primary care practices	4 (10)
Individual health professionals	2 (5)
Hospitals, nursing homes, medical clinics or ICUs	15 (37.5)
Other	4 (10)
Rationale for cluster design^	
Avoid contamination	17 (42.5)
Practical reasons	14 (35)
Cluster level analysis	2 (5)
No justification	10 (25)
Other	10 (25)
Trial design	
Parallel arm	28 (70)
Factorial	3 (7.5)
Cross-over	6 (15)
Stepped wedge	3 (7.5)
Pre-specification documentation availability	
Accessible protocol paper or SAP	16 (40)
Trial registration	33 (82.5)
Neither protocol paper nor trial registration	6 (15)
Randomisation by independent statistician	11 (27.5)
Number of eligibility criterial at the individual level	
<3	16 (40)
>=3	24 (60)
Number of clusters*	
Median (IQR)	24 [12 – 49.5]
Average cluster size*	
Median (IQR)	114 [35 – 456]
Outcome objective	
Yes	30 (75%)
No	10 (25%)

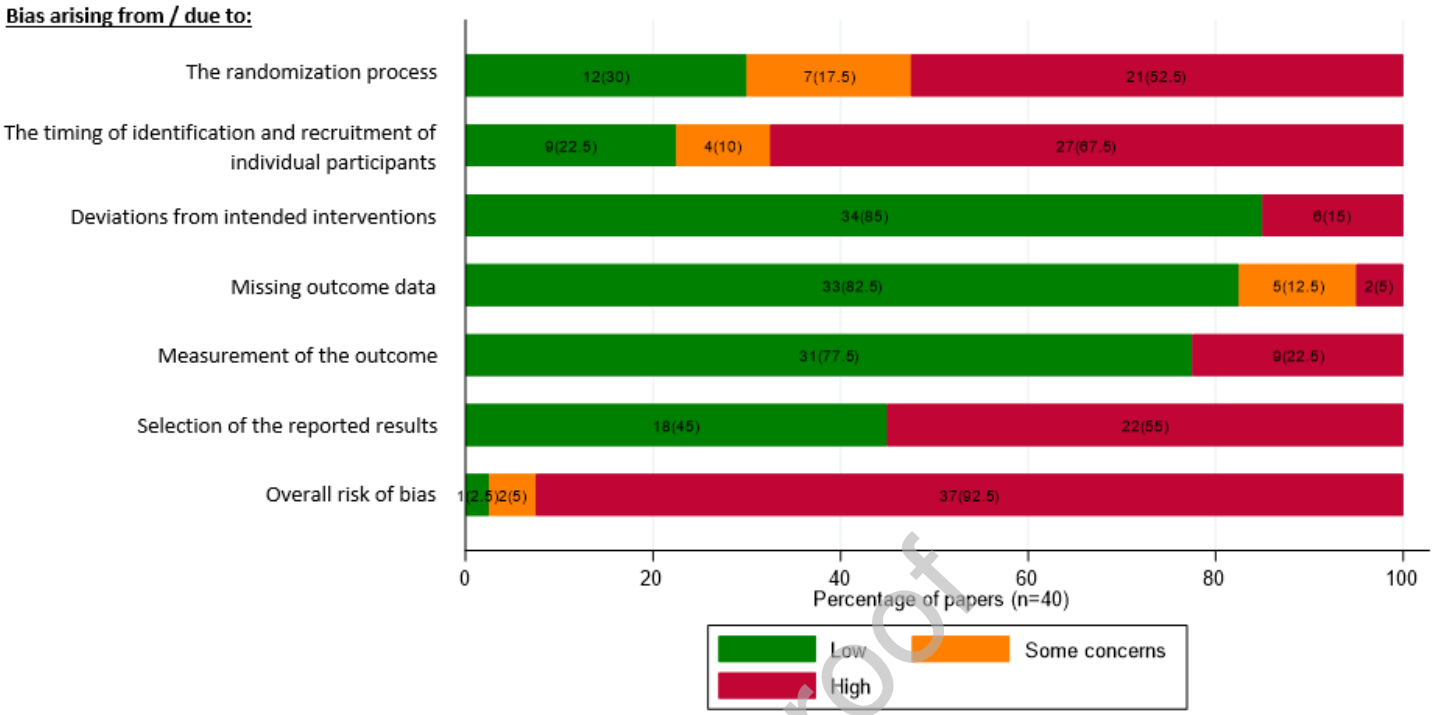
LMIC = Low- or Middle-Income Country; IQR= Interquartile range; ICU: Intensive Care Unit; SAP: Statistical Analysis Plan; *Numbers refer to realised numbers as opposed to those planned in any sample size calculation for example (i.e. the number of clusters randomised and the number of participants on whom baseline measures were taken); ^categories not mutually exclusive.

Tab

Domain	Level of Risk	n (%)	Reliability between reviewers	
		n=40	Gwet's AC (95% CI)	% agreement
1a - Bias arising from the randomization process	Low risk	12(30)	0.46 (0.20,0.72)	50
	Some concerns	7(17.5)		
	High risk	21(52.5)		
1b - Bias arising from the timing of identification and recruitment of individual participants	Low risk	9(22.5)	0.59 (0.37,0.81)	62.5
	Some concerns	4(10)		
	High risk	27(67.5)		
2 - Bias due to deviations from intended interventions	Low risk	34(85)	0.85 (0.74,0.96)	75
	Some concerns	0(0)		
	High risk	6(15)		
3 - Bias due to missing outcome data	Low risk	33(82.5)	0.77 (0.62,0.92)	67.5
	Some concerns	5(12.5)		
	High risk	2(5)		
4 - Bias in measurement of the outcome	Low risk	31(77.5)	0.79 (0.64,0.95)	75
	Some concerns	0(0)		
	High risk	9(22.5)		
5 - Bias in selection of the reported results	Low risk	18(45)	0.44 (0.19,0.70)	57.5
	Some concerns	0(0)		
	High risk	22(55)		
Overall risk of bias judgement[1]	Low risk	1(2.5)	0.92 (0.85,0.99)	82.5
	Some concerns	2(5)		
	High risk	37(92.5)		
Number of domains at high risk	0*	3 (7.5)		
	1	9 (22.5)		
	2	14 (35.0)		
	3	7 (17.5)		
	4	6 (15.0)		
	5	1 (2.5)		

[1] Overall risk of bias judgement: low risk of bias is defined as all domains at low risk of bias; some concerns is defined as at least one domain has some concerns but does not include any high risk of bias for any domain; and high risk of bias is defined as high risk of bias in at least one domain or some concerns for multiple domains; * 0 domains at risk includes 1 at low risk and 2 with some concerns (overall risk).

Figu



Author Statement

KH led the development of the project and wrote the first draft of the paper. MT led the search process and led the identification of studies for inclusion. CE designed and developed the data abstraction tools and conducted the statistical analysis. MT, SE and JT provided important oversight to the project. All authors helped develop the data abstraction tools, provided critical insight, contributed to the data abstraction exercise, and commented on the draft paper.

References

[Bolzern 2018] Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. J Clin Epidemiol. 2018 Jul;99:106-112.

[Brierley 2012] Brierley G, Brabyn S, Torgerson D, Watson J. Bias in recruitment to cluster randomized trials: a review of recent publications. J Eval Clin Pract. 2012 Aug;18(4):878-86.

[CIOMS 2016] International Ethical Guidelines for Health-related Research Involving Humans, Fourth Edition. Geneva. Council for International Organizations of Medical Sciences (CIOMS); 2016.

[Caille 2016] Caille A, Kerry S, Tavernier E, Leyrat C, Eldridge S, Giraudeau B. Timeline cluster: a graphical tool to identify risk of bias in cluster randomised trials. BMJ. 2016 Aug 16;354:i4291.

[Campbell 2012] Campbell MK, Piaggio G, Elbourne DR, Altman DG; CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012 Sep 4;345:e5661.

[Cook 2021] Cook DJ, Rutherford WB, Scales DC, Adhikari NKJ, Cuthbertson BH. Rationale, Methodological Quality, and Reporting of Cluster-Randomized Controlled Trials in Critical Care Medicine: A Systematic Review. Crit Care Med. 2021 Feb 12.

[Dia
residential facilities for older people suggests how to improve quality. BMC Med Res Methodol. 2013 Oct 22;13:127. doi: 10.1186/1471-2288-13-127.

[Edwards 1999] Edwards SJ, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. BMJ. 1999 May 22;318(7195):1407-9.

[Eldridge 2005] Eldridge SM, Ashby D, Feder GS. Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care. Clin Trials. 2005;2(2):91-8.

[Eldridge 2008] Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. BMJ. 2008 Apr 19;336(7649):876-80.

[Eldridge 2009] Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? BMJ. 2009 Oct 9;339:b4006.

[Eldridge 2012] Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester, Wiley, 2012

[Froud 2012] Froud R, Eldridge S, Diaz Ordaz K, Marinho VC, Donner A. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. Community Dent Oral Epidemiol. 2012 Feb;40 Suppl 1:3-14.

[Giraudeau 2009] Giraudeau B, Ravaud P. Preventing bias in cluster randomised trials. PLoS Med. 2009 May 26;6(5):e1000065.

[Gwet 2014] Gwet, K. 2014. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. 4th ed. Gaithersburg, MD: Advanced Analytics.

[Hahn 2005] Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. BMC Med Res Methodol. 2005 Mar 2;5:10.

[Hemming 2019] Hemming K, Carroll K, Thompson J, Forbes A, Taljaard M; SW-CRT Review Group. Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT: crowd-sourcing systematic review. J Clin Epidemiol. 2019 Mar;107:77-88.

[Hemming 2021] Hemming K, Taljaard M, Moerbeek M, Forbes A. Contamination: How much can an individually randomized trial tolerate? Stat Med. 2021 May 7.

[Higgens 2016] Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, Reeves B, Eldridge S. A revised tool for assessing risk of bias in randomized trials In: Chandler J, McKenzie J, Boutron I, Welch V (editors). Cochrane Methods. Cochrane Database of Systematic Reviews 2016, Issue 10 (Suppl 1).

[Hox 2014] Hox JJ, Moerbeek M, Kluytmans A, van de Schoot R. Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. Front Psychol. 2014 Feb 4;5:78. doi: 10.3389/fpsyg.2014.00078. PMID: 24550881; PMCID: PMC3912451.

[Kahan 2015] Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. Trials. 2015 Sep 10;16:405.

[Leyrat 2019] Leyrat C, Caille A, Eldridge S, Kerry S, Dechartres A, Giraudeau B. Intervention effect estimates in cluster randomized versus individually randomized trials: a meta-epidemiological study. Int J Epidemiol. 2019 Apr 1;48(2):609-619.

[Losilla 2018] Losilla JM, Oliveras I, Marin-Garcia JA, Vives J. Three risk of bias tools lead to opposite conclusions in observational research synthesis. J Clin Epidemiol. 2018 Sep;101:61-72.

[Minozzi 2019] Minozzi S, Cinquini M, Gianola S, Castellini G, Gerardi C, Banzi R. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. J Clin Epidemiol. 2019 Aug;112:28-35.

[Mi] for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. 2020 Oct;126:37-44. doi: 10.1016/j.jclinepi.2020.06.015. Epub 2020 Jun 18. PMID: 32562833.

[Moerbeek 2005] Moerbeek M. Randomization of clusters versus randomization of persons within clusters: which is preferable?. The American Statistician 2005; 59(2): 173–179.

[Murray 1998] Murray DM. Design and Analysis of Group Randomized Trials. New York, NY: Oxford University Press Inc; 1998.

[Puffer 2003] Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. BMJ. 2003 Oct 4;327(7418):785-9.

[Spence 2020] Spence O, Hong K, Onwuchekwa Uba R, Doshi P. Availability of study protocols for randomized trials published in high-impact medical journals: A cross-sectional analysis. Clin Trials. 2020;17(1):99-105.

[Taljaard 2017] Taljaard M, Hemming K, Shah L, Giraudeau B, Grimshaw JM, Weijer C. Inadequacy of ethical conduct and reporting of stepped wedge cluster randomized trials: Results from a systematic review. Clin Trials. 2017 Aug;14(4):333-341.

[Taljaard 2020] Taljaard M, Goldstein CE, Giraudeau B, et al. Cluster over individual randomization: are study design choices appropriately justified? Review of a random sample of trials. *Clin Trials*. 2020;17(3):253-263.

[Tielsch 2007] Tielsch JM, Darmstadt GL, Mullany LC, Khatry SK, Katz J, LeClerq SC, Shrestha S, Adhikari R. Impact of newborn skin-cleansing with chlorhexidine on neonatal mortality in southern Nepal: a community-based, cluster-randomized trial. Pediatrics. 2007 Feb;119(2):e330-40.

[Turner 2017a] Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 1-Design. Am J Public Health. 2017 Jun;107(6):907-915.

[Turner 2017b] Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis. Am J Public Health. 2017 Jul;107(7):1078-1086.

[Torgerson 2001]. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ. 2001 Feb 10;322(7282):355-7.

[Wongpakaran 2013] Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013 Apr 29;13:61. doi: 10.1186/1471-2288-13-61

[Yang 2017] Yang R, Carter BL, Gums TH, Gryzlak BM, Xu Y, Levy BT. Selection bias and subject refusal in a cluster-randomized controlled trial. BMC Med Res Methodol. 2017 Jul 10;17(1):94.