Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT

Hemming, Karla; Carroll, K.; Thompson, J; Forbes, A; Taljaard, Monica

DOI:
10.1016/j.jclinepi.2018.11.017

Document Version
Peer reviewed version

Citation for published version (Harvard):
Hemming, K, Carroll, K, Thompson, J, Forbes, A & Taljaard, M 2019, 'Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT: crowd-sourcing systematic review', *Journal of Clinical Epidemiology*, vol. 107. https://doi.org/10.1016/j.jclinepi.2018.11.017

The CONSORT extension for Stepped-Wedge Cluster Randomised Trials: baseline assessment of reporting quality - and assessment of inter-rater reliability using a crowd-sourcing systematic review

K Hemming[1], K Carroll[2], J Thompson[3], A Forbes[4] and M Taljaard[5] and the SW-CRT review group

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

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, Ontario, Canada. kecarroll@ohri.ca;

[3]Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, London, UK. Jennifer.thompson@lshtm.ac.uk;

[4]Biostatistics, Monash University, Melbourne, Australia. andrew.forbes@monash.edu;

[5]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.

Group Information: The SW-CRT Review Group authors appear at the end of the article.

Author contributions

KH led the development of the project, conducted the initial search, developed the data abstraction tools and wrote the first draft of the paper and conducted some of the statistical analysis. MT made a substantial contribution to all stages of the project; including screening studies for inclusion, helping develop the data abstraction tools and providing critical insight. KC invited participants and emailed relevant material; abstracted and compiled basic demographic summary for the included trials; and conducted the survey to elicit basic demographic information on review participants. JT was a reserve reviewer; advised on the statistical analysis; and provided critical insight. AF was a reserve reviewer, planned and conducted part of the statistical analysis; and provided critical insight. All group authors abstracted data from a published paper reporting a trial, participated in the joint assessment exercise, and were invited to comment on the draft paper.

Funding

This research was partly funded by the UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative. Karla Hemming is funded by a NIHR Senior Research Fellowship SRF-2017-10-002. Jennifer Thompson is funded by the Medical Research Council UK.

Summary

The Consolidated Standards Of Reporting Trials (CONSORT) extension for the stepped-wedge cluster randomised trial (SW-CRT) is a recently published reporting guideline for SW-CRTs. We assess the quality of reporting of a recent sample of SW-CRTs according to the 26 items in the new guideline using a novel crowd sourcing methodology conducted independently and in duplicate, with random assignment, by 50 reviewers. We assessed reliability of the quality assessments, proposing this as a novel way to assess robustness of items in reporting guidelines.

Several items were well reported. Some items were very poorly reported, including several items that have unique requirements for the SW-CRT, such as the rationale for use of the design, description of the design, identification and recruitment of participants within clusters, and concealment of cluster allocation (not reported in more than 50% of the reports). Agreement across items was moderate (median percentage agreement was 76% [IQR 64 to 86]). Agreement was low for several items including the description of the trial design and why trial ended or stopped for example.

When reporting SW-CRTs authors should pay particular attention to ensure clear reporting on the exact format of the design with justification, as well as how clusters and individuals were identified for inclusion in the study, and whether this was done before or after randomisation of the clusters, which are crucial for risk of bias assessments. Some items, including why the trial ended might either not be relevant to SW-CRTs, or might be unclearly described in the statement.

Introduction

The stepped-wedge cluster randomised trial (SW-CRT) is a novel cluster randomised trial design that is particularly valuable in implementation research and becoming more commonly used [Brown 2006; Mdege 2011, Martin 2016]. It is particularly relevant for evaluating service innovations in learning healthcare organisations [Hemming 2015]. The SW-CRT involves randomisation of clusters (e.g. primary care units, wards, hospitals) to different sequences that dictate the order in which each cluster will switch to the intervention condition [Hussey and Hughes 2007]. The CONSORT (Consolidated Standards Of Reporting Trials) statement, which outlines key characteristics to be reported in a parallel arm individually randomised trial [Schulz 2010], has recently been extended to provide a reporting guideline specific to SW-CRTs (referred to as the CONSORT extension for the SW-CRT) [Hemming 2018]. Whilst the earlier CONSORT extension for cluster randomised trials provides reporting guidance for trials in which groups of individuals (clusters) are randomised [Campbell 2012], there are many differences between the SW-CRT and the parallel cluster trial.

Key design characteristic of the SW-CRT (which do not typically feature in a parallel CRT conducted at a single cross section) include whether the design repeatedly measures the same individuals [Copas 2015]; whether the study is designed with an equal number of clusters allocated to each sequence; whether the study incorporates a transition period (a time period in which the intervention is embedded into practice) [Hemming 2015]; whether participants are recruited into the study before randomisation; whether cross-sectional samples are selected for outcome assessment; and whether outcomes are ascertained for a complete enumeration of the cluster members using routinely collected data. Furthermore, the design is also susceptible to several risks of biases particular to the SW-CRT: including confounding with time [Hemming 2017]; complex within-cluster correlation patterns [Girling 2016; Hooper 2016; Kasza 2017; Thompson 2018]; the risk of within-cluster contamination [Copas 2015; Hemming 2018]; the possibility of time varying treatment effects [Davey 2015, Hemming 2017]; and risks of recruitment or selection biases [Higgins 2016; Caille 2016].

Several systematic reviews have examined the quality of reporting of SW-CRTs. These have reported lack of clarity of the design, reporting of time adjustments, and key ethical aspects such as consent and ethical oversight [Brown 2006; Mdege 2011; Martin 2016; Grayling 2017; Taljaard 2017]. Whilst only about 40 completed SW-CRTs have been identified by these reviews, there has been an exponential increase in the use of this design over the past few years with an expected further increase in the near future.

Here we report the results of a systematic review of an assessment of the quality of reporting of a recent sample of SW-CRTs according to the newly developed reporting guideline. The aims of this review were to determine the quality of reporting of a recent set of SW-CRTs according to this newly developed guideline, to flag areas that are particularly poorly reported (to encourage immediate improved reporting), and to provide a baseline assessment for future studies examining any changes over time. This assessment of quality of reporting was conducted by 50 reviewers, with random assignment, so that each assessment of quality of reporting was performed independently by two reviewers. This was followed by a discussion to resolve differences, leading to a joint assessment. Whilst not a primary aim, we also assessed the reliability of the independent assessments. Measures of reliability of the independent assessments was used to suggest potential items in the statement where the wording might be unclear.

Methods

We identified a sample of recently published SW-CRT trials, which were then randomly allocated to 50 reviewers for quality assessment. We capitalised on the willingness, interest and skill set of participants (the reviewers) attending a workshop on the reporting of SW-CRTs in London, UK during November 2017. To allow for independent extraction by two reviewers per trial report, and to allow for the possibility of more participants than expected attending the workshop we sought to identify approximately 30 SW-CRT trial reports. Specific objectives were to assess agreement

between reviewers and to provide a joint assessment of the quality of reporting according to the new CONSORT extension for this trial design.

Scope of this review

We included SW-CRTs with a minimum of three sequences (where a sequence is defined as a unique allocation of periods spent in the control condition followed by periods in the intervention condition). One exception was made to this criteria to include trials with two sequences and three periods, which are also considered SW-CRTs by the new reporting guideline. We included only studies comparing two treatment conditions. We did not restrict inclusion to those designs with all clusters initiating in the control condition and ending up in the intervention condition. Only studies using the cluster as the unit of randomisation were included. We included only primary reports of SW-CRTs; protocols and reports of secondary analyses of a previously reported trial were excluded. Reports had to be open access or viewable from either the University of Birmingham or University of Ottawa libraries, and published in English. Focusing on the most recently published trials helped ensure our assessment of reporting quality was current and avoided any overlap with earlier systematic reviews of SW-CRTs. A protocol for the review was not registered with PROSPERO as it was out of remit.

Search strategy

We identified eligible studies in MEDLINE (accessed using PubMed) using a previously published electronic search strategy [Martin 2016, Table S1] run on the 21st November 2017 (including variations of the search terms "stepped wedge", "experimentally stated introduction", "delayed intervention"). We identified and ordered studies in descending order by date listed in Medline. To allow for exclusion of ineligible studies, titles and abstracts of the first 50 studies were identified. The title and abstracts of these 50 studies were screened independently by two authors (KH and MT). Those found not to meet the eligibility criteria were excluded. Any differences were resolved by discussion. Those studies not meeting the eligibility criteria were excluded. For the remaining studies, full copies of the trial reports were then obtained and assessed against the inclusion criteria.

Data abstraction process

Data abstraction was undertaken by delegates attending a workshop in London on the reporting of SW-CRTs on 30 November 2017. The workshop was run as part of an annual methodological meeting (including workshops and short invited and contributed talks) on current developments in cluster randomised and stepped-wedge trials. The workshop was a low cost (£50) event with open attendance, but for which pre-registration was necessary.

In preparation for the event we invited all registered delegates to participate in the review. Participants were informed that attending the workshop would require undertaking a small amount of work in advance and during the event, with an invitation to contribute to the resulting manuscript as a group author. Anyone not wishing to participate in the development of the paper was invited to opt out (by 24 November), and any collected data would be excluded should the participant so wish. Participants were also informed that data on inter-rater reliability as well as the quality of reporting would be anonymously evaluated. Participants were free to not participate in this activity whilst still attending the meeting.

Participants not opting out were randomly allocated to one of the trial reports using computer-generated numbers, so that two participants were allocated to each study. One week in advance of the meeting, a full PDF copy of their allocated study was e-mailed to each participant. Participants were asked to independently assess the quality of reporting according to the 26 items in the newly proposed CONSORT extension for SW-CRTs (using a simple tool provided Appendix 1). Participants were kept blind to the other allocated assessor of the same report until the morning of the workshop. Participants were requested to bring a hard copy of the completed quality of reporting assessment tool with them to the workshop.

At the beginning of the workshop participants listened to a 30-minute presentation (by KH and MT) summarising the most salient points of the new reporting guideline. After this, the two participants assigned to each trial report met for a 30 minute period, discussed any discrepancies and reached a consensus for each of the 26 quality assessment items (hereafter called the joint assessment). Two facilitators (KH and MT) provided advice on any issues of clarity. At the end of the workshop, participants submitted their completed independent assessments (which they were asked not to change during the joint assessment process) and joint assessments. Data were therefore abstracted independently and in duplicate for each trial report.

Data abstracted

For each item, participant reviewers were provided with the wording of the checklist item and asked to assess whether the item was reported in their assigned study (Appendix 1). Reviewers were asked to assess the quality of reporting for each of the 26 items according to a four-point scale (clearly reported in full / clearly but partially reported / unclearly reported / not reported). However, feedback from reviewers following the independent assessment indicated the need for a "not applicable" response in the scale. As a consequence, the response scale was changed to a five-point scale with the addition of "not applicable" for the joint assessment. All reviewers completing the joint assessment were made aware of this change of scale. Some independent assessments also independently decided to add a "not applicable" option; whereas others did not. Several of the 26 items have two or three sub-items, leading to a total of 40 items undergoing assessment. The data were abstracted and entered into an Excel database by one person (KH).

Statistical analysis

First, we described for each item, the reliability of the quality assessment using percentage agreement (*within item across pairs*) and the Gwet A1 statistic [Gwet 2013] using the *kappaetc* command in Stata 14, chosen because the study design was of reviewers nested within articles. Percentage agreements were compared using the four-point scale, and by dichotomising the four-point scale into a two-point scale (clearly reported including clearly reported in full or clearly but partially reported versus not clearly reported which includes unclearly reported or not reported). Independent reviews were excluded from the assessments of reliability if only one of the pair submitted their independent assessments (n=3). Any individual level items which were missing in either one or both independent assessments were also excluded (numbers included provided in tables), as were any assessments of not applicable because this was not included in the independent assessment tool. Items with low agreement might be considered to be items that are less clearly described in the statement than those with higher agreement.

We then described the joint assessment based on the consensus achieved between the reviewer pairs of reporting quality for all 26 items, along with the average number of items clearly reported for each report (or assessed as not applicable). Here, items assessed as not applicable were counted as clearly reported as all reviewers were aware of the inclusion of the category not applicable at the joint assessment stage and because this was deemed the best way not to penalise the assessment of quality of reporting when items were deemed not applicable. Any item assessments missing from the joint assessments were excluded (10 out of a total of 1000 possible assessments).

Finally, for completeness we describe for paper and each pair of reviewers the percentage agreement across items (*within pair across items*) and the corresponding measure of reliability (again excluding any assessments of not applicable, missing independent assessments and missing individual item assessments). Reviewer pairs with very low (or very high) agreement might indicate reviewers with low (or high) expertise or papers that are clearly (or unclearly) reported. Alongside this, we report the joint assessment of quality of reporting for each study.

Results

Identified studies

The initial search identified 437 potential articles for inclusion (Figure 1). The most recently published 50 of these were screened in duplicate. Of these 50 studies, 7 were excluded because they were not randomised; 7 because they were protocols and not full reports; 4 because they were individually randomised; 3 because they reported a secondary analysis; 1 because it was a methods paper; and an additional two because they were not accessible. Of these exclusions, two were excluded in the full paper screen (one secondary analysis; and one because it did not clearly randomise clusters); and the others were excluded on the abstract screen. This therefore meant a total of 24 studies were excluded; leaving 26 for inclusion. One study was later identified as not meeting the inclusion criteria by a reviewer and so was also excluded (individually randomised); leaving 25 trial reports for inclusion in the analysis. A description of the 25 studies is provided in Table 1.

Participant reviewers

A total of 53 participants registered for the workshop. Of these, three participants (AF, JT and group author JM who were all participants of the Delphi group in the development of the extension statement) were asked to be reserve reviewers. Of the remaining 50 participants four dropped out (one dropped out before the day, one was a duplicate registration and two did not attend on the day). One participant on the waiting list to attend the workshop was invited to attend to replace the participant who dropped-out before the event.

The three participants invited to act as reserves all attended the event and each acted as a replacement for the three participants who did not attend the workshop (two of these provided an independent assessment for their subsequently allocated paper, one of which was done un-blinded to the other reviewers assessment). At the end of the event, 47 independent assessments were provided some of which included missing items (two of the invited participants did not return their independent assessment at the end of the workshop) and we received 25 completed joint assessments on 25 trials.

The participants were mainly from the UK (84%); the majority worked in University settings (72%); and most were statisticians although a number were clinicians (18%) spanning a range of career levels; with most having being involved some way in a stepped-wedge trial and a small number having participated in development of CONSORT statements (Supplementary Table S1).

Reliability of quality assessments

Some items showed very high agreement (observer percentage agreement) across the independent assessments, *within items and across reviewer pairs* (Table 2; Supplementary Figure S1). For example, reviewers were in agreement 100% of the time as to whether the title was clearly reported (Item 1a). Reviewers were also very often in agreement for the reporting of Item 5 (Interventions); Item 10c (Consent); Item 23 (Trial registration); Item 24 (Protocol) and item 26 (Funding). Some items showed particularly poor agreement across independent assessments. These included Item 3a (trial design); item 8b (Type of randomisation); Item 11b (If relevant, description of the similarity of the treatments); item 13a (Flow diagram); Item 14b (Why the trial ended) and Item 22 (Interpretation) all in agreement less than 60% of the time. Over all 26 items the median percentage agreement was 53% ([IQR: 43-64]) and the median Gwet's A1 statistic was 0.41 (IQR 0.29-0.55) when comparing agreement across the four possible assessment categories. Median agreement increased to 76% [IQR: 67-86] when the comparison was made across a two point scale. The median Gwet's A1 statistic across all items on the two point score was 0.72 [IQR: 0.59 to 0.90]. As expected, reliability was higher on the two point scale compared to the four point scale.

Quality of reporting of SW CONSORT items in the 25 trial reports based on the joint assessment of reviewers

There was variability among trial reports in the quality of reporting (Table 3; Supplementary Figure S2), with the median number of items reported (clearly or partially) was 28 [IQR: 23 to 30] of a total of 40 separate items. Several items were assessed as being well reported in the joint assessment. These included Item 1b (structured summary); Item 4a (eligibility criteria) and Item 4b (setting and locations where data were collected); Item 5 (description of intervention and control conditions); Item 6a (outcomes); Item 15 (baseline table); Item 17a (results); Item 25

(sources of funding and other support, role of funder); and Item 26 (whether the study was approved by a research ethics committee); all clearly reported more than 80% of the time.

Other items were poorly reported. In particular, the following items were clearly reported less than 50% of the time (in order of appearance in the statement): Item 2a (rationale and justification for the trial design); Item 3a (description of trial design); Item 3b (important changes to methods after trial commencement); Item 6b (any changes to trial outcomes after the trial commenced); Item 11a (blinding); Item 19 (Important harms); Item 21 (generalisability); and Item 24 (where the full trial protocol can be accessed, if available). All items involving randomisation were also poorly reported. This included Item 8a (methods used to generate the random allocation); Item 8b (type of randomisation); Item 9 (specification that allocation was based on the clusters; description of any methods used to conceal the allocation from the clusters until after recruitment); Item 10a (who generated the randomisation); and Item 10b (mechanism by which individual participants were included in the trial (such as complete enumeration, random sampling)).

Quality of reporting for each paper and reliability within pairs of reviewers

Making comparisons *within reviewer pairs and across items*, the percentage agreement (on a two-point scale) across all items ranged between 38% (reviewer pair 9) and 83% (reviewer pair 22). The median Gwet's A1 statistic across all pairs was 0.42 [IQR: 0.36 to 0.51]. (Supplementary Table S2) The median number of items (out of the total of 40 sub-items) assessed as being clearly reported was similar between the independent assessments (median [IQR] are 26 [IQR 21 to 29] not shown in tables) and 28 [IQR 23 to 30] for the joint assessments - see above.

Discussion

Principal findings

We have reported, using an innovative review approach capitalising on the interests and expertise of delegates at a workshop, an assessment of quality of reporting of SW-CRTs according to the new extension of the CONSORT statement for this trial design. Perhaps unsurprisingly we have identified sub-optimal reporting. Items particularly poorly reported were those items detailing the justification and design of the study and information on the method of randomisation, including the method of recruitment and allocation of clusters and participants. Some items with low reporting are specific to the SW-CRT (such as justification of the design); others are not specific and include such things as identification of the trial protocol. Two items, Item 11b (description of the similarity of the treatments) and Item 14b (why the trial ended or stopped) might not be relevant to many SW-CRTs despite their inclusion in the updated statement.

Implications

Clear reporting of who recruited or identified participants and whether or not such individuals were blind to allocation is important so readers can determine the risks for recruitment and selection biases [Higgins 2016; Caille 2016]. We identified that a description of any methods used to conceal the allocation from the clusters until after recruitment was poorly reported, as was the mechanism by which individual participants were included in clusters for the purposes of the trial (such as complete enumeration or random sampling; continuous recruitment or ascertainment, or recruitment at a fixed point in time), including who recruited or identified participants [Copas 2015; Hemming 2018]. In parallel cluster trials, it has recently been proposed that trials should report a timeline diagram to allow clear identification of the timing of randomisation with respect to recruitment of clusters and participants [Caille 2016]. Whether such a diagram could be modified for the SW-CRT remains to be investigated.

We also identified suboptimal reporting of the format of the design and justification for use of the SW-CRT. Clear reporting of the exact form of the design is necessary to determine whether the appropriate sample size and analysis

methods have been used [Hooper 2016]. Clear reporting of the justification for use of the design is important to determine if the study is ethically appropriate [Hemming 2017b].

Whilst some basic items were well reported, such as the title, sources of funding, and ethics review, a substantial minority of reports failed to report these essential details and many failed to report where the full trial protocol can be assessed. We identified that 32% of this very recent sample of SW-CRTs did not report trial registration clearly. Trial registration is known to increase transparency of reporting of primary outcomes and will deter against selective reporting [Mathieu 2009; Killeen 2014; Azar 2015]. Related to this, an earlier in-depth review of the quality of reporting of the ethical conduct and reporting of SW-CRTs identified poor reporting and identification of the research participant, informed consent and ethical review [Taljaard 2017]. In light of this, when the CONSORT extension for the SW-CRT was developed it introduced an extra item to encourage transparent reporting of ethical review. This review echoes previous findings and highlights how a significant minority of SW-CRTs fail to report any ethical review processes or provide details of the trial protocol; despite these being endorsed by the International Committee of Medical Journal Editors. These findings might indicate a misconception that SW-CRTs are different to conventional trials and do not need to be reported conducted with similar standards.

What is already known

The CONSORT statement describes a minimum set of items that should be reported in any clinical trial to ensure transparent reporting. Without transparent reporting the validity of the methods and the results cannot be assessed and this can have wide ranging medical implications [Rennie 2001]. Studies published in endorsing journals are more likely to report this minimum set of items than those that do not [Turner 2012]; and reviews have suggested an increase in quality of reporting post publication of reporting guidelines [Ivers 2011; Chan 2017].

There have been several reviews of the quality of reporting of SW-CRTs. Early reviews however did not assess quality of reporting against the CONSORT items [Beard 2015]; other reviews, whilst assessing reporting against the CONSORT extension for cluster randomised trials, gave in-depth reviews, for example assessing quality of reporting of sample size [Martin 2016] or analysis items only [Davey 2015]. Reviews assessing quality of reporting against all of the items in the CONSORT extension for cluster trials have demonstrated, not surprisingly, sub-optimal reporting [Grayling 2017]. This sub-optimality might be explained partly by the necessary inclusion of older trials (to avoid an overly small sample size) and the lack of relevance or coherence of items written with an individually randomised or parallel cluster trial in mind.

Strengths and limitations

This methodological review used a novel approach to data abstraction, capitalising on the interest and willingness of a group of methodologists attending a small conference to undertake quality assessments. The number of articles reviewed was small (25) but represents a substantial proportion of the total number of stepped-wedge trials undertaken to date (40 completed trial reports up to February 2015) [Grayling 2017]. Furthermore, the date range was selected to be mutually exclusive with other recent methodological reviews [Martin 2016; Grayling 2017]. Being the most recent year at the time of the review, our results reflect current reporting practices. There was some variation across reviewers in degrees of expertise and knowledge of the design, but all participants were given a short briefing via the introductory presentation. Reports were randomly allocated to the reviewers and data were abstracted independently to the other reviewers and then in duplicate. One important limitation is that participants were asked to assess quality of reporting on a four-point scale for their independent assessment; this was changed to a five-point scale by adding the category “not applicable” for the joint assessment. On the four-point scale agreement was low; when agreement was considered on a two-point scale (i.e. reported partially or fully vs. not reported or unclearly reported) agreement improved. However, even on this two-point scale agreement was poor

for some items. Items with low agreement might be less clearly worded in the trial reports than items with higher agreement.

Whilst this review did not set out to explicitly identify items with unclear wording, some of our results might be suggestive of items with insufficient wording clarity. Any future revisions of the CONSORT statement for SW-CRTs might pay particular attention therefore to the wording and relevance of Items 11b (If relevant, description of the similarity of treatments) and 14b (Why the trial ended or was stopped). Any robust attempt to study the clarity of the wording of CONSORT items would ideally be pre-planned, and include a range of diverse stakeholders (trialists, principal investigators, clinicians etc.) and likely include a qualitative assessment as well as a quantitative assessment of agreement.

We also observed variability in agreements within pairs, with some pairs of reviewers consistently agreeing across items and other pairs of reviewers mostly disagreeing across items. Agreement may have been improved with additional training of assessors (there was no association between quality of the study report as assessed by the joint assessment and level of agreement of independent assessments – Table S2).

Conclusions

The CONSORT extension for SW-CRTs provides specific and tailored guidance for the reporting of SW-CRTs. Clear reporting is crucial to allow assessment of the risks of bias and transparent interpretation of the results of studies. Whilst this review echoes the findings of many other assessments of reporting reviews, that reporting needs to be improved, we provide clear guidance of items which require particular attention as they are frequently poorly reported. Of note, researchers should pay careful attention to the reporting of how clusters and participants are recruited or sampled; and whether this was done without knowledge of the randomised allocation. Exactly how the design, results and interpretation of a SW-CRT should be reported is different to that of individually randomised and parallel cluster trials. Any future updates of the CONSORT extension for SW-CRTs should pay careful attention to the wording and relevance of items with low agreement between reviewers, including why the trial stopped, which may not be relevant to many SW-CRTs. Future developments of guidelines for reporting might assess reliability before publication and as a final stage in the development of the reporting guideline.

SW-CRT review group authors:

Susan J Dutton Oxford Clinical Trials Unit, Centre for Statistics in Medicine, University of Oxford susan.dutton@csm.ox.ac.uk;
Vichithranie Madurasinghe Queen Marys University of London v.madurasinghe@qmul.ac.uk;
Katy Morgan London School of Hygiene and Tropical Medicine katy.morgan@lshtm.ac.uk;
Beth Stuart University of Southampton bls1@soton.ac.uk;
Katherine Fielding London School of Hygiene and Tropical Medicine katherine.fielding@lshtm.ac.uk;
Victoria Cornelius Imperial College London v.cornelius@imperial.ac.uk;
Elizabeth L. Turner Duke University liz.turner@duke.edu;
Richard Hooper Queen Marys University of London r.l.hooper@qmul.ac.uk;
Bruno Giraudeau Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours, France bruno.giraudeau@univ-tours.fr;
Paul T. Seed King's College London paul.seed@kcl.ac.uk;
Alecia Nickless University of Oxford alecia.nickless@phc.ox.ac.uk;
Michael Grayling Medical Research Council Biostatistics Unit mjg211@cam.ac.uk;
Mélanie Pragu Bordeaux University, Inria, Inserm U1219, France melanie.prague@inria.fr;
Sally Kerry Queen Marys University of London s.m.kerry@qmul.ac.uk;
Lauren Bell London School of Hygiene and Tropical Medicine Lauren.Bell@lshtm.ac.uk;
Eila Watson Oxford Brookes University ewatson@brookes.ac.uk;
Rafael Gafoor King's College London rafael.gafoor@kcl.ac.uk;
Nadine Marlin Queen Marys University of London n.marlin@qmul.ac.uk;
Emel Yorganci King's College London Cicely Saunders Institute emel.yorganci@kcl.ac.uk;
Lesley Smith Oxford Brookes University lesleysmith@brookes.ac.uk;
Murielle Mbekwe INSERM U1246, Tours, France muriellembekwe@yahoo.fr;
Steven Teerenstra Radboud University Nijmegen z824116@umcn.nl;
Claire Chan Queen Marys University of London c.l.chan@qmul.ac.uk;
Mirjam Moerbeek Utrecht University m.moerbeek@uu.nl;
Pamela Jacobsen King's College London pamela.jacobsen@kcl.ac.uk;
Simon Bond Cambridge Clinical trials Unit simon.bond@addenbrookes.nhs.uk;
Ben Jones Plymouth University ben.jones@plymouth.ac.uk;
John Preisser University of North Carolina, USA jpreisse@bios.unc.edu;
Mona Kanaan University of York mona.kanaan@york.ac.uk;
Catherine Hewitt University of York catherine.hewitt@york.ac.uk;
Christina Easter; University of Birmingham C.L.Easter@bham.ac.uk;
Tracy Pellatt-Higgins University of Kent t.pellatt-higgins@kent.ac.uk;
Laura Pankhurst NHS Blood and Transplant laura.pankhurst@nhsbt.nhs.uk;
Schadrac C. Agbla London School of Hygiene and Tropical Medicine schadrac.agbla@lshtm.ac.uk;
Sandra Eldridge Queen Marys University of London s.eldridge@qmul.ac.uk;
Robin G. Lerner University of Oxford robin.lerner@ndorms.ox.ac.uk;
Clémence Leyrat Department of Medical Statistics, London School of Hygiene and Tropical Medicine clemence.leyrat@lshtm.ac.uk;
Mark Pilling University of Cambridge mark.pilling@medschl.cam.ac.uk;
Julia R. Forman King's College London julia.forman@kcl.ac.uk;
Indrani Bhattacharya ICR-CTSU indrani.bhattacharya@icr.ac.uk;
Nicholas Magill King's College London nicholas.magill@kcl.ac.uk;
Jane Candlish University of Sheffield jane.candlish@sheffield.ac.uk;
Cliona McDowell NICTU cliona.mcdowell@nictu.hscni.net;
James Martin Birmingham University J.T.Martin@bham.ac.uk;
Caroline Kristunas University of Leicester cak21@le.ac.uk;
Elizabeth Allen London School of Hygiene and Tropical Medicine elizabeth.allen@lshtm.ac.uk;
Nadine Seward London School of Hygiene and Tropical Medicine nadine.seward@lshtm.ac.uk;
Elaine Nicholls Keele University e.nicholls@keele.ac.uk;
Bryony Dean Franklin, School of Pharmacy, University College of London bryony.deanfranklin@ucl.ac.uk;

Table1. **Characteristics of the included stepped-wedge cluster randomized trial study reports**

Characteristic	Frequency (%) N = 25 reports in total
Publication year	
2017	13 (52%)
2016	12 (48%)
Journal Impact Factor	
Median (IQR)	2.8 [2.1-9.8]
Country of study	
High income country	
UK or Ireland	5 (20%)
Netherlands	5 (20%)
Canada or US	3 (12%)
Australia	2 (8%)
Other	3 (12%)
Low income country	5 (20%)
Middle income country	1 (4%)
Multiple countries of different levels	1 (4%)
Type of setting	
Health-care	20 (80%)
Non health-care	5 (20%)
Type of clusters	
General practice	3 (12%)
Hospital/Ward/Specialities	9 (36%)
Nursing homes	2 (8%)
Other cluster in health setting	3 (12%)
Geographical unit	5 (20%)
Other / Unclear	3 (12%)
Number of sequences	
Three	5 (20%)
Four	6 (24%)
Five	2 (8%)
Six or more	11 (44%)
Unclear	1 (4%)
Median (IQR)	5 (4-8)
Number of clusters	
Three	1(4%)
Four	4 (16%)
Five – Ten	5 (20%)
Eleven - Fifteen	5 (20%)
Greater than Fifteen	10 (40%)
Median (IQR)	13 [6-20]

IQR: interquartile range.

Table 2: Percentage agreement for each item across reviewer pairs

		N (reviewers)	Four-point scale		Two-point scale	
			Agreement	Gwet's A1 statistic (95% CI)	Agreement	Gwet's Al statistic (95% CI)
Title and abstract	Item 1a	44	82%	0.78 (0.52: 1.00)	100%	1.00 (0.84: 1.00)
	Item 1b	44	55%	0.43 (0.13: 0.73)	86%	0.81 (0.54: 1.00)
Background and objectives	Item 2a	44	41%	0.25 (-0.05: 0.55)	73%	0.49 (0.08: 0.91)
	Item 2b	42	52%	0.41 (0.09: 0.73)	81%	0.68 (0.33: 1.00)
Trial design	Item 3a	44	45%	0.29 (0.00: 0.59)	59%	0.23 (-0.23: 0.68)
	Item 3b	34	65%	0.56 (0.22: 0.91)	76%	0.55 (0.10: 1.00)
Participants	Item 4a	44	50%	0.35 (0.05: 0.65)	64%	0.39 (-0.05: 0.83)
	Item 4b	42	43%	0.30 (-0.01: 0.60)	76%	0.64 (0.27: 1.00)
Interventions	Item 5	44	59%	0.49 (0.20: 0.77)	95%	0.94 (0.75: 1.00)
Outcomes	Item 6a	42	43%	0.29 (-0.02: 0.60)	76%	0.69 (0.36: 1.00)
	Item 6b	32	69%	0.64 (0.29: 0.99)	75%	0.61 (0.16: 1.00)
Sample size	Item 7a	42	52%	0.37 (0.06: 0.68)	71%	0.44 (0.01: 0.86)
	Item 7b	16	75%	0.72 (0.17: 1.00)	75%	0.69 (0.12: 1.00)
Randomisation	Item 8a	44	68%	0.60 (0.31: 0.88)	82%	0.66 (0.30: 1.00)
	Item 8b	42	43%	0.26 (-0.04: 0.57)	57%	0.16 (-0.31: 0.63)
Allocation	Item 9	44	59%	0.46 (0.16: 0.76)	86%	0.73 (0.40: 1.00)
Implementation	Item 10a	44	59%	0.51 (0.21: 0.81)	82%	0.75 (0.45: 1.00)
	Item 10b	42	43%	0.24 (-0.07: 0.55)	67%	0.35 (-0.10: 0.80)
	Item 10c	42	52%	0.38 (0.06: 0.69)	90%	0.82 (0.52: 1.00)
Blinding	Item 11a	30	60%	0.51 (0.12: 0.89)	80%	0.69 (0.26: 1.00)
	Item 11b	18	56%	0.44 (-0.05: 0.93)	56%	0.21 (-0.45: 0.88)
Statistical analysis	Item 12a	44	55%	0.41 (0.11: 0.72)	77%	0.63 (0.27: 0.99)
	Item 12b	40	50%	0.37 (0.04: 0.70)	85%	0.74 (0.39: 1.00)
Flow diagram	Item 13a	44	27%	0.05 (-0.22: 0.31)	59%	0.26 (-0.21: 0.72)
	Item 13b	42	29%	0.06 (-0.21: 0.33)	71%	0.47 (0.05: 0.89)
Recruitment	Item 14a	44	50%	0.34 (0.04: 0.64)	86%	0.73 (0.40: 1.00)
	Item 14b	28	36%	0.19 (-0.16: 0.55)	50%	0.02 (-0.55: 0.58)
Baseline data	Item 15	44	64%	0.54 (0.25: 0.83)	86%	0.81 (0.53: 1.00)
Numbers analysed	Item 16	42	19%	-0.08 (-0.31: 0.16)	62%	0.26 (-0.21: 0.72)
Outcomes and estimation	Item 17a	42	57%	0.45 (0.14: 0.75)	71%	0.55 (0.15: 0.95)
	Item 17b	30	40%	0.22 (-0.14: 0.59)	67%	0.34 (-0.17: 0.86)
Auxiliary analysis	Item 18	38	47%	0.30 (-0.02: 0.63)	68%	0.38 (-0.09: 0.85)
Harms	Item 19	34	71%	0.66 (0.32: 1.00)	88%	0.82 (0.48: 1.00)
Limitations	Item 20	40	45%	0.29 (-0.02: 0.59)	75%	0.57 (0.18: 0.96)
Generalisability	Item 21	36	56%	0.41 (0.06: 0.75)	67%	0.34 (-0.13: 0.82)
Interpretation	Item 22	40	40%	0.25 (-0.06: 0.56)	60%	0.37 (-0.09: 0.83)
Registration	Item 23	42	81%	0.77 (0.51: 1.00)	90%	0.82 (0.51: 1.00)
Protocol	Item 24	38	79%	0.74 (0.44: 1.00)	95%	0.91 (0.63: 1.00)
Funding	Item 25	42	76%	0.73 (0.46: 1.00)	90%	0.88 (0.63: 1.00)
Research ethics	Item 26	42	71%	0.62 (0.31: 0.93)	76%	0.59 (0.20: 0.98)
	All items (median IQR)		53% [43%, 64%]	0.41 [0.29, 0.55]	76% [67%, 86%]	0.62 [0.38, 0.74]

Four-point scale classifies each item as: clearly and fully reported / clearly but partially recorded / unclearly reported / not reported. Two-point scale classified each item as: clearly and fully reported or partially recorded / unclearly or not reported. Any “not applicable” are excluded. IQR: interquartile range. CI: Confidence Interval. N: Number of reviewers.

Table 3: Joint assessment of the quality of reporting for 26 items of CONSORT extension for the SW-CRT

Item	Detailed description	Clearly reported* Reported /N (%)
Item 1a	Identification as a stepped-wedge cluster randomised trial in the title.	15/25 (60%)
Item 1b	Structured summary of trial design, methods, results, and conclusions.	20/25 (80%)
Item 2a	Scientific background. Rationale for using a cluster design and rationale for using a stepped-wedge design	10/25 (40%)
Item 2b	Specific objectives or hypotheses	17/25 (68%)
Item 3a	Description and diagram of trial design including definition of cluster, number of sequences, number of clusters randomised to each sequence, number of periods, duration of time between each step, and whether the participants assessed in different periods are the same people, different people, or a mixture	13/24 (54%)
Item 3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons.	12/25 (48%)
Item 4a	Eligibility criteria for clusters and participants.	21/25 (84%)
Item 4b	Settings and locations where the data were collected.	21/23 (91%)
Item 5	The intervention and control conditions with sufficient details to allow replication, including how and when they were administered; whether the intervention was delivered at the level of the cluster, the individual, or both.	21/25 (84%)
Item 6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed.	20/25 (80%)
Item 6b	Any changes to trial outcomes after the trial commenced, with reasons	10/25 (40%)
Item 7a	How sample size was determined. Method of calculation and relevant parameters with sufficient detail so the calculation can be replicated. Assumptions made about correlations between outcomes of participants from the same cluster.	14/25 (56%)
Item 7b	When applicable, explanation of any interim analyses and stopping guidelines.	15/25 (60%)
Item 8a	Method used to generate the random allocation to the sequences of treatments.	10/24 (42%)
Item 8b	Type of randomisation; details of any constrained randomisation or stratification if used.	12/25 (48%)
Item 9	Specification that allocation was based on clusters; description of any methods used to conceal the allocation from the clusters until after recruitment.	10/25 (40%)
Item 10a	Who generated the randomisation schedule, who enrolled clusters, and who assigned clusters to sequences	5/25 (20%)
Item 10b	Mechanism by which individual participants were included in clusters for the purposes of the trial (such as complete enumeration, random sampling; continuous recruitment/ascertainment, or recruitment at a fixed point in time), including who recruited or identified participants.	11/25 (44%)
Item 10c	Whether, from whom and when consent was sought and for what; whether this differed between treatment conditions.	14/25 (56%)
Item 11a	If done, who was blinded after assignment to sequences (for example, cluster level participants, individual level participants, those assessing outcomes) and how.	7/25 (28%)

Item 11b	If relevant, description of the similarity of treatments.	16/24 (67%)
Item 12a	Statistical methods used to compare treatment conditions for primary and secondary outcomes including how time effects, clustering and repeated measures were taken into account.	17/25 (68%)
Item 12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses.	17/25 (68%)
Item 13a	For each treatment condition or allocated sequence the numbers of clusters and participants who were assessed for eligibility, were randomly assigned, received intended treatments and were analysed for the primary outcome.	16/25 (64%)
Item 13b	For each treatment condition or allocated sequence, losses and exclusions for both clusters and participants with reasons.	14/25 (56%)
Item 14a	Dates defining the steps, initiation of intervention and deviations from planned dates. Dates defining recruitment and follow-up for participants.	14/25 (56%)
Item 14b	Why the trial ended or was stopped	17/25 (68%)
Item 15	Baseline characteristics for the individual and cluster levels as applicable for each treatment condition or allocated sequence.	22/25 (88%)
Item 16	The number of observations and clusters included in each analysis for each treatment condition and whether the analysis was according to the allocated schedule.	16/25 (64%)
Item 17a	For each primary and secondary outcome, results for each treatment condition, and the estimated effect size and its precision (such as 95% confidence interval); any correlations and time effects estimated in the analysis.	21/25 (84%)
Item 17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended.	18/25 (72%)
Item 18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory.	16/24 (67%)
Item 19	Important harms or unintended effects in each treatment condition (for specific guidance see CONSORT for harms).	9/25 (36%)
Item 20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses.	16/24 (67%)
Item 21	Generalisability (external validity, applicability) of the trial findings. Generalisability to clusters and/or individual participants (as relevant).	11/24 (46%)
Item 22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence.	19/24 (79%)
Item 23	Registration number and name of trial registry.	17/25 (68%)
Item 24	Where the full trial protocol can be accessed, if available.	7/25 (28%)
Item 25	Sources of funding and other support (such as supply of drugs), role of funders.	23/24 (96%)
Item 26	Whether the study was approved by a research ethics committee, with identification of the review committee(s). Justification for any waiver or modification of informed consent requirements.	20/25 (80%)
All	Total (out of 40) number of items clearly or fully reported (median [IQR])	28 [23 to 30]

*Includes not applicable;

Table S1: Characteristics of participant reviewers performing data-abstraction in quality assessment review

	N (%) N=50
Highest Career Level	
PhD student (completed or in progress)	9(18%)
MSc student (completed or in progress)	6(12%)
Post doc researcher	8(16%)
Lecturer	4(8%)
Senior lecturer/ Reader	7(14%)
Professor	13(26%)
*Other	3(6%)
Main Occupation	
Methodologist (Statistician)	39 (78%)
Trialist	2(4%)
Other (Clinician, other type of researcher)	9(18%)
Type of Setting	
Health-care	2(4%)
University	36(72%)
Health-care and University	12(24%)
Country of Work	
United Kingdom	42(84%)
France	3(6%)
United States	2(4%)
Netherlands	2(4%)
Australia	1(2%)
Previous involvement in SW-CRT CONSORT statement	
Participated in Delphi Survey	6(12%)
Co-author	7(14%)
No	37(74%)
Previous involvement in other CONSORT statements	
Yes	7(14%)
No	43(86%)
Previous experience with SW-CRTs	
Yes; one trial	15(30%)
Yes; two trials	6(12%)
Yes; three or more trials	8(16%)
No experience	21(42%)

*Statistician (non-academic); research statistician; medical professional

Table S2: Percentage agreement for each reviewer pair across all items

		Independent assessments		Joint assessment
Paper ID	**Reviewers**	**Agreement**	**Gwet's A1 statistic**	**Number items clearly reported^ (max 40)**
Study ID 2	Pair 1	56.76%	0.47 (0.25: 0.69)	11 (28%)
Study ID 3	Pair 2			
Study ID 7	Pair 3	55.56%	0.47 (0.26: 0.68)	27 (68%)
Study ID 8	Pair 4			
Study ID 9	Pair 5	43.24%	0.33 (0.13: 0.53)	*22 (61%)
Study ID 12	Pair 6	60.53%	0.53 (0.33: 0.73)	23 (58%)
Study ID 18	Pair 7	51.43%	0.42 (0.21: 0.63)	*26 (67%)
Study ID 20	Pair 8			
Study ID 21	Pair 9	38.24%	0.26 (0.05: 0.47)	29 (73%)
Study ID 22	Pair 10	41.03%	0.30 (0.10: 0.50)	*19 (49%)
Study ID 23	Pair 11	50.00%	0.41 (0.20: 0.62)	30 (75%)
Study ID 29	Pair 12	48.65%	0.39 (0.18: 0.59)	29 (73%)
Study ID 31	Pair 13	50.00%	0.42 (0.21: 0.62)	30 (75%)
Study ID 32	Pair 14	39.39%	0.28 (0.07: 0.50)	23 (58%)
Study ID 33	Pair 15	61.54%	0.55 (0.35: 0.74)	28 (70%)
Study ID 34	Pair 16	58.33%	0.52 (0.31: 0.73)	34 (85%)
Study ID 37	Pair 17	44.44%	0.34 (0.13: 0.55)	26 (65%)
Study ID 39	Pair 18	54.29%	0.47 (0.25: 0.69)	12 (30%)
Study ID 43	Pair 19	56.76%	0.49 (0.28: 0.70)	27 (68%)
Study ID 44	Pair 20	64.52%	0.60 (0.37: 0.83)	32 (80%)
Study ID 45	Pair 21	51.35%	0.42 (0.21: 0.62)	*28 (74%)
Study ID 46	Pair 22	82.86%	0.80 (0.62: 0.97)	26 (65%)
Study ID 47	Pair 23	48.28%	0.41 (0.17: 0.65)	9 (23%)
Study ID 49	Pair 24	45.16%	0.35 (0.12: 0.58)	29 (73%)
Study ID 50	Pair 25	75.00%	0.71 (0.55: 0.88)	10 (25%)
All	Average (median [IQR])		0.42 [0.36, 0.51]	

Each pair represents two reviewers who independently reviewed the same paper and assessed quality of reporting on the two point scale for 26 items. Each pair reviewed a different paper. Any items rated as not applicable by reviewers are treated as a missing assessment (as this was not included on the original quality assessment form). Pairs with missing agreement returned only one independent assessment and not two. *At least one quality assessment for one item was missing in the joint assessment and denominator for percentage not 40. ^In joint assessment.

References

[Azar 2015] Azar M, Riehm KE, McKay D, Thombs BD. Transparency of Outcome Reporting and Trial Registration of Randomized Controlled Trials. PLOS ONE. 2015 Nov 18;10:e0142894.

[Beard 2015] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials. 2015 Aug 17;16:353.

[Brown 2006] Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol. 2006 Nov 8;6:54.

[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; for the CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012 Sep 4;345:e5661.

[Chan 2017] Chan CL, Leyrat C, Eldridge SM. Quality of reporting of pilot and feasibility cluster randomised trials: a systematic review. BMJ Open. 2017 Nov 8;7(11):e016970.

[Copas 2015] Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials. 2015 Aug 17;16:352.

[Davey 2015] Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials. 2015 Aug 17;16:358.

[Grayling 2007] Grayling MJ, Wason JM, Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials. 2017 Jan 21;18(1):33.

[Gwet 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;13:61. doi:10.1186/1471-2288-13-61.

[Hemming 2015] Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ. 2015 Feb 6;350:h391.380.

[Hemming 2017] Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials. 2017 Mar 4;18(1):101.

[Hemming 2017b] K Hemming, S Eldridge, G Forbes, C Weijer, M Taljaard How to design efficient cluster randomised trials BMJ. 2017; 358: j3064.

[Hemming 2018] Hemming K, Taljaard M, et al. The CONSORT extension for Stepped-Wedge Cluster Randomised Trials. BMJ *to appear*

[Girling 2016] Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med. 2016 Jun 15;35(13):2149-66.

[Higgins 2016] Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. 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;10(Suppl 1).

[Hooper 2016] Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med. 2016 Nov 20;35(26):4718-28.

[Hussey 2007] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007 Feb;28(2):182-91.

[ICMJE] International Committee of Medical Journal Editors [http://www.icmje.org/]. Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals [16/05/2017] Available from: http://www.ICMJE.org.

[Ivers 2011] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, Skea Z, Brehaut JC, Boruch RF, Eccles MP, Grimshaw JM, Weijer C, Zwarenstein M, Donner A. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ. 2011 Sep 26;343:d5886. doi: 10.1136/bmj.d5886.

[Ivers 2012] Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, Upshur R, Zwarenstein M. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. Trials. 2012 Aug 1;13:120. doi: 10.1186/1745-6215-13-120.

[Kasza 2017] Kasza J, Hemming K, Hooper R, Matthews J, Forbes AB; ANZICS Centre for Outcomes & Resource Evaluation (CORE) Committee. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Methods Med Res. 2017 Jan 1:962280217734981.

[Killeen 2014] Killeen SMDF, Sourallous PM, Hunter IAPF, Hartley JEMDBF, Grady HLOMDF. Registration Rates, Adequacy of Registration, and a Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials Published in Surgery Journals. Ann Surg. 2014:259(1):193-6

[Kottner 2011] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011 Jan;64(1):96-106.

[Mathieu 2009] Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA. 2009;302(9):977-84.

[Martin 2016] Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ Open. 2016 Feb 4;6(2):e010166.

[Martin 2017] Martin J. Advancing knowledge in stepped-wedge cluster randomised trials (Unpublished doctoral thesis). University of Birmingham, UK. 2017.

[Mdege 2011] Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol. 2011 Sep;64(9):936-48.

[Moher 2010] Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 2010 Feb 16;7(2):e1000217.

[Prost 2015] Prost A, Binik A, Abubakar I, Roy A, De Allegri M, Mouchoux C, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials. 2015 Aug 17;16:351.

[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.

[Thompson 2017] Thompson JA, Fielding KL, Davey C, Aiken AM, Hargreaves JR, Hayes RJ. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. Stat Med. 2017 Oct 15;36(23):3670-3682.

[Turner 2012] Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Syst Rev. 2012 Nov 29;1:60.

[Rennie 2001] Rennie D. CONSORT revised--improving the reporting of randomized trials. JAMA. 2001 Apr 18;285(15):2006-7.

[Schulz 2010] Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.