BMJ

RESEARCH

Internal and external validity of cluster randomised trials: systematic review of recent trials

Sandra Eldridge, professor of biostatistics,[1] Deborah Ashby, professor of medical statistics,[2] Catherine Bennett, statistician,[1] Melanie Wakelin, lecturer in medical statistics,[1] Gene Feder, professor of primary care research and development[1]

[1]Centre for Health Sciences, Barts and The London School of Medicine and Dentistry, London E1 2AT

[2]Wolfson Institute of Preventive Medicine, Barts and The London School of Medicine and Dentistry, London EC1M 6BQ

Correspondence to: S Eldridge
s.eldridge@qmul.ac.uk

doi:10.1136/bmj.39517.495764.25

ABSTRACT

Objectives To assess aspects of the internal validity of recently published cluster randomised trials and explore the reporting of information useful in assessing the external validity of these trials.

Design Review of 34 cluster randomised trials in primary care published in 2004 and 2005 in seven journals (*British Medical Journal*, *British Journal of General Practice*, *Family Practice*, *Preventive Medicine*, *Annals of Internal Medicine*, *Journal of General Internal Medicine*, *Pediatrics*).

Data sources National Library of Medicine (Medline) via PubMed.

Data extraction To assess aspects of internal validity we extracted data on appropriateness of sample size calculations and analyses, methods of identifying and recruiting individual participants, and blinding. To explore reporting of information useful in assessing external validity we extracted data on cluster eligibility, cluster inclusion and retention, cluster generalisability, and the feasibility and acceptability of the intervention to health providers in clusters.

Results 21 (62%) trials accounted for clustering in sample size calculations and 30 (88%) in the analysis; about a quarter were potentially biased because of procedures surrounding recruitment and identification of patients; individual participants were blind to allocation status in 19 (56%) and outcome assessors were blind in 15 (44%). In almost half the reports, information relating to generalisability of clusters was poorly reported, and in two fifths there was no information about the feasibility and acceptability of the intervention.

Conclusions Cluster randomised trials are essential for evaluating certain types of interventions. Issues affecting their internal validity, such as appropriate sample size calculations and analysis, have been widely disseminated and are now better addressed by researchers. Blinding of those identifying and recruiting patients to allocation status is recommended but is not always carried out. There may be fewer barriers to internal validity in trials in which individual participants are not recruited. External validity seems poorly addressed in many trials, yet is arguably as important as internal validity in judging quality as a basis for healthcare intervention.

INTRODUCTION

In cluster randomised trials, groups or clusters of individuals, rather than individuals themselves, are randomised. These trials are increasingly common in health services research, being particularly appropriate for evaluating interventions aimed at changing behaviour in patients or practitioners or changing organisation of services. Clusters might, for example, consist of patients in general practices or older people in nursing homes. Cluster randomised trials are pragmatic, measuring effectiveness rather than efficacy[1] and should therefore be both internally and externally valid.[2]

Internal validity

Internal validity refers to the extent to which differences identified between randomised groups are a result of the intervention being tested. It thus depends on good design, conduct, and analysis of the trial, with minimal bias.[3-5] In addition, without a sufficient sample size, differences that do exist between randomised groups that are a result of the intervention being tested might not be detected; sufficient sample size can also be considered a marker of internal validity.[5] For cluster randomised trials, statisticians have repeatedly emphasised the importance of accounting for the clustered nature of the data in sample size calculations and analysis[6-9] but investigators have not always heeded this guidance.[10-14]

A potential barrier to internal validity highlighted more recently is lack of blinding to allocation status of those identifying or recruiting individuals into a cluster randomised trial.[15 16] Concealment of allocation from those recruiting and randomising participants is well recognised as a cornerstone of internal validity for individually randomised trials.[17] In cluster randomised trials there are two levels of participant: the cluster and the individual. Identification or recruitment of individuals, or both, often takes place after randomisation (of clusters) and if those carrying out the identification or recruitment of patients at this post-randomisation stage are not blind to allocation status, bias can occur. Puffer and colleagues recommend that reports include a clear statement about when individual participants are

identified and whether or not those recruiting are blind to allocation status.[16]

Lack of other types of blinding is associated with poor internal validity in individually randomised trials[18] and might result in poor internal validity in cluster randomised trials. Lack of blinding in outcome assessment is usually considered the most serious potential source of bias[19]; in most cluster randomised trials it is possible to assess outcomes blind to allocation status. The nature of the intervention in most of these trials, however, means that it is rarely possible to blind those delivering components of the intervention to individual participants. For example, an intervention might involve educational outreach to all clinical staff in intervention general practices (clusters); these staff, who must then deliver enhanced care to patients, cannot be blind to whether or not they receive education.[w1] In addition, it is not always possible to blind individual participants to the fact that they are receiving an intervention—for example, if they are receiving leaflets[w2]—although this does not necessarily mean that they know their allocation status. This inability to blind health professionals (and sometimes individual participants) is a distinctive feature of these trials.

External validity

External validity refers to the extent to which study results can be applied to other individuals or settings. Several frameworks have been developed that are helpful in assessing this.[20-22] The RE-AIM framework (table 1) was developed by Glasgow and colleagues to characterise the public health impact of inter ventions.[22 23] The framework has been used to assess the external validity of evaluations of interventions common in cluster randomised trials,[23-25] although none of the previously published assessments specifically focuses on cluster randomised trials. Four features of RE-AIM are related to external validity: reach, adoption, implementation, and maintenance. We have focused on adoption and implementation because these factors can operate differently in individually and cluster randomised trials and are amenable to assessment from trial reports.

To judge adoption (the extent to which the settings included are representative of a wider population of settings and adequately described), a reader needs information on eligibility criteria for clusters, numbers of clusters randomised and analysed, and a discussion of generalisability of trial findings to clusters as well as individuals, all factors recommended in the extension to the CONSORT statement for cluster randomised trials.[26] Cluster recruitment rate also contributes to an assessment of adoption. The implementation of an intervention as intended requires the cooperation of the clusters in potentially two distinct ways. Firstly, health professionals in clusters must comply with any intervention targeted at them—for example, an educational programme. Secondly, they must deliver components of the intervention they are supposed to be actively involved in—for example, extra counselling sessions to patients. Using terms defined by Bonell and colleagues in a framework for assessing generalisability, we refer to compliance with programmes targeted at health professionals in clusters as acceptability, and delivery of intervention as intended as feasibility (table 1).[21]

Current study

We reviewed recent cluster randomised trial reports to assess the extent to which trial investigators have ensured internal validity through appropriate sample size calculations and analyses, blinding of those identifying and recruiting individual participants to allocation status, and blinding of patients and of outcome assessors. We explored the reporting of information useful in assessing external validity—namely, adoption through cluster eligibility, inclusion, retention, and generalisability, and implementation through the feasibility and acceptability of the intervention to health providers in clusters.

METHODS

Data

We included only trials in primary care to facilitate comparisons with results of trials selected from a previous review of cluster randomised trials.[13] We defined primary care using a hybrid of the definitions used in the United Kingdom[27] and the United States[28]; the rationale for this being that a definition that worked for these two different health services would also work in other countries. The definition is "accessible, often first contact, health care, usually provided within the community, which is either comprehensive, co-ordinated care involving sustained partnership with patients, or undifferentiated by age, gender, disease

Table 1 | The RE-AIM framework

Dimension	Assessed in this study
Reach—extent to which patients included in evaluation are representative of population of interest and adequately described	No
Efficacy—success rate of intervention if it is implemented as in guidelines	No
Adoption—extent to which settings included are representative of wider population of settings and are adequately described	Cluster eligibility, numbers approached, recruited, analysed
Implementation—extent to which intervention is implemented as intended in real world	Acceptability (adherence to any intervention components targeted directly at health professionals in clusters); feasibility (extent to which health professionals delivered intervention components to patients as intended)
Maintenance—extent to which programme is sustained over time	No

or organ. This includes comprehensive, co-ordinated care to particular subsets of the population sometimes for a fixed period, or care which focuses on sustaining health rather than treating illness."[13]

We included trials reporting primary evaluations of effectiveness where randomisation was by cluster (for example, general practices) as long as there were some outcomes collected from observational units at a level below the randomisation unit (for example, individual patients). We excluded reports that referenced main trial findings elsewhere or did not report outcomes or where individual participants were randomised. SE searched the National Library of Medicine (Medline) database electronically for primary care trials published (including e-publications) in 2004 and 2005 in seven current journals that our previous research identified as publishing six or more cluster randomised trials in primary care in an earlier period (1997-2000) (*British Medical Journal*, *British Journal of General Practice*, *Family Practice*, *Preventive Medicine*, *Annals of Internal Medicine*, *Journal of General Internal Medicine*, *Pediatrics*). SE identified cluster randomised trials by examining the abstracts and, when necessary, full texts. On the basis of previous trends, we estimated that we needed toidentify 40 trials, enough to provide sensible estimates of proportions of trials in certain categories. Two reviewers (SE and CB or MW) independently extracted appropriate information and resolved discrepancies by discussion or by referral to GF and DA.

Internal validity

To assess the extent to which investigators had followed recommendations about adequate power and appropriate analyses, we calculated proportions of reports correctly accounting for clustering in design and analysis. We compared these with similar proportions from trial reports in the same seven journals in 1997-2000 (unpublished data from our previous review). To assess the extent of blinding of those identifying or recruiting individual patients to allocation status, we grouped the trials into four categories:

Possibility of bias in recruitment/identification of participants—Bias was possible if those identifying or recruiting patients were not blind to allocation status and could have had an impact on who was identified or recruited or could have relayed information to patients to make them more or less likely to consent or if information given to patients at consent was clearly different in different intervention groups.

Bias unlikely in recruitment/identification of participants—Bias was unlikely if those identifying and recruiting patients were blind to allocation status or criteria for patient entry were such that recruiters could not have had a substantial impact on who was recruited, or both.

No possibility of bias in recruitment/identification of participants—If identification was blind to allocation status and there was no recruitment of individual participants bias could not exist. This can happen if, for example, general practices are recruited and outcomes from individual participants are assessed via routine data.[w3]

Unclear—Used if we could not put a trial into one of the above categories based on the trial report.

Many trials in our review would have started before publication of the key paper that highlighted inadequate blinding at recruitment of patients as a barrier to internal validity[16] and investigators might not have been fully aware of this issue at identification and recruitment of patients. We therefore also assessed whether or not investigators seemed to be aware of the issue at the time they published, as evidenced by appropriate discussion within their trial report. To assess other types of blinding we recorded whether the reports indicated that patients and those who assessed the primary outcome were blind to allocation status, not blind, or whether this was unclear. We defined the primary outcome as that specified by authors or, if not specified, the outcome used in the calculation of sample size or, if there was no sample size calculation, the first outcome presented in the abstract.

External validity

To assess adoption we extracted information reported on cluster eligibility and numbers approached, recruited, and lost to follow-up; when possible we calculated cluster recruitment and attrition rates. We compared results with those for trials from the same seven journals in 1997-2000. We also extracted any phrases investigators used to discuss cluster generalisability. To assess implementation we identified whether investigators reported the extent of adherence to any components of the intervention targeted directly at health professionals in clusters (acceptability) and the extent to which health professionals delivered any of these components to patients as intended (feasibility). In this sense, feasibility is not specific to cluster randomised trials but might be particularly important in these trials where interventions are often multifaceted and complex. We also identified whether investigators reported any lack of adherence to trial protocol as an issue in their trial. In addition, we assessed whether there was evidence of a substantial evaluation of trial processes to try to ascertain and understand acceptability and feasibility.

RESULTS

We identified 40 potential eligible trials and excluded six (in one clusters were not fully randomised, two referenced main trial findings elsewhere, two did not report outcomes, one was primarily a report of an individually randomised trial). We reviewed the 34 trials involving various cluster types and interventions (table 2).[w1-w34] Most disagreements on data extraction were resolved by discussion between data abstractors.

Internal validity

All reports contained information on analysis and 29 on sample size calculations. One report mentioned a sample size calculation reported elsewhere (we categorised this as not clear whether sample size calculation accounted for clustering). Sixty two per cent (21/34) definitely accounted for clustering in sample size

calculations and 88% (30/34) in analyses compared with 15% (9/60) and 73% (44/60), respectively, for trials in the same seven journals in 1997-2000 (unpublished data from previous review) (table 3).

Bias caused by lack of blinding of those identifying and recruiting individual participants was impossible or unlikely in 62% of trials (21/34) and possible in 21% (7/34) (table 3). In 14 trials individual participants (usually patients) were not recruited; we judged that selection bias was impossible in 12 and possible in one where general practitioners identified relevant patients after randomisation,[w2] and one trial report was not clear enough for us to make a judgment[w12] (table 4). Where individual participants were recruited (20 trials), we judged that bias was unlikely in nine, possible in six, and that we could not judge in five. Five reports commented on the possibility of bias in participant recruitment or identification; this was more likely if we had judged that there was a possibility of such bias in the trial (three out of seven trials).[w7 w22 w28] Individual

Table 2 | Clusters randomised and interventions used in trials

Trial	Type of cluster	Description of intervention
Aittasalo[w29]	Physicians	Training for two hours in procedure for prescription based physical activity counselling, users' guide for physicians
Byng[w8]	General practices	Quality improvement programme
Crump[w10]	Family compounds	Flocculant disinfectant or sodium hypochlorate given to family compounds to add to water supply
Dey[w13]	Health centres	Educational strategy to promote guidelines
Dietrich[w25]	Primary care practices	Care managers (supervised by psychiatrists) provided support to patients; clinicians and practice staff received education
Edwards[w17]	General practitioners	Training in shared decision making and use of simple aids for risk communication
Fairall[w1]	Primary care clinics	Senior nurses to deliver three to four educational outreach sessions to all clinical staff over three month period
Gattellari[w3]	Practices (general practitioners at same address)	Three telephone administered peer coaching sessions delivered by medical peer educators, enhanced by information packages including material for patients and general practitioners
Glasgow[w20]	Physicians	Computer assisted intervention aimed at patients before visit related to diabetes that produced outputs for patients, physicians, and "care manager" (nurse or medical assistant assigned by clinic and trained to use patient centred self management approach)
Griffiths[w27]	General practices	Asthma liaison nurses, education to practices, template to prompt review of patients, peak flow meters, and plans given to patients
Harmsen[w6]	General practitioners	Education of general practitioners (2.5 days of training) on intercultural communication; education of patients (video in waiting room) in communication in general
Herbert[w4]	Groups of physicians	Four arm trial, two interventions: individualised prescribing feedback; evidence based education
Hilberink[w30]	General practices	Four hour group training session on chronic obstructive pulmonary disease, smoking, and smoking cessation, support materials for professionals and patients; patient visit and education
Jellema[w26]	Practices	General practitioners explored presence of psychosocial prognostic factors, discussed these factors, set specific goals for reactivation, and provided educational booklet
Kendrick[w22]	Midwives	Education to midwives; structured discussions between midwife and mother to be; postcard and fridge magnet to mother
Kenealy[w18]	General practitioners	Four arm trial, two interventions: diabetes risk sheet filled in by patients and given to practitioner during consultation; computer icon flashed for patients eligible for screening
Kools[w5]	Child health centres	Breast feeding promotion programme
Laurant[w12]	Groups of general practices	Nurse practitioners to work collaboratively with clinicians following agreed guidelines
Margolis[w21]	Private paediatric and family practices	Education and process improvement methods through monthly meetings to support implementation of office systems for delivery of preventive care, tools provided to accelerate testing
Midlov[w31]	Practices	Educational outreach visits to general practices
Mitchell[w14]	General practices	Three arm trial, two interventions: anonymised feedback on practice performance compared with average for all practices; anonymised feedback plus feedback on individual patients
Mohan[w11]	Primary health centres	Doctors trained in counselling, communication, and clinical skills
Myers[w32]	Practices	Reminder to consider complete diagnostic evaluation (CDE) for appropriate patients; academic detailing visits to practice; feedback on CDEs; letter and telephone call to primary care practitioners
Ornstein[w15]	Primary care practices	Multi-method quality improvement intervention
Powell[w9]	Nutrition clinics	Community health aids received training and demonstrated play activities aimed at psychosocial stimulation to mothers at weekly home visits
Ruffin[w33]	Primary care practices	Four arm trial, two interventions: providing patients with their screening history and cues to future screening; providing staff with patient's screening history and current screening recommendations at every patient contact
Sandora[w24]	Child care centre	Supply of hand sanitiser to families in intervention childcare centres
Seligman[w19]	Physicians	Notification to physicians of patients' health literacy
Smith[w16]	General practices	Structured shared care service for diabetes implemented through educating general practitioners and nurses; introduction of specialist nurse; routine reviews in primary care; fast track referral system
Sondergaard[w2]	General practitioners	Two educational visits, feedback forms on baseline performance, guidelines, and patients' handouts
van Boeijan[w28]	Practices	All participants received individual treatment for 12 weeks based on cognitive behavioural therapy principles delivered in three different ways: guided self help; guidelines to general practitioners who then delivered simplified cognitive behaviour therapy (CBT); CBT from therapists
Watson[w23]	Families	Safety advice and safety equipment provided to families by health visitors
Welschen[w7]	Peer review groups of practitioners	Education for peer review groups and assistants; copy of guidelines; feedback on prescribing behaviour; educational material for patients
Witt[w34]	Practices	Academic detailing around guideline

participants were reported to be blind to allocation status in 56% (19/34) of trials. This was the case in all trials in which participants were not recruited except for two which randomised families or family compounds[w10 w23] and in seven out of 20 trials in which participants were recruited. In all of the latter seven trials, investigators reported making a specific effort to ensure that individual participants were not given information about allocation status (table 4). Primary outcome assessment was blind to allocation status in 44% (15/34) of trials (table 3); blinding was more likely if participants were recruited (10/20), but this effect could have arisen by chance (odds ratio 1.8, 95% confidence interval 0.4 to 7.3).

External validity

Most reports contained some information about cluster eligibility. We attempted to judge generalisability based on this and information about cluster inclusion and retention but found it difficult to do. Only 59% (20/34) of trial reports contained full information on numbers of clusters approached, recruited, and analysed (table 3); the comparable figure for trials from the same journals in 1997-2000 was 31% (19/60). We calculated cluster recruitment rates for 23 trials (median 50%, interquartile range 30-100%) and attrition rates for 27 (median 0%, 0-5%) (comparable figures for 1997-2000 trials were median 72%, 29-88%, and median 0%, 0-6%). Of the 18 trials with recruitment rates below 85%, only six reported a comparison of the characteristics of clusters approached and recruited (see table A on bmj.com). Two trials lost over a quarter of clusters after recruitment: one because of lack of eligible patients,[w29] the other because some clusters did not allow data collection to be completed.[w33]

Fifty three per cent (18/34) of trial reports contained a discussion of cluster generalisability (table 3). This was more likely if they also reported full information on numbers of clusters approached, recruited, and analysed (13/20 *v* 5/14), although again this effect could have arisen by chance (odds ratio 3.3, 0.8 to 13.9). Most suggested that generalisability might be restricted, but only four explained how clusters included might differ from those not included: more interested, motivated, familiar with training methods, ready to change.[w6 w17 w26 w33] None of these trials showed evidence of effectiveness of the intervention for the whole trial population and primary outcome.

Only two trials did not involve clusters in either an intervention targeted at them that they could opt out of or active involvement in intervention delivery; both assessed the effect of giving information to health professionals.[w14 w19] Fifteen trials reported information about levels of intervention implementation, and four discussed it (see table A on bmj.com). In most of these trials implementation was less than optimal. No reasons were given for health professionals in the clusters not fully adhering to the intervention targeted at them (lack of acceptability). The most common reason given for less than optimal delivery of the intervention (lack of feasibility) was lack of time. Eight reports mentioned additional specific research, usually qualitative, which explored trial processes, acceptability, or feasibility.

When we divided the trials according to whether they were published in the *BMJ* or elsewhere, the *BMJ* scored higher than other journals on eight of the 10 criteria in table 3. The difference in the proportions of trials in which the primary outcome was assessed blind to allocation status (81% in the *BMJ* and 26% in other journals, odds ratio 12.7, 2.1 to 76.6) was particularly striking. All other differences could have arisen by chance.

Table 3 | Proportions of trials (n=34 unless otherwise stated) following procedures to enhance internal and external validity. Figures are numbers (percentages)

Procedure	Followed procedure	Unclear if followed	Not followed
Internal validity			
Accounting for clustering in sample size calculation	21 (62)	8 (24)*	5 (15)
Accounting for clustering in analysis	30 (88)	2 (6)	2 (6)
Protected against recruitment/identification bias when identifying/recruiting patients	21 (62)†	6 (18)	7 (21)
Blinding of individual participants to allocation status	19 (56)	11 (33)	4 (12)
Assessment of primary outcome blind to allocation status	15 (44)	13 (38)	6 (18)
External validity			
Full information on number of clusters approached, recruited, and analysed	20 (59)	NA	14 (41)
Comparison of characteristics of clusters recruited and those not recruited	6 (33)‡	NA	12 (67)
Discussion of cluster generalisability	18 (53)	NA	16 (47)
Discussion of how the clusters analysed might differ from other clusters	4 (12)	NA	30 (88)
Some information about acceptability and/or feasibility	19 (59)§	NA	13 (41)

NA=not applicable (for external validity we assessed whether or not certain information was reported in trial report; by definition, it was never unclear whether information was reported).
*Five reports did not include sample size calculations; three did not provide adequate information in the sample size calculation.
†Includes those trials where we judged that selection bias was impossible or unlikely.
‡n=18 (judged only for those recruiting <99% of clusters).
§n=32 (judged only for trials in which clusters had option to opt out of intervention targeted at them or had to deliver part of intervention to patients).

DISCUSSION

Main findings

The time trends in our data suggest an encouraging improvement in the extent to which investigators account for clustering in the design and analysis of cluster randomised trials. About a quarter of the trials were potentially biased because of procedures for selecting patients. Blinding of individual participants to allocation status was almost universal in trials in which individual participants were not recruited, but much less common in trials when individual participants were recruited. In less than half of the trials assessment of the primary outcome was blind to allocation status. In two fifths of reports there was no information about the implementation of the intervention; where there was information, implementation was almost always less than optimal. The reporting of information relating to cluster generalisability might have improved since the late 1990s but remains poor in almost half of the trials we reviewed. We were not able to assess time

trends in procedures for selecting patients, blinding, or reporting of implementation because we had no data from earlier trials, but there seems to be considerable room for improvement. Because of small numbers of trials we are not able to make substantive conclusions about the differences in quality between journals, although our results suggest that trials reported in the *BMJ* might be of higher quality than trials in many other journals in respect of blinding those who assessed the primary outcomes.

Strengths and limitations

We focused on recent trials and had rigorous review procedures. We could not, however, judge the extent of some of the barriers to internal and external validity because of lack of reporting and might have underestimated the extent to which investigators recognised and dealt with some barriers as a result. In addition, we did not consider all possible barriers to validity, in particular inadequate descriptions of interventions, lack of generalisability of patients, and lack of maintenance of effect. A consideration of adequate description of the intervention was beyond the scope of our study, but previous research suggests that many interventions of the sort evaluated in cluster randomised trials are not described in enough detail to enable their adoption in other settings[29] and makes recommendations for description.[30] Although we limited our review to trials in primary care to facilitate comparison with an earlier review, we have no reason to think that our general conclusions are not more widely applicable. Limitation of the review to trials published in journals that are more familiar with this type of trial design might have led to an overoptimistic assessment of quality in comparison with the quality of trials in other journals.

Previous research

There have been several previous reviews of cluster randomised trials.[10-14 16 31-34] Most have indicated poor quality in relation to accounting for clustering in sample size and analysis. Previous statistical publications could have contributed to the increase in trials correctly accounting for clustering.[1 8 9 35-38] Few reviews have explored the other aspects of internal and external validity that we considered. Using slightly different methods, Puffer et al found similar levels of evidence of bias in selection of patients in 36 trials published in the *BMJ*, *Lancet*, and *New England Journal of Medicine* in 1997-2002.[16] In reviewing eight experimental and quasi-experimental studies of HIV prevention, Bonell et al found that none commented on the extent to which study samples were representative of the targeted populations.[21] Our research concurs with their more general conclusion that few studies assessed the generalisability of their results. Recent research suggests that evaluation of process in trials of complex interventions, such as those described here, is important[39]; such evaluations could facilitate an understanding of generalisability.[21] Although we did not identify many trials that had separate process evaluations, we looked for evidence of this only within the trial reports.

Implications—internal validity

Cluster randomised trials are essential for evaluating certain types of intervention and often afford an important advantage over individually randomised trials in terms of internal validity because they are less prone to contamination bias. Nevertheless, other design features of such trials might compromise internal validity, largely through lack of blinding of those delivering care or identifying and recruiting participants or of the individual participants themselves. Sometimes such lack of blinding is inevitable, and sometimes it can be avoided.

To avoid bias, trial investigators should ideally ensure that those who identify or recruit individual participants, or both, are blinded to allocation status. If knowledge of allocation status is unlikely to influence the characteristics of individual participants identified or recruited (for example, if the inclusion process is computerised or unlikely to be subverted for other reasons), investigators should report this. As suggested previously,[16 26] investigators should report identification and recruitment strategies transparently, particularly in relation to the timing of randomisation and intervention delivery, who identifies and recruits individuals, and whether they are blind to allocation status. Investigators should also detail the information given to participants. Full information about the trial might lead to later unblinding of patients, and possibly performance bias, when they are exposed to a particular intervention, while different information given to intervention groups might result in differential recruitment or expectation bias in participants.[40] A few reports we reviewed detailed information given to patients at recruitment; all those that did suggested that patients were given identical information regardless of intervention group, and in many cases an effort was made to ensure that they did not know their allocation status.

This strategy, which might reduce bias, is nevertheless at odds with the generally accepted ethical principle of fully informed consent that proposes that patients should be given full information about the trial that they are participating in.[40] Trial investigators should be aware that this conflict between science and ethics is also present in trials in which individual participants are not recruited; blinding of participants is easy to maintain, but participants receive no information about the trial. When individual participants, those identifying or recruiting them, and outcome assessors cannot be blind to allocation status, this might or might not have serious consequences for internal validity; as some issues seem distinct in these trials we cannot necessarily assume that results regarding factors that affect bias transfer from individually randomised trials to cluster randomised trials. Our study was too small to assess whether these various potential barriers to internal validity actually lead to biased results. Further studies are needed to explore

Table 4 | Selection processes for individual participants and potential for bias (our judgment) because of methods of recruiting/identifying participants

Trial	Description of patient selection process	Potential for bias	Bias mentioned by investigators
Crump[w10]	All family (cluster) members participated, no separate recruitment of family members	None	No
Gattellari[w3]	Records of tests ordered from routine pathology data, no recruitment	None	No
Herbert[w4]	Records of relevant prescriptions from routine data, no recruitment	None	No
Kenealy[w18]	Computerised identification of patients, no recruitment	None	No
Margolis[w21]	Repeat random samples of children's records, no recruitment	None	No
Midlov[w31]	Repeat records of relevant prescriptions from routine data, no recruitment	None	No
Mitchell[w14]	Electronic data collection from practices, no recruitment	None	No
Myers[w32]	Identification centrally, no recruitment	None	No
Omstein[w15]	Patients identified quarterly by computer, no recruitment	None	No
Ruffin[w33]	Repeat random samples of relevant (computer identified) patients' records, no recruitment	None	No
Watson[w23]	Medical records of all children <5 in families (clusters), no separate recruitment of family members	None	No
Witt[w34]	Records of relevant prescriptions from routine data, no recruitment	None	No
Fairall[w1]	Patients recruited by fieldworkers independent from those delivering intervention, both blind to intervention status	Unlikely	No
Griffiths[w27]	Patients identified through routine secondary care data, recruited by researchers after randomisation	Unlikely	No
Hilberink[w30]	Patients identified by computer search, recruitment probably by researchers*†	Unlikely	No
Kools[w5]	All relevant patients identified from intake list, recruited through letter with identical information for all intervention groups	Unlikely	Yes (bias unlikely)
Mohan[w11]	Researchers blind to allocation status identified and recruited fixed number of relevant participants after randomisation and gave both intervention groups same information	Unlikely	No
Powell[w9]	Used clinic records to estimate number of children available in advance of randomisation, probably recruited from this list, but not clear by whom	Unlikely	No
Sandora[w24]	All individual participants approached, recruitment probably by researchers*	Unlikely	No
Seligman[w19]	Researcher identified and recruited relevant patients in waiting rooms after randomisation and did not discuss allocation status with them	Unlikely	No
Smith[w16]	Patients identified from disease registers before randomisation, recruitment probably by researchers*	Unlikely	No
Aittasalo[w29]	Receptionist who had not received intervention identified and recruited patients prospectively after randomisation but was not able to approach all relevant patients	Unclear	Yes (bias possible)
Byng[w8]	List of relevant patients created in each practice with contribution from GP, unclear when created, or who recruited, patients unaware of their practice status	Unclear	No
Edwards[w17]	Patients identified from practice registers by practice staff using standard protocol with help from researchers, recruited by mail, timing not clear	Unclear	No
Glasgow[w20]	Standard protocol used to generate relevant lists of patients and recruit patients after randomisation, but unclear who used it and how; when invited to participate patients received a brochure "describing the project"	Unclear	No
Harmsen[w6]	Patients eligible if they visited their GP on specific days; unclear if or how patients were recruited but they were ignorant of GP assignment	Unclear	No
Laurant[w12]	Doctors reported number of relevant consultations, no recruitment	Unclear	No
Dey[w13]	GPs in clusters identified relevant patients (acute low back pain) throughout study, research assistant confirmed eligibility, unclear if recruited or not	Possible	No
Dietrich[w25]	Clinicians in clusters identified relevant patients (starting or changing treatment for depression) throughout study, probably recruited by clinicians but unclear	Possible	No
Jellema[w26]	GPs identified and recruited consecutive relevant patients during study period, patients kept unaware of study groups	Possible	No
Kendrick[w22]	Eligible patients (women ≥28 weeks' gestation) identified and recruited by cluster based midwives who started intervention at recruitment	Possible	Yes (evidence of bias)
Sondergaard[w2]	Data collection based on relevant patients (consulting for ischaemic heart disease) identified by GPs, no recruitment	Possible	No
van Boeijan[w28]	GPs in clusters identified relevant patients by questionnaire after randomisation	Possible	Yes (evidence of bias)
Welschen[w7]	GPs registered all relevant patients (presenting with acute symptoms of respiratory tract) after randomisation	Possible	Yes (bias possible but unlikely)

*Text was not explicit but was written in such a way that we assumed recruitment was by researchers.
†GPs were asked to confirm eligibility of patients.

this. We suggest that at the design stage of their trials investigators should systematically identify potential biases arising from lack of blinding, the anticipated relative importance of these biases, and whether there is any potential for avoidance.

Implications—external validity

Judgment about external validity can be facilitatedby the reporting of readily available information about numbers and characteristics of clusters approached, recruited, and analysed, and a discussion of generalisability. Information about the characteristics of included health professionals and organisations might be more important in cluster randomised trials, where clusters can have considerable impact on an intervention's effect, than in individually randomised trials, where those delivering the intervention often have minimal impact on its effect. Nevertheless, we found it difficult to judge the generalisability of findings, even with this information. Indeed, a judgment about whether an intervention could be used in a different setting might depend on detailed knowledge

WHAT IS ALREADY KNOWN ON THIS TOPIC

Cluster randomised trials have not always been well designed and analysed

Lack of blinding in the identification and randomisation of individual participants can be a problem

WHAT THIS STUDY ADDS

The extent to which investigators are designing and analysing these trials appropriately has improved

Some trials still do not blind those recruiting and identifying participants

Information relating to cluster generalisability is generally poorly reported

of the area being researched, the setting and healthcare system of the country in which the trial takes place, and the setting and healthcare system to which the intervention might be transferred. Thus, while appropriate guidelines can govern how to assess internal validity, we might be able to assess only whether investigators have presented information that could be used to judge external validity. While frameworks for generalisability developed recently are helpful in this respect, uncertainty about external validity can still remain even when all the parameters of these frameworks are complied with. For cluster randomised trials one key element of this uncertainty is the current lack of knowledge about how clusters with different characteristics respond to different types of intervention. Indeed, most of the trial investigators in our review were not specific about the likely effect of the clusters included on external validity. In individually randomised drug trials, judgments about differences in health status and morbidity between trial participants and other populations are generally easier to make and routine monitoring of drug use after licensing can facilitate a judgment of generalisability.

Although it has already been recommended,[41] no monitoring system exists to assess the wider effectiveness of complex interventions such as those aimed at clusters. Studies to assess the implementation and impact of similar interventions in different types of cluster and setting[21] and exploration and synthesis of empirical evidence from existing trials could also help to fill this knowledge gap. This will mean exploiting the developing science of evidence synthesis; meta-analyses of complex interventions are often not credible and narrative analyses do not provide estimates of the influence of patient or cluster factors on effect sizes. For most cluster randomised trials, investigators should discuss the implementation of the intervention. Again, a better understanding of factors affecting implementation in different circumstances and among different clusters, possibly through evaluations of trial process,[40] would clarify implications for external validity. Our study is too small to form any substantive conclusions about the relation between statistical significance and external validity, although it may be that reporting of certain aspects of external validity is influenced by the statistical significance of findings; this is an issue for future research.

Further observations on validity

In individually randomised trials there is usually a clear distinction between internal and external validity. For example, selection of individual participants into a trial affects external validity, while allocation of individual participants affects internal validity; implementation of the intervention by health professionals affects external validity. In cluster randomised trials, however, this distinction becomes blurred. Lack of blinding to allocation status at identification and recruitment of individual participants might affect internal validity through differential recruitment in two groups but might also affect external validity through the overall profile of participants. Similarly, as health professionals in intervention clusters generally have to implement a wider range of components of an intervention than those in control clusters, failure to implement components will probably be more common in intervention clusters and this might affect internal validity. Thus, while we have focused on internal and external validity, these are to some extent arbitrary distinctions in these trials. Our concern is, nevertheless, to highlight features of these trials that are potential barriers to their validity, both internal and external.

Conclusion

Cluster randomised trials are essential for evaluating certain types of intervention and there are often strong scientific reasons to conduct them. Issues relating to the internal validity of these trials, such as appropriate calculations of sample size and analysis, have been widely disseminated and are now better addressed by the research community. The importance of blinding those who identify and recruit patients has been raised but, as yet, is not always well addressed. There might be fewer barriers to internal validity in trials in which individual participants are not recruited. External validity has not been discussed previously in the literature and seems to be poorly addressed in many trials, yet is arguably as important as internal validity in judging the quality of trials as a basis for healthcare policy.

Contributions: SE conceived the idea for the study, led the research, and wrote the initial draft. GF and DA contributed to design and interpretation. MW and CB extracted data. All authors contributed to the final paper. SE is guarantor.

Funding: SE received a HEFC promising research fellowship.

Competing interests: None declared.

Ethics approval: Not required.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Donner A, Klar N. *Design and analysis of cluster randomised trials in health research*. London: Arnold, 2000.

2 Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med Res Methodol* 2003;3:28.

3 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.

4 Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58:635-41.

5 Kjaergard LL, Gluud C. Funding, disease area, and internal validity of hepatobiliary randomized clinical trials. *Am J Gastroenterol* 2002;97:2708-13.
6 Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;114:906-14.
7 Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *J Clin Epidemiol* 1996;49:435-9.
8 Kerry SM, Bland JM. Trials which randomize practices I: how should they be analysed? *Fam Pract* 1998;15:80-3.
9 Kerry SM, Bland JM. Trials which randomize practices II: sample size. *Fam Pract* 1998;15:84-7.
10 Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* 1995;85:1378-83.
11 Chuang JH, Hripcsak G, Jenders RA. Considering clustering: a methodological review of clinical decision support system studies. *Proc AMIA Symp* 2000:146-50.
12 Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health* 2004;94:393-9.
13 Eldridge S, Ashby D, Feder G, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomised trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004;1:80-90.
14 Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. *Int J Epidemiol* 1990;19:795-800.
15 Farrin A, Russell I, Torgerson D, Underwood M. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clin Trials* 2005;2:119-24.
16 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.
17 Schulz KF, Altman DG, Moher D. Allocation concealment in clinical trials. *JAMA* 2002;288:2406-7.
18 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.
19 Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet* 2002;359:696-700.
20 Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials* 2006;1:e9.
21 Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ* 2006;333:346-9.
22 Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health* 1999;89:1322-7.
23 Glasgow RE, McKay HG, Piette JD, Reynolds KD. The RE-AIM framework for evaluating interventions: what can it tell us about approaches to chronic illness management? *Patient Educ Couns* 2001;44:119-27.
24 Dzewaltowski DA, Estabrooks PA, Klesges LM, Bull S, Glasgow RE. Behavior change intervention research in community settings: how generalizable are the results? *Health Promot Int* 2004;19:235-45.
25 Rabin BA, Brownson RC, Kerner JF, Glasgow RE. Methodologic challenges in disseminating evidence-based interventions to promote physical activity. *Am J Prev Med* 2006;31(4 suppl):S24-34.
26 Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702-8.
27 Mant D. *National working group on R & D in primary care: final report*. London: NHS Executive, 1997.
28 Donaldson MS, Yordy KD, Lohr KN, Vaneslow NA, eds. *Defining primary care: an interim report*. Washington, DC: Institute of Medicine/National Academy Press, 1994.
29 Nation M, Crusto C, Wandersman A, Kumpfer KL, Seybolt D, Morrissey-Kane E, et al. What works in prevention. Principles of effective prevention programs. *Am Psychol* 2003;58:449-56.
30 Davidson KW, Goldstein M, Kaplan RM, Kaufmann PG, Knatterud GL, Orleans CT, et al. Evidence-based behavioral medicine: what is it and how do we achieve it? *Ann Behav Med* 2003;26:161-71.
31 Smith PJ, Moffatt ME, Gelskey SC, Hudson S, Kaita K. Are community health interventions evaluated appropriately? A review of six journals. *J Clin Epidemiol* 1997;50:137-46.
32 Hayes RJ, Alexander ND, Bennett S, Cousens SN. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat Methods Med Res* 2000;9:95-116.
33 Isaakidis P, Ioannidis JP. Evaluation of cluster randomized controlled trials in sub-Saharan Africa. *Am J Epidemiol* 2003;158:921-6.
34 Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;4:21.
35 Murray DM. *Design and analysis of group randomised trials*. New York: Oxford University Press, 1998.
36 Kerry SM, Bland JM. Sample size in cluster randomisation. *BMJ* 1998;316:549.
37 Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ* 1998;317:1171-2.
38 Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assess* 1999;3:iii-92.
39 Oakley A, Strange V, Bonell C, Allen E, Stephenson J. Process evaluation in randomised controlled trials of complex interventions. *BMJ* 2006;332:413-6.
40 Eldridge SM, Ashby D, Feder GS. Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care. *Clin Trials* 2005;2:91-8.
41 Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694-6.

Accepted: 25 February 2008