The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting

K Hemming,[1] T P Haines,[2] P J Chilton,[1] A J, Girling,[1] R J Lilford[3]

[1]University of Birmingham, Birmingham B15 2TT, UK

[2]Monash University, Victoria 3800, Australia

[3]University of Warwick, Coventry CV4 7AL, UK

Correspondence to: K Hemming k.hemming@bham.ac.uk

Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/BMJ.h391)

Cite this as: *BMJ* 2015;350:h391
doi: 10.1136/bmj.h391

Accepted: 24 November 2014

The stepped wedge cluster randomised controlled trial is a relatively new study design that is increasing in popularity. It is an alternative to parallel cluster trial designs, which are commonly used for the evaluation of service delivery or policy interventions delivered at the level of the cluster. The design includes an initial period in which no clusters are exposed to the intervention. Subsequently, at regular intervals (the "steps") one cluster (or a group of clusters) is randomised to cross from the control to the intervention under evaluation. This process continues until all clusters have crossed over to be exposed to the intervention. At the end of the study there will be a period when all clusters are exposed. Data collection continues throughout the study, so that each cluster contributes observations under both control and intervention observation periods. It is a pragmatic study design, giving great potential for robust scientific evaluations that might otherwise not be possible.

SUMMARY POINTS

The stepped wedge cluster randomised trial is a novel research study design that is increasingly being used in the evaluation of service delivery type interventions. The design involves random and sequential crossover of clusters from control to intervention until all clusters are exposed.

It is a pragmatic study design which can reconcile the need for robust evaluations with political or logistical constraints. While not exclusively for the evaluation of service delivery interventions, it is particularly suited to evaluations that do not rely on individual patient recruitment. As in all cluster trials, stepped wedge trials with individual recruitment and without concealment of allocation (or blinding of the intervention) are at risk of selection biases.

In a stepped wedge design more clusters are exposed to the intervention towards the end of the study than in its early stages. This implies that the effect of the intervention might be confounded with any underlying temporal trend. A result that initially might seem suggestive of an effect of the intervention may therefore transpire to be the result of a positive underlying temporal trend. Sample size calculations and analysis must make allowance for both the clustered nature of the design and the confounding effect of time.

The stepped wedge cluster randomised trial is an alternative to traditional parallel cluster studies, in which the intervention is delivered in only half the clusters with the remainder functioning as controls. When the clusters are relatively homogeneous (that is, the intra-cluster correlation is small), parallel studies tend to deliver better statistical performance than a stepped wedge trial. However, if substantial cluster-level effects are present (that is, larger intra-cluster correlations) or the clusters are large, the stepped wedge design will be more powerful than a parallel design, even one in which the intervention is preceded by a period of baseline control observations.

Brief history of the stepped wedge cluster randomised trial

The stepped wedge cluster randomised trial has been used across several settings for some years, but early stepped wedge designs were sometimes described in other terms such as "waiting list designs" or "phased implementations." The Gambia hepatitis intervention study (example 1) is probably the earliest and most widely known stepped wedge study.[1]

Two systematic reviews, determining the number and breadth of stepped wedge studies, have recently been conducted.[2 3] These reviews reveal that the use of this study design is on the increase and that areas of use are diverse and include HIV, cancers, healthcare associated infections, social policy, and criminal justice.

In 2007 Hussey and Hughes[4] first described methods to determine statistical power available when using a stepped wedge design. However, there is a dearth of literature on the more general methodological aspects, such as the rationale for, and conduct of, stepped wedge studies. In this article we illustrate how this new study design differs from the conventional parallel design and its variations. We also give several examples and consider several design and methodological issues, including rationale, sample size, and efficiency compared with competing designs, and highlight some important reporting and analysis considerations.

Study rationale

The stepped wedge is a pragmatic study design that reconciles the constraints under which policy makers and service managers operate with the need for rigorous scientific evaluations. While researchers may believe an evaluation of an intervention is required, it is decision makers (that is, politicians and managers) who control resources for system change. In order to get the research done, the researcher must be alive to the concerns of other stakeholders.

First, it may be the case that a key stakeholder (such as hospital manager or government minister) thinks that there is already sufficient evidence of effectiveness, whereas the researcher might take a different view. For example, when the UK government announced a new flagship programme called Sure Start to provide support for preschool children in deprived neighbourhoods, there was already some evidence in favour of the intervention,[5–7] but value for money had not been proven to everyone's satisfaction.

Second, decision makers may perceive that their credibility will be threatened or that they may hand their opponents a public relations scoop if they set up a traditional experiment. The political payoff from the Sure Start programme might have been attenuated had participating communities been divided into intervention and control clusters with no scheduled intervention date. The Sure Start programme is not alone here, indeed "the history of public policy experiments is littered with evaluations torpedoed by politicians appropriately attentive to the short term desires of their constituents, such as those who wind up in the control group without new services

EXAMPLE 1: THE GAMBIA HEPATITIS INTERVENTION STUDY

The Gambia hepatitis intervention study used a stepped wedge cluster randomised design in the 1980s to investigate the effectiveness of a vaccine for hepatitis B in preventing liver disease.[1] The vaccine had established effectiveness at preventing hepatitis B, though at the time no randomised evidence existed to show that it protected against chronic liver disease. Conclusive evidence of effectiveness against liver disease would require very long term studies (in the region of 30 years). Given the preliminary evidence of efficacy against hepatitis B, the vaccination was going be rolled out in the national infant vaccination programme. However, in order to obtain evidence on long term benefit, a phased but random implementation of the hepatitis B vaccination was initiated. Under a sequential rollout, geographically defined areas of the country were randomly allocated to incorporate the vaccination into the existing childhood vaccination schedule. A new region was randomly allocated in steps of 10–12 week intervals, such that complete national coverage was obtained after about four years. Follow-up of the cohort for liver disease outcomes is ongoing.

or who cannot imagine why a government would randomly assign citizens to government programmes."[8]

Third, there may be logistical constraints: complex interventions can rarely be implemented en bloc but must be rolled out sequentially. The stepped wedge study then fulfils a dual role, serving as a scientific tool that incorporates a fair way to determine the order of rollout under logistic constraints.[9]

The above political, logistical, and ethical constraints tend to coexist. In such circumstances, the alternative to a stepped wedge design may not be a parallel cluster trial but a weaker, non-experimental design. Under such a scenario the stepped wedge design is "naturalistic" in that the implementation may proceed much as it would have done had the evaluation not been in place while allowing randomised evidence of effectiveness.

The stepped wedge trial has other advantages over parallel cluster designs of a technical nature, but it also has disadvantages. In the rest of this paper we discuss these statistical and technical features and identify situations where this design is more or less suitable than alternatives.

How the stepped wedge cluster randomised trial relates to other cluster designs

In the evaluation of interventions delivered at the level of a general practice, hospital ward, or hospital where it is not possible to randomise individuals, randomisation may be carried out at the level of the cluster.[10] There are broadly three types of cluster trials to choose from (illustrated in fig 1). In the conventional (parallel) cluster randomised trial, clusters are randomised to either the intervention or control arm at the start of the trial and remain in that arm for the duration of the study (figure 1a).

This design may be elaborated into a cluster randomised trial with a baseline period (fig 1b).[11] Under this design observations are made during a baseline period (before any cluster is randomised to receive the intervention) and again in a post-intervention period (where clusters randomised to the intervention have switched to receiving the intervention). This design (with the addition that control clusters received the intervention at the end of the study) was used for the evaluation of the Mexican Universal Health Insurance Programme, described in example 2.

In a stepped wedge study, the design is extended so that every cluster provides before and after observations and every cluster switches from control to become exposed to the intervention, but not at the same point in time (fig 1c). The stepped wedge study takes its name from the stepped wedge shape that is apparent in the schematic illustrations.

As with other types of cluster design, the outcome data in a stepped wedge trial can derive either from single measurements taken from individual participants, but different participants at each step in the study (cross-sectional data); from repeat measurements on the same cohort of individuals recruited at the start and followed up throughout the study (longitudinal data); or from mixtures of the two (probably best described as an open cohort design). The depression management trial (example 3) is an example of an open cohort study design, with some patients

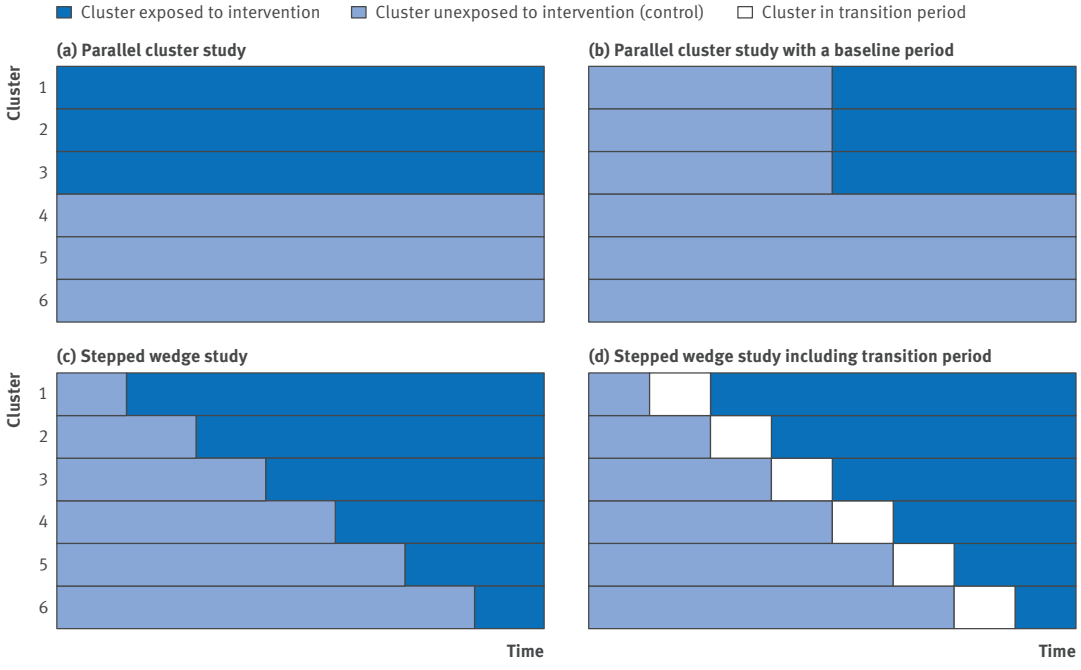

Fig 1 | Schematic illustration of the conventional parallel cluster study (with variations) and the stepped wedge study

remaining in the study for the duration, and others joining the study when they become a resident of a care home. The EPOCH trial (example 4), based on hospital emergencies, is a good example of a cross-sectional study.

Design considerations

When designing a stepped wedge cluster randomised trial, the number of clusters, number and length of steps, and number of clusters randomised at each step need to be determined. These are generally influenced by logistical considerations. For example, the availability of suitable clusters may limit the number of clusters included.

EXAMPLE 2: EVALUATION OF THE MEXICAN UNIVERSAL HEALTH INSURANCE PROGRAMME

A major change in the method used to finance healthcare in Mexico was evaluated by a phased and random implementation.[8] A randomised evaluation on such a large scale represents a major achievement in the robust evaluation of a public policy. The Harvard research team was tasked with an evaluation at the request of the Mexican Ministry of Health (in the expectation that if the intervention was successful it would survive any change of government). Seventy four clusters were matched in pairs so that one received the intervention and the other acted as control (as illustrated in fig 1b). In this particular case, an undertaking was made to make the intervention available to control clusters on completion of the study.

EXAMPLE 3: MULTI-STRUCTURED DEPRESSION MANAGEMENT IN NURSING HOMES

A stepped wedge cluster randomised trial, using an open cohort design, was used across 17 nursing homes in the Netherlands, with homes randomly assigned to one of five dates to introduce an intervention to promote the diagnosis and management of depression.[12] The trial ran from May 2009 to April 2011, with three or four homes randomised approximately every four months. Individual informed consent was elicited from residents, though residents and staff were blinded to the allocation. Most residents were recruited at the start of the trial and followed up over the five steps; others were recruited during the trial and followed up for any remaining steps. Characteristics of the clusters and individuals were summarised by group of randomisation. Adherence to the intervention was high (82%). The primary outcome was prevalence of depression, and this was analysed using a linear mixed model, adjusted for calendar time and with a random effect for nursing home and allowing for repeated measures on the same individuals (using a random effect). The adjusted (for calendar time) effect size was −7.3% (95% confidence interval −13.7 to −0.9), which means an estimated 7% reduction in prevalence of depression after the introduction of the intervention.

EXAMPLE 4: THE EPOCH TRIAL

The EPOCH trial is a cross-sectional stepped wedge cluster randomised trial of a service delivery intervention to improve the care of patients undergoing emergency laparotomy.[14] The intervention package is a complex intervention, including quality improvement and an integrated care pathway. The intervention will be rolled out sequentially to 90 hospitals, with six clusters of 15 geographically close hospitals (that is, clustering within clustering) switching from control to intervention every 5 weeks at 15 different time points (fig 2). The design incorporates a 5 week transition period in each cluster. The primary outcome is 90 day mortality, and no individual patient recruitment is needed. It is expected that approximately 18 patients will meet the inclusion criteria (> 40 years old and undergoing emergency laparotomy) per hospital per 5 week epoch, equating to a total sample size of about 27 500 patients. The trial is powered to detect a change in 90 day mortality from 25% to 22% at 90% power and 5% significance. Implementation of this power calculation used the Stata function "steppedwedge"[15] and followed the Hussey and Hughes approach.[4]

When the motivation for using the stepped wedge design is that it may be impossible to intervene simultaneously in all clusters (as in example 3), the study duration is dictated by the system's capacity to implement the service change. The number of observations per cluster is often determined by the number of participants meeting eligibility criteria. Design features that are not fixed by logistical constraints can be chosen with the aim of maximising statistical efficiency (a point to which we return). The chosen design can then be illustrated schematically, as in example 4, the EPOCH trial (fig 2).

The clusters switching at each step are usually independent of one another; however, they might be related in some way (such as different wards within the same hospital), in which case a multilevel element is introduced into the stepped wedge design (as was used in the EPOCH study, example 4).[13] Stepped wedge studies are usually designed so that approximately equal numbers of clusters switch at each step. Some designs make special allowance for the length of time it takes to embed the intervention into a cluster (example 4). During such transition periods the cluster cannot be considered as either exposed or not exposed (fig 1d).

A variation on the stepped wedge design has also been described to evaluate disinvestment rather than investment decisions. Here, instead of rolling out a new intervention, an existing intervention that was routinely provided is removed sequentially (thus reversing the roles of control and intervention periods).[16] Other variations include a group of clusters that remain exposed or not exposed to the intervention throughout the study period.[17]

Once the layout of the design has been determined, the individual clusters (or groups of clusters) should be randomised to their positions within the design. In the EPOCH trial, hospitals were divided into geographical groups (to facilitate simultaneous rollout within groups), and these groups were randomised to the implementation start date (that is, the "rows" of the design in fig 2). In any stepped wedge design the steps of the wedge divide the study duration naturally into separate observation periods.

Sample size and power calculations

Formal methods for sample size and power calculations have been described only for cross-sectional stepped wedge designs.[4 13] For this reason, a cross-sectional study is assumed throughout the following paragraphs. Similar considerations apply to cohort designs, but a reliable sample size algorithm for these designs has not yet been established.

All cluster trials should allow for correlations between individuals in the same cluster.[18] A consequence is that a parallel cluster trial will require a larger sample size than a corresponding individually randomised trial. The standard approach here assumes that any two observations from the same cluster will have a constant correlation between them, a quantity known as the intra-cluster correlation. Then the increase in sample size for the simple parallel cluster trial is determined by a simple multiplicative factor (the "design effect"), which depends both on the magnitude of the intra-cluster correlation and on the number of subjects in each cluster. For the parallel cluster trial with a baseline period, a variation on this design effect is available.[11]

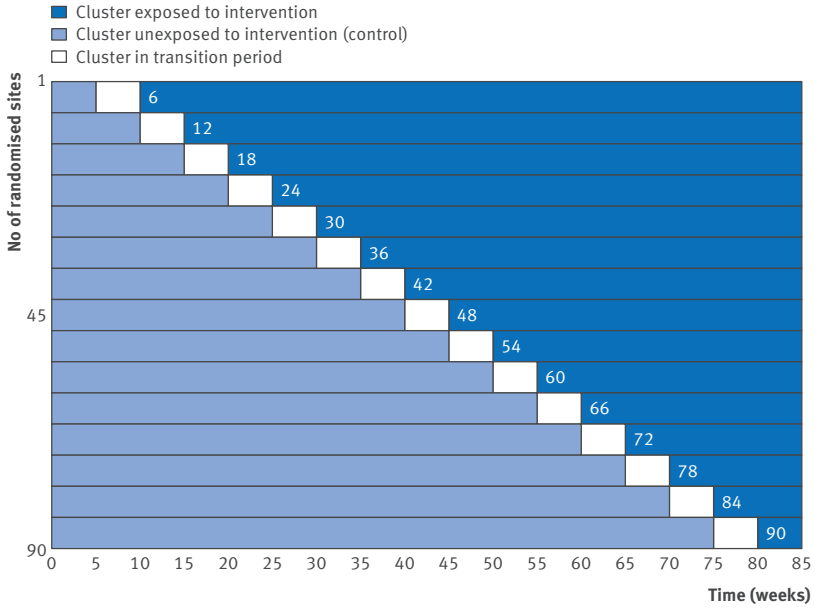

Fig 2 | Schematic representation of the EPOCH stepped wedge study (example 4). The trial will be conducted over 85 weeks, with six hospitals crossing from control to intervention approximately every 5 weeks until week 85, when all 90 hospitals will be exposed to the intervention

In a stepped wedge study, the sample size calculation is complicated by the need to allow for the confounding effect of calendar time (an issue discussed later in more detail), and this means that the standard design effect is no longer applicable. Compared with a simple parallel study, where no such confounding occurs, the time effect tends to degrade the precision of the study and increase the sample size needed to achieve adequate power. On the other hand, each cluster in a stepped wedge study contributes both exposed and unexposed observations, and so, to some extent, acts as its own control. This feature tends to enhance the precision of the study compared with a simple parallel study if substantial cluster effects are present (that is, if the intra-cluster correlation is large). Indeed, a similar enhancement can occur even for small intra-cluster correlations if the individual clusters are large.[19] In summary, when the intra-cluster correlation is small, a simple parallel design (as in fig 1a) tends to deliver more statistical power (per measurement taken) than a stepped wedge design (fig 1c), but for larger intra-cluster correlations a stepped wedge study will tend to be more powerful.

Comparison with a parallel trial with a baseline period (fig 1b) is also revealing. This design entails a group of clusters in which both exposed and unexposed observations are taken alongside a dedicated control group of clusters. To this extent, it combines features of both the stepped wedge and the parallel design. Certainly its performance improves on that of the simple parallel study if the intra-cluster correlation is large. However, it delivers less power than a stepped wedge design (with four or more steps) whatever the value of the intra-cluster correlation.[19]

These properties are illustrated in Table 1, which shows how power depends on the intra-cluster correlation in a cross-sectional study. The example uses the method of Hussey and Hughes[4 13] and refers to a trial with 20 clusters and total cluster size of 50 designed to detect a standardised effect size of 0.3 (at 5% significance). The simple parallel cluster trial achieves the highest level of power when the intra-cluster correlation is small (0.01). However, when the intra-cluster correlation is large (0.1) the power available under the simple parallel design drops to just 50%, while the power available under both the parallel design with a baseline period and the stepped wedge design (with four steps) retains a value close to 80%. In this example, the power under the stepped wedge cluster trial is only slightly larger than that under the parallel cluster trial with a baseline period.

In practice, power calculations and comparative efficiency for stepped wedge trials depend not only on the intra-cluster correlation but also the number of clusters in the study, the number of observations in each cluster, and the detailed structure of the design (as depicted in fig 1, panels c and d). Methods that determine statistical power for a stepped wedge trial of fixed size[4 20 21] have been implemented in the statistical software package Stata.[15] The methods assume a constant sampling rate—that is, equal numbers per observation period in each cluster. However, the possibility of transition periods with no observations at all (as in the EPOCH trial, example 4) is incorporated into the Stata function. As yet, there is no specific adaptation of design effects, for calculating the power or sample size in a cohort stepped wedge trial, nor implementation in a statistical package for this design.

Conduct

The conduct of the stepped wedge cluster randomised trial bears much in common with the main alternatives—the parallel cluster trial and the parallel cluster randomised trial with a baseline period. Since all these designs are used to study similar policy and service delivery interventions, they raise many of the same issues, particularly those relating to selection and concealment.

Recruitment of individual participants is not typically necessary when policy or service delivery interventions are studied, and where cross-sectional designs based on anonymous data, such as death and morbidity rates, are used (as in example 4). Particular care is needed, however, when individuals are recruited within each cluster to take part in the study. Here, as in the case of parallel designs, steps should be taken to mitigate the risk that participants will vary systematically across exposed and unexposed observation periods. In particular, participants should be recruited before allocation (to unexposed or exposed period) is known or recruited completely blind to the exposure status (as in the depression screening trial, example 3).[22 23] In stepped wedge studies with recruitment extended over time or without blinding to the intervention this may not be possible, creating a risk of selection bias, even though clusters are allocated randomly over time.

The stepped wedge cluster randomised trial, while pragmatic, also requires cooperation and commitment from the clusters. Clusters will have to be ready to cross to the intervention as and when the randomisation order dictates. Most stepped wedge trials to date have given ample notice of the crossover date to clusters. However, this does not rule out the possibility that some clusters will not be able to initiate the intervention as and when the randomisation schedule dictates.

Analysis

In a standard parallel trial the intervention is allocated to some clusters and not to others and analysis compares

Table 1 | Comparison of power* available under different cluster randomised trial designs by intra-cluster correlation: a simple parallel design, a parallel design with baseline period, and a stepped wedge design (all cross-sectional designs)

	Intra-cluster correlation 0.01			Intra-cluster correlation 0.1		
	Simple parallel trial	Parallel trial with baseline period	Stepped wedge trial	Simple parallel trial	Parallel trial with baseline period	Stepped wedge trial
Number of clusters	20	20	20	20	20	20
Cluster size	50	50	50	50	50	50
Total sample size	1000	1000	1000	1000	1000	1000
Number of steps	0	1	4	0	1	4
Number of clusters per step		10	5		10	5
Power*	0.97	0.87	0.88	0.50	0.77	0.82

*Power to detect a moderate effect size of 0.3 (SD 1) at 5% significance (power under an individual randomisation is 0.9973).

intervention arms. In a stepped wedge study, exposed (intervention) and unexposed (control) observation periods take the place of "arms" in parallel cluster trials. Thus, the distribution of results across unexposed observation periods is compared with that across the exposed observation periods. As with any randomised comparison, characteristics of the individuals and clusters should be summarised by exposure status so as to allow consideration of selection biases and lack of balance. Where there are a small number of steps, these characteristics can be compared by randomisation group (as in example 3). This should include the numbers analysed, the average cluster size, cluster characteristics, and important patient characteristics. The actual design, showing numbers of observations per cluster, can be schematically presented, as shown in fig 3.

Following an intention to treat principle, clusters should be analysed according to their randomised crossover time irrespective of whether crossover was achieved at the desired time. Some studies have included both an analysis by the intention to treat schedule and by that which actually occurred.[24]

Under the stepped wedge design the evaluation happens over a period of time, during which the proportion of clusters exposed to the intervention gradually increases. This means that unexposed observations will, on average, be from an earlier calendar time than exposed observations. Additionally, in evaluations of policy changes and service delivery interventions, other external changes may occur in the way care is delivered, which may have an impact on the outcome under evaluation. Thus, calendar time is associated with both the exposure to the intervention and also possibly the outcome, and so is a potential confounder and should be adjusted for in the analysis.

In the situation of an underlying temporal trend, an intervention that at first seems to be effective might no longer be when adjustment for calendar time has been made. There are several possible explanations for this. It might be that, external to the study, there was a general move towards improving patient outcomes, perhaps the very initiative which prompted study investigators to instigate the intervention in question. This phenomenon has been described as "a rising tide."[25] On the other hand, an intervention may be effective, yet there still may be a real underlying temporal trend, although this may be attributable to contamination. In the Matching Michigan study (example 5) these explanations were explored as possible reasons for the finding of no effect of an intervention that in other settings had been very positive.

Adjusting for the systematically different observation periods and for clustering in the data is accomplished by fitting an appropriate generalised linear mixed model or using generalised estimating equations. Hussey and Hughes specify models in which time is included as a fixed effect for each step.[4] So, for continuous (and normally distributed) outcomes, this would mean a linear model with random effect for cluster and fixed effect for each step; and, for binary outcomes, a logistic regression model with random effect for cluster and fixed effect for each step.

In a cohort design, some acknowledgment for the dependence between individual measurements over the course of the study will be needed. The simplest option is perhaps to introduce an additional random effect for individuals in the study (as in example 3).

The estimated intra-cluster correlation and time effect from the fitted model, although not of direct importance in the interpretation of the effect of the intervention, should be reported both for use in the design of future trials and to allow appreciation of any underlying confounding effects of calendar time. Other options for analysis include using within cluster comparisons only (although this does not adjust for any confounding effect of time), and treating the study as a series of (unbalanced) parallel cluster trials.[26]

A stepped wedge design also allows investigators to examine the way in which the impact of the intervention develops (over time) once it is introduced into a cluster. This might be important where an intervention needs an initial period of adjustment before becoming fully embedded in the setting. In such cases the length of the period (up to the current observation) during which the cluster has been exposed to the intervention can be included in the model as an effect modifier.

Finally, a stepped wedge design also allows exploration of heterogeneity in treatment effects between clusters, using within cluster comparisons of exposed and unexposed periods. Although the design may not be powered for these analyses, they can inform an interesting secondary investigation.

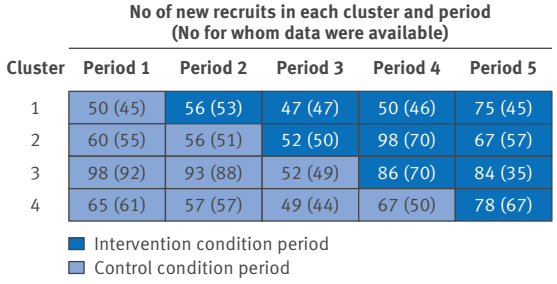

Fig 3 | Example of study size presentation that could be applied to a cross-sectional stepped wedge cluster randomised trial

EXAMPLE 5: THE MATCHING MICHIGAN STUDY

The Matching Michigan study is a non-randomised stepped wedge trial evaluating a complex intervention to reduce central venous catheter bloodstream infections in intensive care units.[25] The study included 215 intensive care units (out of a possible 223) in the UK and obtained complete outcome data for 147 (66%) of units. Four groups of units initiated the intervention at four separate points in time between April 2009 and March 2011. The intervention was based on a similar intervention that had been hailed as reducing intensive care infections by 80% in a non-randomised, before and after design in Michigan.

The Matching Michigan study identified secular trends and no evidence of any intervention effect, even though at first the intervention looked to be a success. Reasons for the secular trend are probably multifaceted and include, but are not limited to, a rising tide of activities directed at improving patient safety and the contamination of the intervention in clusters waiting to be crossed over.

Reporting of stepped wedge cluster randomised trials

Reporting guidelines specific to stepped wedge cluster randomised trials do not exist, and so we recommend that reporting should follow the Consort 2010 extension to cluster randomised trials.[27] Here we recommend some minor additions or modifications that can be used until specific guidelines for a stepped wedge design are formalised (Table 2).

Recommendations

The stepped wedge is a novel cluster randomised controlled trial design, emerging in the field of service delivery as well as policy evaluations. This design can be considered as an extension of the a parallel cluster trial with a baseline period and other variations of the conventional parallel cluster trial, including non-randomised designs, for the evaluation of service delivery interventions or other forms of interventions delivered at the level of the cluster.

We recommend the stepped wedge cluster trial as a potentially efficient and pragmatic randomised study design (although efficiency depends on the intra-cluster correlation and cluster size) for the evaluation of service delivery interventions where outcomes are based on routinely collected data (and so eliminating the need for individual participant recruitment). When outcomes are not based on routinely collected data or when individual recruitment is required, as in all cluster trials, special consideration should be given to minimising selection biases.

When planning a stepped wedge cluster randomised trial, consideration needs to be given not only to the sample size (which will depend on the intra-cluster correlation and number of steps) but also to the method of analysis, the possibility of repeated measures on individuals (that is, clarity of cohort, open cohort, or cross-sectional design), and reporting of adjusted (for time) treatment effects as the primary analysis. In addition to the conventional flow diagram, we recommend that a design diagram should illustrate how many participants are within each cell of the design.

Policy implications

Evaluation of drug therapies has long been deemed essential in accordance with evidence based medicine. The evaluation of non-pharmaceuticals, such as policy changes or service delivery methods, has unfortunately been less rigorously evaluated. It has been argued that service redesign should be evaluated by rigorous quasi-experimental design.[28] Quasi-experimental designs typically include (non-randomised) before and after studies, which are known to be confounded by temporal trends (which can't be adjusted for), and controlled before and after studies (which are subject to other confounding biases). However, given that most policies are rolled out over a period of time, the stepped wedge cluster randomised trial offers a fair (as the order of the rollout is determined at random) and randomised evaluation. Policy makers ought to take advantage of this pragmatic study design to evaluate effectiveness of policy changes.

Further research

There are many potential variations to the simple stepped wedge cluster randomised trial that are yet to be investigated, and which we have not considered here. These include design and analysis of the cohort stepped wedge trial (which has to contend with change over time at the site or cluster level as well as within participants), clusters within clusters in a trial (that is, wards within hospitals), trials with more than two arms, restricted randomisation such as pairing, the effect of varying cluster sizes and varying step sizes, and the hybrid design (which is a mixture of the conventional parallel design and the stepped wedge design).[17]

In cases where random allocation would not be possible outside the stepped wedge design, then the stepped

Table 2 | Suggested modifications to the Consort 2010 cluster extension for reporting of stepped wedge cluster randomised trials

Section and topic	Modified checklist item
Title and abstract	
Title	Identification as a stepped wedge cluster randomised trial
Introduction	
Rationale	Stakeholders not amenable to parallel randomisation
	Need for sequential rollout
	Desire for all clusters to receive the intervention
	Evidence of preliminary effectiveness
	Likely to be an efficient design for anticipated intra-cluster correlation and cluster size
Methods	
Trial design	Definition of the cluster
	Cluster size distinguished from cluster size per observation period
	Length of the steps (observation periods)
	Number of clusters randomised at each step
	Cohort (repeated measures on individuals), cross-sectional design (different individuals), or mixture (open cohort)
	Schematic representation of the trial design
Sample size justification	Allowance for clustering
	Allowance for the number of steps
	Allowance for any repeated measures on individuals
	Clear reference to the methods used
Analysis	Allowance for clustering (that is, random effect model)
	Allowance for the number of steps (that is, fixed effect for step)
	Allowance for repeated measures on individuals, if appropriate
Results	
	Characteristics of sample reported by exposed and unexposed observation periods, or by randomisation group
	Adjusted (for time) treatment effect and 95% CI should be interpreted as unbiased estimate of effect size
	Schematic representation of actual study design
	Intention to treat analysis should follow the randomised design and might be different to that which actually transpired

wedge cluster randomised trial should be preferable to a non-random study design. Further work is required to establish empirical evidence on systematic or random bias associated with stepped wedge studies, especially in those with individual patient recruitment. Finally, quality and reporting of all study designs is generally low when the designs are first introduced and so consensus guidelines on reporting and analysis are urgently needed.[29]

We thank Jon Deeks, Richard Riley, and Oyinlola Oyebode for their comments on early drafts.

Contributors: KH conceived the idea for the paper, wrote the first draft, and led the writing of the paper. KH is the guarantor. AG read and commented on drafts and provided critical insight into the efficiency section. RL read and commented on draft and provided critical insight into the rationale section. PC and TH read and critically commented on drafts. All authors have approved the final version.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: RJL and PJC had financial support for the submitted work from the National Institute for Health Research (NIHR) Collaborations for Leadership in Applied Health Research and Care (CLAHRC) for West Midlands; KH, RJL, and AJG had financial support from the Medical Research Council (MRC) Midland Hub for Trials Methodology Research (grant No G0800808); TPH is supported by a Career Development Fellowship from the Australian National Health and Medical Research Council (1069758). No financial relationships with any organisations that might have an interest in the submitted work in the previous three years. No other relationships or activities that could appear to have influenced the submitted work.

RECOMMENDATIONS

The stepped wedge cluster randomised trial offers a randomised method of evaluation of an intervention delivered at the level of the cluster. In cases where randomisation to either control or intervention arm is precluded but randomisation to a date of initiation is possible, it offers a means of a randomised evaluation in place of a less robust design. However, the design requires the fitting of more complex models than parallel designs and must adjust for underlying temporal trends.

A stepped wedge cluster randomised trial is likely to be the preferable study design when all or some of the below hold:

- There is evidence already in support of the intervention (for example, known to be effective at individual level but uncertain about policy level), or there is resistance to a parallel design in which only half of the clusters receive the intervention.
- The intervention is a service delivery or policy change that can be implemented without the need for individual participant consent. The outcome, or at least some important outcomes, may be available from routinely collected data, so that individual participation for outcome data collection is not required (that is, no patient questionnaires).
- The intra-cluster correlation is anticipated to be high or cluster sizes large so that a cross-sectional stepped wedge design is likely to be more efficient than the simple parallel cluster design.

Reasons to be cautious about using a stepped wedge cluster randomised trial:

- When the intra-cluster correlation is low (or the cluster size small) the stepped wedge cross-sectional study can be an inefficient design compared with a simple parallel cluster design.
- The study has a cohort or open cohort design, for which there are currently no methods developed to determine power available.
- When the outcome requires individual participant data collection (without blinding), lack of concealment of allocation is likely to mean a risk of differential selection of participants between arms.
- It is unlikely that clusters will be able to follow the randomisation schedule.

1 The Gambia Hepatitis Study Group. The Gambia hepatitis intervention study. *Cancer Res* 1987;47:5782–7.
2 Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006;6:54.
3 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;64:936–48.
4 Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;28:182–91.
5 Camps L, Long T. Origins, purpose and future of Sure Start children's centres. *Nurs Child Young People* 2012;24:26–30.
6 Glass N. Sure Start: the development of an early intervention programme for young children in the United Kingdom. *Child Soc* 1999;13:257–64.
7 Smith MS, Bissell JS. Report analysis: the impact of Head Start. *Harv Educ Rev* 1970;40:51–104.
8 King G, Gakidou E, Ravishankar N, Moore RT, Lakin J, Vargas M, et al. A "politically robust" experimental design for public policy evaluation, with application to the Mexican universal health insurance program. *J Policy Anal Manage* 2007;26:479–506.
9 Urlings-Strop LC, Stijnen T, Themmen AP, Splinter TA. Selection of medical students: a controlled experiment. *Med Educ* 2009;43:175–83.
10 Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Med Res Methodol* 2011;11:102–11.
11 Teerenstra S, Eldridge S, Graff M, de HE, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;31:2169–78.
12 Leontjevas R, Gerritsen DL, Smalbrugge M, Teerenstra S, Vernooij-Dassen MJ, Koopmans RT. A structural multidisciplinary approach to depression management in nursing-home residents: a multicentre, stepped-wedge cluster-randomised trial. *Lancet* 2013;381:2255–64.
13 Hemming K, Lilford RJ, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med* 2015;34:181–96.
14 Pearse R, Pedan C, Bion J, Faiz O, Holt P, Girling A, et al. HS&DR - 12/5005/10. Enhanced Peri-Operative Care for High-risk patients (EPOCH) trial: a stepped wedge cluster randomised trial of a quality improvement intervention for patients undergoing emergency laparotomy [protocol]. 2013. www.nets.nihr.ac.uk/projects/hsdr/12500510
15 Hemming K, Girling A. A menu driven facility for sample size for power and detectable difference calculations in stepped wedge randomised trials. *Stata J* 2014;14:363–80.
16 Haines T, O'Brien L, McDermott F, Markham D, Mitchell D, Watterson D, et al. A novel research design can aid disinvestment from existing health technologies with uncertain effectiveness, cost-effectiveness, and/or safety. *J Clin Epidemiol* 2014;67:144–51.
17 Hughes J, Goldenberg RL, Wilfert CM, Valentine M, Mwinga KG, Guay LA, et al. Design of the HIV Prevention Trials Network (HPTN) Protocol 054: A cluster randomized crossover trial to evaluate combined access to Nevirapine in developing countries. 2003. *UW Biostatistics Working Paper Series* Working Paper 195. http://biostats.bepress.com/uwbiostat/paper195
18 Campbell MJ, Walters SJ. *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley, 2014.
19 Girling A, Hemming K. Statistical efficiency and optimal designs for stepped cluster studies. *Stat Med* [submitted].
20 Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013;66:752–8.
21 Hemming K, Girling A. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. *J Clin Epidemiol* 2013;66:1427–8.
22 Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ* 2009;339:b4006.
23 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;327:785–9.
24 Craine N, Whitaker R, Perrett S, Zou L, Hickman M, Lyons M. A stepped wedge cluster randomized control trial of dried blood spot testing to improve the uptake of hepatitis C antibody testing within UK prisons. *Eur J Public Health* 2014;cku096 [Epub ahead of print].
25 Bion J, Richardson A, Hibbert P, Beer J, Abrusci T, McCutcheon M, et al. 'Matching MichiganÉ: a 2-year stepped interventional programme to minimise central venous catheter-blood stream infections in intensive care units in England. *BMJ Qual Saf* 2013;22:110–23.
26 Hughes JP. Stepped wedge design. In: *Wiley Encyclopedia of Clinical Trials*. John Wiley, 2008:1–8.
27 Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012;345:e5661.
28 Moberly T. Service redesign should be tested as rigorously as new treatments, NHS chief says. *BMJ* 2014;348:g3744.
29 Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;4:21.