PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A systematic review finds major deficiencies in sample size methodology and reporting for stepped wedge cluster randomised trials
AUTHORS	Martin, James; Taljaard, Monica; Girling, Alan; Hemming, Karla

VERSION 1 - REVIEW

REVIEWER	Mike Campbell University of Sheffield
REVIEW RETURNED	27-Jul-2015

GENERAL COMMENTS	Stepped wedge cluster trials are an exciting and developing field of research. This appears as a well conducted and written study. However, this paper identified only 32 trials published in 27 years, which shows that they are a very rare design indeed. Admittedly there were only 28 trials and protocols before 2012, compared with 32 after, so it is a burgeoning field (albeit still very rare). There have already been two reviews in the area although neither looked at the CONSORT statement. 1) The message of this paper is not very new. It is well established already that reporting of a complex methodology is poor. 2) There have been many other reviews of this type in different areas of trial methodology. For example. a similar review was conduct and published in the BMJ with regard sample size (1) . It would be useful to cite this to justify publication in BMJ. 3) The publication of protocols is relatively new, so this could account for some of the increasing trend. It wasn't clear whether some of the protocols were in fact for trials which were also included in the review, which would mean the outcomes were not completely independent and which may affect the p-values in Tables 3 and 4. 4) The p-values are of little use without some comment on the multiple testing issue, and the fact that there is little power to detect

	differences. It is BMJ convention to give the difference in proportions and a confidence interval, not just a p-value. Minor 1. Refs 1 and 7 are the same. Since refs 2 and 3 are books I suspect ref 1 should be Campbell MJ, Walters SJ. How to design, analyse and report cluster randomised trials in medicine and health services research. John Wiley and Sons, 2014. 2 Ref 13 and 18. The author list not standard. Reference 1 Clark T, Berger U and Mansmann U Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review. BMJ 2013;346:f1135

REVIEWER	Toby Prevost King's College London
REVIEW RETURNED	28-Jul-2015

GENERAL COMMENTS	1) Impact - Three factors are highlighted in the objectives - accounting for (1)clustering, (2) time effects and (3) repeated measurements on individuals. Neglecting to account for (1) clustering when determining a sample size can easily have a major under-estimation effect (e.g. two-fold) and provide a wholly inadequate trial sample size. Uncertainty in anticipating the degree of clustering (i.e. the size of the eventual ICC) can also affect power to this degree of magnitude. How serious are (2) and (3)? Are the effects of these equivalent to losing maybe 10% power generally and so not as important? It would be good to know how relatively important these two factors were across the range of design parameters in these 60 trials. Novelty in the findings lies less in (1). 2) In practice at the design stage, how easy is it reliably to pre-specify an accurate ICC, time effect and correlations spread over time periods. Are there implications for pilot work objectives, or setting and adapting the parameters (sample size, clusters, time periods) to emerging estimates. If it is not easy then will any methodological development be implemented (properly). 3) Abstract conclusions - the conclusions include a need for methodological development. Is there a greater need to encourage further good practice in trial design, and so should the conclusion also encourage further improvements in dissemination and training into practice and reporting. Does the methodological development refer only to accounting for repeated measurements in individuals in the sample size calculation and is this enabled only when good estimates of correlations over time periods are reliably available? 4) Page 10 line 54 "normality". For sample size calculation, is the use

	of normal distribution methods for means always that bad (or inaccurate) for proportions (which are themselves means of binary outcomes) given that trials have quite decent sample sizes and event information? e.g. Alan Donner developed sample size methods for the clustered difference in proportions between trial arms. 5) Early in the article, it would be useful to have a schematic representation of the most common designs that people are adopting in the literature of stepped wedge trials. Not all readers will know the different between a cohort design and an open cohort design, and maybe such a graph could illustrate key factors, e.g. time effect and repeated measures, for the readers. It is at quite a late stage (Page 13 line 5) that is becomes clearer that the review "only included designs in which is was intended that all clusters would ultimately receive the intervention", and omitting a third (sizeable or not?) group of trials that fall within the Stepped Wedge design family targeted by the article. 6) Although the focus in on the sample size (design end) it would be interesting to report more on the analysis end. For example, although neglecting to account for a time trend in the mis-calculation of sample size, the effect of omitting it also (or instead) at the analysis end would cause a bias which could be more important than inaccurate precision (Page 11 line 53). How many of the trials had such a potential bias i.e. did not report including a time factor in the analysis? Similarly, for another of the 3 key factors in the objectives, how many trials forgot to achieve increased variance (and wider confidence intervals) from the correlation in repeated measures on the same individual. 7) When comparing designs (parallel CRT versus SW-CRT), Page 11 line 58, the description was too brief to be able to be clear about what was kept constant and what else (data-wise) was different apart from the design. e.g. in this comparison, are the number of clusters the same and the total number of observed measurements the same, or the number of participants the same. Is the trial achieved in shorter research time under one design than the other. 8) The results were very interesting, revealing low and improving reporting adherence to recommended practice, with room for further improvement, increasing use of the design, uncertainty to research, and the review seems very thorough.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1 Mike Campbell

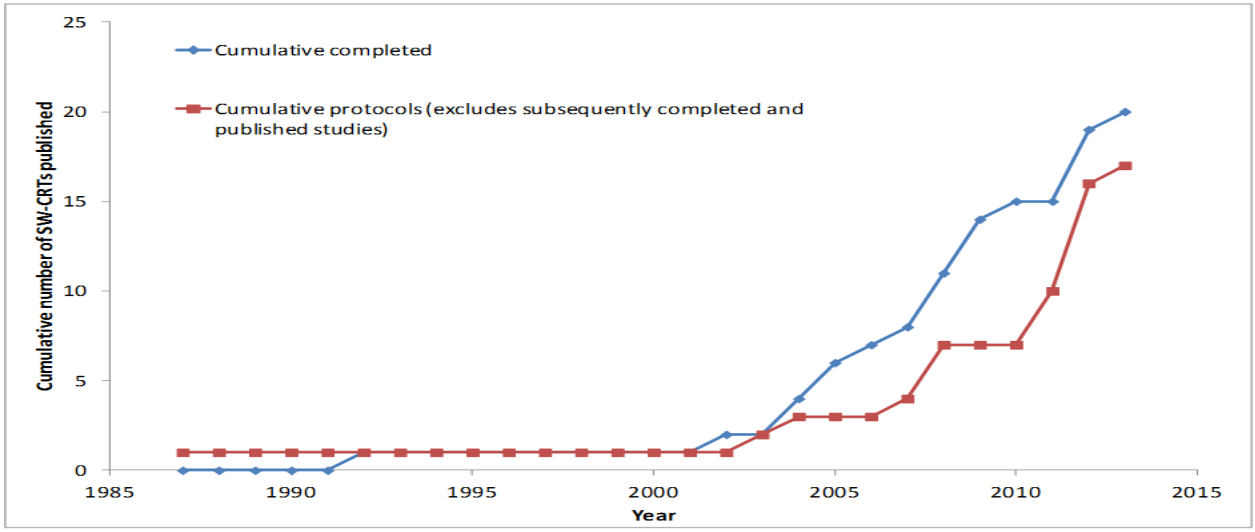
Comments:

Stepped wedge cluster trials are an exciting and developing field of research. This appears as a well conducted and written study.

The reviewer is correct in identifying that stepped wedge studies are a new and developing field of research and that our review has been conducted to a high standard.

However, this paper identified only 32 trials published in 27 years, which shows that they are a very rare design indeed. Admittedly there were only 28 trials and protocols before 2012, compared with 32 after, so it is a burgeoning field (albeit still very rare). There have already been two reviews in the area although neither looked at the CONSORT statement.

Whilst we did identify only 32 fully published trials, we identified a similar number of published protocols and of these almost all had been published in the past few years. The rate at which the trial design is being used is increasing rapidly. Furthermore, this number of papers is sufficient already to identify poor practice. Waiting until there have been more poorly reported trials, and a greater number of papers thus to be included, will only result in more studies being poorly conducted. Highlighting these poor practices now, can only be beneficial.



The message of this paper is not very new. Reporting of a complex methodology is poor.

Whilst trials of complex interventions share many similarities with other similar designs, there are some very specific reporting issues which relate specifically to the SW-CRT. Highlighting areas of poor performance in the use of this design early-on, will help mitigate poor practice becoming coming place. This review will also identify items for inclusion in a CONSORT extension, or a future update of the CONSORT extension for cluster randomised trials.

There have been many other reviews of this type in different areas of trial methodology. For example, a similar review was conduct and published in the BMJ with regard sample size. It would be useful to cite this to justify publication in BMJ.

The BMJ has published several reviews of this type, albeit for different study designs. As the reviewer points out the quality of reporting (and actually conduct too) of complex trial designs is poor. Publishing review papers highlighting this in high impact journals has the greatest potential of reaching researchers and having an impact. Publishing papers highlighting poor quality of reporting in low impact journals is unlikely to have an impact on practice. The review mentioned will be cited. There are specific issues relating to the SW-CRT which are different to other designs.

The publication of protocols is relatively new, so this could account for some of the increasing trend. It wasn't clear whether some of the protocols were in fact for trials which were also included in the review, which would mean the outcomes were not completely independent and which may affect the p-values in Tables 3 and 4.

We have additionally looked at the differences between protocols and full reports, these results are mentioned but not presented in the paper. The differences between protocols and full reports do not explain our findings. We have also clarrifed here that we did not "double count", i.e., we never had the abstractions for the same study (protocol+final report) - so all the results are independent and p-values are not affected in Talbes 3 and 4.

The p-values are of little use without some comment on the multiple testing issue, and the fact that there is little power to detect differences. It is BMJ convention to give the difference in proportions and a confidence interval, not just a p-value.

We agree and have added confidence intervals for absolute differences to the manuscript. Other minor points identified by the reviewer but not listed here have been corrected.

Reviewer: 2 Toby Prevost

Comments:

Impact - Three factors are highlighted in the objectives - accounting for (1)clustering, (2) time effects and (3) repeated measurements on individuals. Neglecting to account for (1) clustering when determining a sample size can easily have a major under-estimation effect (e.g. two-fold) and provide a wholly inadequate trial sample size. Uncertainty in anticipating the degree of clustering (i.e. the size of the eventual ICC) can also affect power to this degree of magnitude. How serious are (2) and (3)? Are the effects of these equivalent to losing maybe 10% power generally and so not as important? It would be good to know how relatively important these two factors were across the range of design parameters in these 60 trials. Novelty in the findings lies less in (1).

Whereas sample size calculations that don't allow for the effect of clustering are likely to lead to underpowered SW-CRTs, those that do not allow for the effect of time might lead to studies being either under- or over-powered. For the case where the ICC is low, designing a SW-CRT using methodology for a parallel study is likely to lead to an under-powered study. Absolute differences in magnitude of power might be fairly low, for example, in the region of 10%. However, when the ICC is higher, designing a SW-CRT using methodology for a parallel design is bound to lead to an over powered design at the expense of including vast numbers of observations which may contribute little to the power. We have references to support this.

In practice at the design stage, how easy is it reliably to pre-specify an accurate ICC, time effect and correlations spread over time periods. Are there implications for pilot work objectives, or setting and adapting the parameters (sample size, clusters, time periods) to emerging estimates. If it is not easy then will any methodological development be implemented (properly).

Researchers need to specify estimates of ICCs in advance — as with any other cluster trial. Whilst allowance for time effects are needed in the power calculation, these do not require any judgements or estimations, but are simply based on setting the number of steps, the number of clusters randomised per step and the average size of the cluster per step.

Abstract conclusions - the conclusions include a need for methodological development. Is there a greater need to encourage further good practice in trial design, and so should the conclusion also encourage further improvements in dissemination and training into practice and reporting. Does the methodological development refer only to accounting for repeated measurements in individuals in the sample size calculation and is this enabled only when good estimates of correlations over time periods are reliably available?

As expected, the quality of reporting of sample size items in stepped-wedge trials is sub-optimal. The vast majority of studies are using a sample size methodology that doesn't match their design. There is an urgent need for dissemination of the appropriate power methodology, guidelines for reporting and

methodological development to match the proliferation of the use of this design in practice. Time effects and repeated measures should be considered in all SW-CRT power calculations; and there should be clarity in reporting trials as cohort or cross-sectional designs.

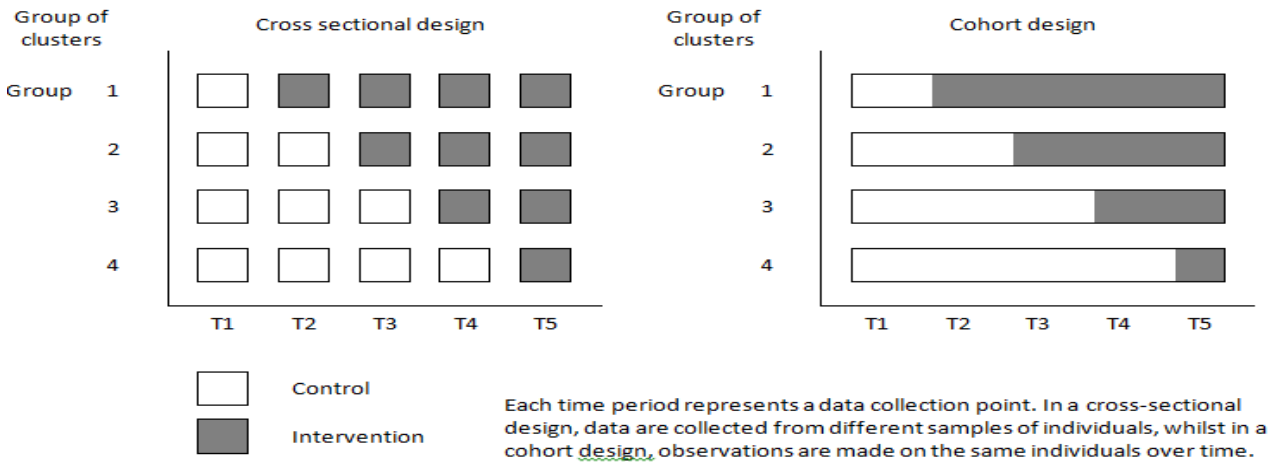
Page 10 line 54 "normality". For sample size calculation, is the use of normal distribution methods for means always that bad (or inaccurate) for proportions (which are themselves means of binary outcomes) given that trials have quite decent sample sizes and event information? e.g. Alan Donner developed sample size methods for the clustered difference in proportions between trial arms.

We have added some brief information on the sample size calculations for binary outcomes.

Early in the article, it would be useful to have a schematic representation of the most common designs that people are adopting in the literature of stepped wedge trials. Not all readers will know the different between a cohort design and an open cohort design, and maybe such a graph could illustrate key factors, e.g. time effect and repeated measures, for the readers. It is at quite a late stage (Page 13 line 5) that is becomes clearer that the review "only included designs in which is was intended that all clusters would ultimately receive the intervention", and omitting a third (sizeable or not?) group of trials that fall within the Stepped Wedge design family targeted by the article.

A schematic representation, focusing on the difference between the cohort and cross-sectional design has been added to the paper. We have also clarified the description in the text as all stepped wedge studies have been included, contrary to the impression we seem to have provided to the reviewer.

Figure 1 Schematic illustration of the stepped-wedge cluster randomised trial



Although the focus in on the sample size (design end) it would be interesting to report more on the analysis end. For example, although neglecting to account for a time trend in the mis-calculation of sample size, the effect of omitting it also (or instead) at the analysis end would cause a bias which could be more important than inaccurate precision (Page 11 line 53). How many of the trials had such a potential bias i.e. did not report including a time factor in the analysis? Similarly, for another of the 3 key factors in the objectives, how many trials forgot to achieve increased variance (and wider confidence intervals) from the correlation in repeated measures on the same individual.

This is an interesting point. Bias in estimates of treatment effects is viewed by many to be more important than any lack of precision; and biases will not be prevalent unless the study did not take into account the time effects at the analysis stage. Whether or not allowances for time effects at the

analysis stage are more frequent than at the design stage is not something we have considered, though it would seem unlikely that these mistakes are rectified if they are missed at the design stage. We did extract this information in our review and found that 53% of the studies adjusted for time effects at the analysis stage. This could be incorporated into a revision of the paper.

When comparing designs (parallel CRT versus SW-CRT), Page 11 line 58, the description was too brief to be able to be clear about what was kept constant and what else (data-wise) was different apart from the design. e.g. in this comparison, are the number of clusters the same and the total number of observed measurements the same, or the number of participants the same. Is the trial achieved in shorter research time under one design than the other.

We have clarified this in our paper.

The results were very interesting, revealing low and improving reporting adherence to recommended practice, with room for further improvement, increasing use of the design, uncertainty to research, and the review seems very thorough.

We agree with this very positive comment.

VERSION 2 – REVIEW

REVIEWER	Dr Yannan Jiang The University of Auckland New Zealand
REVIEW RETURNED	24-Oct-2015

GENERAL COMMENTS	As part of the development of an extension to the CONSORT statement for stepped-wedge cluster randomised trials (SW-CRTs), this paper led by Hemming et al. has undertaken a methodological review of SW-CRTs published between 1987 and 2014 with specific focus on the sample size methodology used. Basic trial demographics of included studies (28 trial protocols and 32 full reports) and realised design features were summarised. Adherence to reporting each of the sample size items recommended in the CONSORT 2010 statement and the 2012 extension to cluster randomised trials were assessed. The quality of reporting and methodological rigor of sample size calculation were compared pre and post 2012. In conclusions, the authors recommended specific areas for improvement, with an urgent need for further methodological development. Some minor comments below for consideration. The title doesn't reflect the main objective of this paper. With a series of systematic reviews published to date, how this review differs from those already carried out needs to be highlighted here. In the Results, the reported number 1,753 has no reference. Figure 2 may be better presented with the total 274 assessed for eligibility but excluded, and the numbers of full reports (32) and trial protocols (28) for inclusion added to the box. In Table 2, Number of steps ranged from two, three or four, more than five, not reported. Also, median [IQR] was not indicated for Number of measurement points.

	In Table 3, the median numbers of items reported in 1987-2012 and 2013-14 were 4 [IQR 1-6] and 6 [IQR 5-6] respectively. The absolute difference was reported as -1.22 with 95% CI [-2.36, -0.07] (p-value=0.067). In Table 5, the number of studies allowing for time effects increased from 17% to 44% with a p-value of 0.053. This was reported differently in the last sentence of the Results section. Please check that all numbers are reported consistently. The absolute differences in Table 3-5 were reported for proportions and medians to 1-2 decimal places. These values, however, cannot be directly calculated from the reported proportions and medians which are mostly presented as integers.

VERSION 2 – AUTHOR RESPONSE

1.The title doesn't reflect the main objective of this paper. With a series of systematic reviews published to date, how this review differs from those already carried out needs to be highlighted here.
We thank the reviewer for this suggestion. We have changed the title to:
“A systematic review finds major deficiencies in sample size methodology and reporting for stepped wedge cluster randomised trials”
2. In the Results, the reported number 1,753 has no reference. Figure 2 may be better presented with the total 274 assessed for eligibility but excluded, and the numbers of full reports (32) and trial protocols (28) for inclusion added to the box.

Thank you for spotting this over-sight. We have changed this sentence to read:
“The searches identified 3,248 studies of which 1,218 were immediately identified as duplicates and 1,696 were excluded on the initial abstract screen, leaving 334 full text articles which were assessed for eligibility.”

We have also as per the suggestion added the number of full reports (32) and protocols (28) to Figure 2, along with a total of the number of exclusions after full assessment (274).

3. In Table 2, Number of steps ranged from two, three or four, more than five, not reported. Also, median [IQR] was not indicated for Number of measurement points.

Table 2 has been amended so that the median and IQR are indicated for the variable the number of measurement points. We have also corrected the labelling of the item the number of steps, to two, three of four, more than four, not reported.

4. In Table 3, the median numbers of items reported in 1987-2012 and 2013-14 were 4 [IQR 1-6] and 6 [IQR 5-6] respectively. The absolute difference was reported as -1.22 with 95% CI [-2.36, -0.07] (p-value=0.067).
We have corrected this error. The correct absolute difference is 1.22 with 95% CI (0.07, 2.36).

5. In Table 5, the number of studies allowing for time effects increased from 17% to 44% with a p-value of 0.053. This was reported differently in the last sentence of the Results section. Please check that all numbers are reported consistently.

We have corrected this mistake. The correct number was reported in the table and this has been made consistent in the text. We have also carefully screened the paper for correctness. The sentence in the paper now reads:

“There was an increase over time in the percentage of studies allowing for time effects from 17% pre 2012 to 44% post 2013 (P-value 0.063).”

6. The absolute differences in Table 3-5 were reported for proportions and medians to 1-2 decimal places. These values, however, cannot be directly calculated from the reported proportions and medians which are mostly presented as integers.

We have amended Tables 3 to 5 so that all percentages are reported to 1 decimal place.

Additional changes
We have in-addition made a few small changes to the paper, to improve clarity and to amend a couple of typing errors we have spotted. We have noted these by submitting a tracked changes version of the paper, as well as a clean version.