

# An epistemology of patient safety research: a framework for study design and interpretation. Part 2. Study design

C Brown,<sup>1</sup> T Hofer,<sup>2</sup> A Johal,<sup>1</sup> R Thomson,<sup>3,4</sup> J Nicholl,<sup>5</sup> B D Franklin,<sup>6</sup> R J Lilford<sup>1</sup>

## See Editorial, p 154

<sup>1</sup> Department of Public Health and Epidemiology, University of Birmingham, Birmingham, UK;

<sup>2</sup> University of Michigan Medical School, Ann Arbor, Michigan, USA; <sup>3</sup> National Patient Safety Agency, London, UK;

<sup>4</sup> Newcastle upon Tyne Medical School, Newcastle upon Tyne, UK; <sup>5</sup> University of Sheffield, Sheffield, UK; <sup>6</sup> London School of Pharmacy, London, UK

Correspondence to:  
Dr C Brown, Research Methodology Programme, Department of Public Health and Epidemiology, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK; [c.a.brown@bham.ac.uk](mailto:c.a.brown@bham.ac.uk)

Accepted 1 January 2008

## ABSTRACT

This is the second in a four-part series of articles detailing the epistemology of patient safety research. This article concentrates on issues of study design. It first considers the range of designs that may be used in the evaluation of patient safety interventions, highlighting the circumstances in which each is appropriate. The paper then provides details about an innovative study design, the stepped wedge, which may be particularly appropriate in the context of patient safety interventions, since these are expected to do more good than harm. The unit of allocation in patient safety research is also considered, since many interventions need to be delivered at cluster or service level. The paper also discusses the need to ensure the masking of patients, caregivers, observers and analysts wherever possible to minimise information biases and the Hawthorne effect. The difficulties associated with masking in patient safety research are described and suggestions given on how these can be ameliorated. The paper finally considers the role of study design in increasing confidence in the generalisability of study results over time and place. The extent to which findings can be generalised over time and place should be considered as part of an evaluation, for example by undertaking qualitative or quantitative measures of fidelity, attitudes or subgroup effects.

This is the second in our four-part series of articles on the epistemology of patient safety research. In Part 1, we suggested that there is no clear water between safety and quality issues and identified a causal chain to describe how an ameliorative intervention may impact on an organisation's processes and outcomes. In this second part we will discuss the strengths and weakness of different study designs for patient safety research. Although it could be argued that there is already sufficient literature on study design for health services research, several factors suggest that a focus on study design for patient safety research is warranted. First, patient safety interventions are often "complex interventions" which require a carefully planned evaluation.<sup>1</sup> Second, patient safety interventions are often implemented at cluster, rather than individual level. Cluster-level evaluations require sample size adjustments and cluster randomised controlled trials (RCTs) have additional reporting requirements<sup>2</sup> that are not always met.<sup>3</sup> Third, patient safety interventions are often anticipated to do more good than harm, implying that professional equipoise may be absent<sup>4</sup> and that traditional study designs such as a parallel RCT may not be ethical. It is therefore not

surprising that Ovretveit reports a lack of strong evidence regarding the effectiveness of patient safety interventions,<sup>5</sup> whereas Leape and colleagues suggest that this lack of evidence means that "the traditional evidence-based approach cannot be the sole source of information for advancing patient safety".<sup>6</sup>

The nub of our argument in this article is that the various study designs have different strengths and weaknesses according to the type of intervention being evaluated. We also consider the issue of the unit of allocation and comparison, highlighting the consequences of cluster-based studies. A relatively innovative design, known as the "stepped wedge" design will then be described, since this has particular advantages for the evaluation of many patient safety interventions and deals with the issue of lack of equipoise identified above. We discuss two more methodological issues that are cognate to study design: masking and the ability of the design to allow generalisation beyond study sites. The issue of end points within studies will subsequently be tackled in Part 3 of this series.

## OVERVIEW OF STUDY DESIGNS

Our starting point is a general classification of study designs applicable to any scientific study. At the most fundamental level, studies may be uncontrolled or controlled. If controlled, they may be controlled only for time and/or they may have contemporaneous controls. We will refer to those controlled for time under the umbrella term "before and after studies". Such studies may involve many time points before, during and after the intervention phase. Depending on how the data are analysed these designs may be referred to as time series analysis or statistical process control. The latter is often used for quality control, although the results can also be used for research purposes. Contemporaneous comparisons may be made only after an intervention has been put in place or they may be made both before and after the intervention phase (controlled before and after studies). The stepped wedge design, as we shall see, is a particular kind of controlled before and after study.

Studies with contemporaneous controls may be non-experimental (natural experiments or "quasi-experimental" studies<sup>7</sup>) or they may involve an experimental design where intervention and control units are chosen at random—RCTs. Figure 1 summarises the range of study designs. The study designs identified in fig 1 generally assume that the "intervention" being evaluated is fully developed

## Developing research and practice

and will not evolve after roll-out. This may be the case where interventions have undergone careful preimplementation evaluation, as discussed in Part 1 of this series. However it is possible to “track” the effects of an intervention that changes over time. An example is the two-stage approach to reducing errors made by emergency doctors in interpreting radiographs.<sup>8</sup> The methodological aspects of such “tracker” studies have been described in more detail elsewhere.<sup>9</sup>

### Before and after studies

In some cases contemporaneous controls cannot be generated and so neither RCTs nor natural experiments are possible. This happens when new policies are introduced simultaneously across an entire service—for example, a national directive to remove potassium chloride concentrate from hospital wards or to promulgate a common protocol to avoid wrong site surgery. In such cases it is not possible to conduct comparative studies within a service and it is necessary to rely on before and after comparisons. The results of a before and after evaluation of the National Patient Safety Agency’s Patient Safety Alert on correct site surgery has recently been published through the Patient Safety Research Portfolio (PSRP).<sup>10</sup> Before and after studies are also common when interventions are introduced under the control of a single organisation such as one hospital. Although before and after studies may be the only practical method of evaluation in many cases, they are a relatively weak method to distinguish cause and effect even when a change is statistically significant. This weakness is important if any observed change could plausibly be attributed to developments in the service other than the intervention of interest. Many factors determine how much confidence can be placed on the results of before and after studies:

- *The number of before and after measurements.* A statistically significant interruption in a long series of observations (time series; “control chart”) is more impressive evidence of cause

and effect than differences between single before and after observations. Regression to the mean is less likely if serial observations show that the improvement was not preceded by a random “high”. For example, the dramatic and sustained drop in maternal mortality in developed countries when antibiotics and blood transfusions became widespread around 1943 suggests that these interventions were highly efficacious.

- *The magnitude and rate of the change.* The size of the above drop in maternal mortality seemed incompatible with other possible explanations—for example, emancipation of the female work force during the second world war or better nutrition for poorer women under the egalitarian influence of rationing.
- *The plausibility of the intervention.* Haemorrhage and infection were the major causes of maternal mortality before the introduction of antibiotics and blood transfusion and these interventions were highly effective in other contexts.
- *Compatibility within other contemporaneous evidence.* Mortality fell in equal proportion in all social classes (reducing the credibility of a social explanation), irrespective of birth order and particularly in the categories of haemorrhage and infection deaths. We return to this idea of “triangulation” of multiple sources of evidence later in this series of articles.

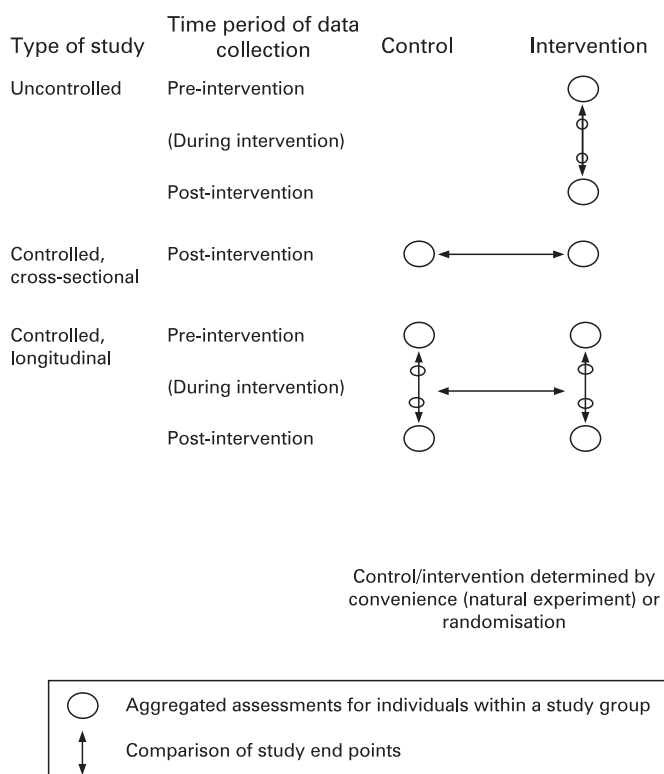
Before and after studies can provide convincing evidence of effectiveness as in the example of maternal mortality. Recently, a similar example has emerged in the patient safety field where a before and after study provided evidence that a multifaceted intervention dramatically reduced central line infections in Michigan intensive care units.<sup>11</sup>

### Controlled comparative studies

In the many cases where before and after studies are potentially biased, comparative studies between sites exposed to the intervention and controls which are not exposed provide a much better basis for inferences about effectiveness. Here, the “intervention” and “control” groups consist of one of two fundamentally different units of comparison: individuals and “clusters” and we consider this issue below.

### The unit of comparison

In Part 1 of this series we argued that service delivery interventions are usually targeted at the service within which patients are cared for, rather than the care itself. Since such interventions will affect a group of patients, cluster studies are often necessary.<sup>12–15</sup> At the limit, it is simply physically impossible to target the intervention at the individual patient level: adopting a distinctive human resource policy, reducing staff working hours or altering the nurse/patient ratio, for example. Even when it is possible to hypothecate an intervention on individuals, cluster studies may need to be considered. This happens when there is likely to be a high degree of contamination (when an intervention intended for members of the trial arm of a study is received by members of the control arm). For example, it may be possible to toggle a decision aid facility on and off in a randomised sequence but clinicians may be influenced by previous exposure to the system even when the aid is disabled. The bias due to contamination occurs in one direction only: it dilutes the measured intervention effect. Enhancing sample size can counteract this bias. The drawback of cluster studies is the loss in power that results from greater similarity across individuals within a cluster than across individuals between clusters. Again, this has implications for sample size, as we discuss below.



**Figure 1** Basic study designs.

Most evaluations of computer-based interventions to improve safety have been based on individual randomisation but have nevertheless shown positive results, implying either that contamination was not widespread or that the underestimated effect size was still significant.<sup>14 15</sup> However, when contamination exceeds about 30% a cluster study is generally more efficient.<sup>16</sup> Clusters typically consist of different sites (hospitals, general practices, etc.) but other types of cluster may be used. For example, clusters may comprise patients treated by doctors exposed to different interventions: Landrigan and colleagues<sup>17</sup> randomised clinicians to different types of on-call duty rota. The cluster was the group of patients treated by a particular clinician.

As with other studies, sample size requirements in cluster studies depend on the size of the effect sought and the risks of false positive and false negative study results that can be accepted. However, the sample size also depends on the extent to which end points tend to cluster within an organisation, and this is measured by the intraclass correlation (ICC). The ICC ranges between 0 and 1 and will increase as the variance in end points between individuals within clusters falls. If the ICC = 0, there is no correlation between individuals in a cluster and the study is effectively a parallel design. If the ICC = 1, the responses of individuals within a cluster are identical and the effective sample size is the number of clusters.<sup>18</sup> ICC values are generally between 0.01 and 0.02 for human studies.<sup>18</sup> A typical two-arm comparative study might include 20 clusters in each arm with 40 units/individuals in each cluster—such a study would be sufficient to detect a statistically significant reduction in error rate from 10% to 5% assuming an ICC of 0.02 ( $\alpha = 0.05$ ;  $\beta = 0.2$ ). The total sample size of 1600 is almost twice the requirement of 868 if clustering is not taken into account. Figure 2 shows the trade-off between the number of clusters and cluster size required to detect a difference in error rate from 10% to 5% with different ICC values. One important conclusion to draw from the figure is the diminishing marginal returns to increasing cluster size on the power of the study.<sup>19</sup>

Ukoununne *et al*<sup>20</sup> identify two types of cluster design: cohort designs, in which repeated measures are taken from the same subjects in each cluster; and repeated cross-sectional designs, in which the sample of subjects within each cluster

changes with each measurement occasion. The latter design affords an opportunity for the measurement of terminal end points similar to those used in clinical research. Terminal end points are those that can only happen once: death is the most obvious example. In clinical research, such end points can be measured once only, whereas terminal end points can be measured repeatedly in safety/service delivery and organisational (SDO) research using cluster (cross-sectional) designs. Thus the effect of a safety intervention on mortality or rates of central line infection or clinical error can be measured in a before and after design.

Clearly arranging an adequately powered cluster study poses logistical difficulties, and, in particular, it is necessary to win collaboration from the managers and policy makers who control the purse strings and hence who can put an intervention into effect around an evaluation framework.<sup>21</sup> Nevertheless, some shining examples of such studies exist. For example, Hillman and colleagues<sup>22</sup> randomised Australian hospitals to have or not have an educational intervention to promote the early recognition and treatment of the deteriorating patient.

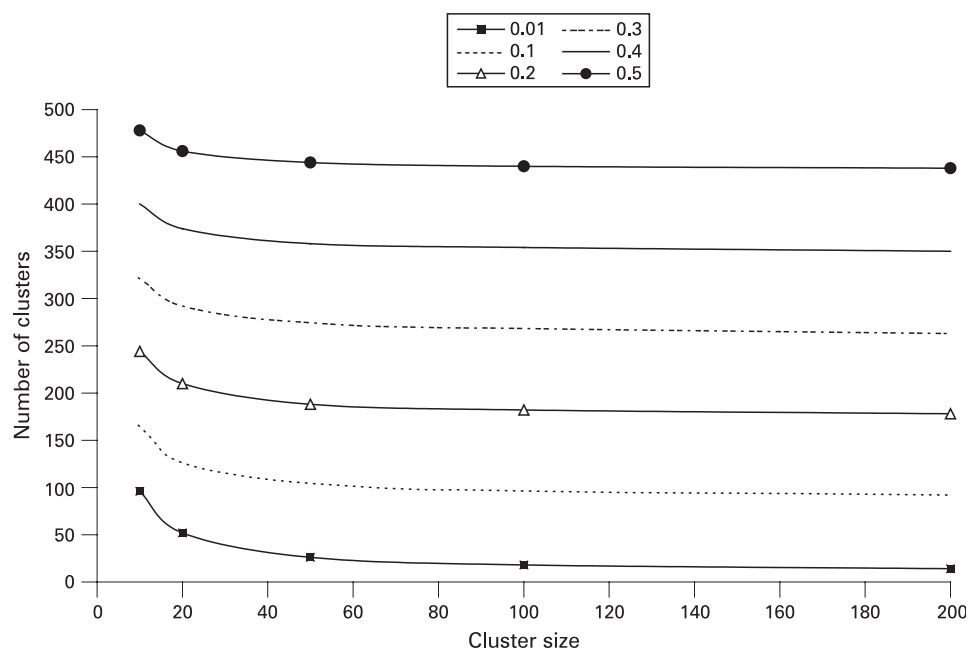
### Design elements of controlled comparative studies

We have distinguished two design variables that distinguish controlled comparative studies:

- ▶ before and after observations versus observations made only after an intervention has been put in place;
- ▶ randomised comparisons versus natural experiment.

Clearly non-randomised postintervention comparisons are the least reliable of these designs because of possible inherent differences between intervention and control sites. Statistical adjustment for confounders can only take into account any known and observed confounding variables and bias frequently originates in hidden variables.<sup>23</sup> Randomised studies with preintervention and postintervention measures are arguably the strongest design (see table 1), but non-randomised comparative studies with before and after measurements may be nearly as good because rates of change are less confounded than cross-sectional data. This is because the confounding factor would need to affect the propensity to change in response to an intervention, net of any baseline differences—the baseline

**Figure 2** Clustered sample size calculations for given ICC values (box). The calculations are based on detecting a difference in error rate between control (0.1) and intervention (0.05), with  $\alpha = 5\%$  and power = 80%.



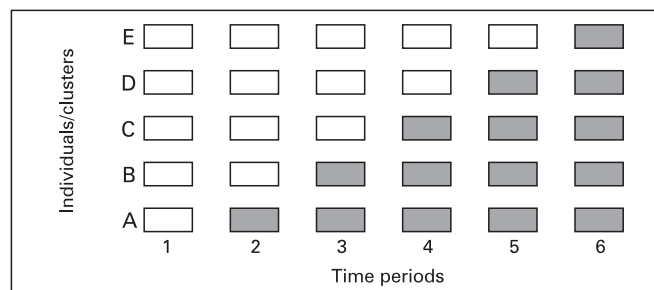
## Developing research and practice

differences themselves are controlled by the before and after nature of the study. Whether randomised or not, a controlled before and after design affords control for secular change, since the intervention effect is estimated as a rate of change above and beyond any background rate of change. With an appropriate statistical model both secular trends and intervention effects can be modelled.<sup>24</sup> An empirical study to compare natural and randomised before and after studies was recently attempted under the sponsorship of the NHS Methodology Programme, but the results were inconclusive.<sup>25</sup>

Before and after comparisons yield not only greater accuracy but also greater precision than cross-sectional studies with only postintervention measurements. Sample size calculations for comparisons of rates of change in cluster studies do not need to take account of the ICC of the end points, but of the ICC of propensity to change net of baseline differences. Murray and Blitstein<sup>30</sup> provide both theoretical justification and practical evidence that these latter ICCs based on pre–post differences are smaller than ICCs based on cross-sectional data alone. Hence in studies where the primary end point is based on rates of change, the effect of clustering on sample size requirements is less pronounced than in cluster studies based on end points collected only in the postintervention stage. However, although there are databases of empirical evidence of ICCs, ICCs are not given for pre–post differences (with most ICCs calculated using pre-intervention data), even where these are focused on studies aimed at changing professional practice (eg, Health Services Research Unit, University of Aberdeen; <http://www.abdn.ac.uk/hsru/epp/cluster.shtml> (accessed 21 April 2008)). Hence the point that comparing changes over time requires use of the ICC of change is insufficiently appreciated—and did not arise, for example, in a systematic review of cluster studies.<sup>20</sup> The idea that change is both more precise and less subject to bias than static measurements has an important implication for research policy: an evaluation framework should be put in place before a new service is implemented, even if an RCT is politically inappropriate or practically unobtainable.<sup>31</sup>

### Stepped wedge trial designs

The stepped wedge is a particular controlled comparative study design that may be appropriate for evaluations of certain patient safety interventions. In a stepped wedge design, an intervention is rolled-out sequentially to the trial participants (either individuals or clusters) over a number of time periods.<sup>32</sup>



**Figure 3** Stepped wedge trial design. Each cell represents a data collection point. Blank cells represent control periods and shaded cells represent intervention periods. At the beginning of the second and subsequent time periods, the next individual/cluster “crosses” from control to intervention.

The order in which the different individuals or clusters receive the intervention is typically determined at random and, by the end of the allocation, all individuals or groups will have received the intervention. Data on key outcomes are usually collected at regular intervals throughout the study, whenever a new group receives the intervention. Figure 3 shows an example of the logistics of a stepped wedge trial design.

Cook and Campbell were possibly the first authors to consider the potential for *experimentally staged introduction* in a situation when an intervention cannot be delivered concurrently to all units.<sup>33</sup> The first empirical example of this design is the Gambia Hepatitis Study, which was a long-term effectiveness study of hepatitis B vaccination in the prevention of liver cancer and chronic liver disease.<sup>34</sup> More recently, the design has been used to evaluate the introduction of a critical care outreach service on in-hospital mortality and length of stay in a general acute hospital.<sup>35</sup>

The stepped wedge design differs from both parallel and cross-over designs in that, by the end of the trial, all participants will have received the intervention, without the intervention being withdrawn from any participants. There are two primary motivations for using a stepped wedge design in evaluating a particular patient safety intervention:

- From an ethical point of view, the stepped wedge design may be appropriate when there is a prior belief that the intervention will do more good than harm.<sup>36</sup> This may often be the case in evaluations of public health and epidemiological policies such as vaccination, screening and training or,

**Table 1** Controlled study design matrix

Phasing	Allocation	
	Randomised	Natural experiment
Postintervention	Variance in end points due to baseline differences will yield an imprecise result unless the number of participants is large <i>Example:</i> RCT in which nurses were randomly assigned to act as dedicated medication nurses or continue as general nurses, with a comparison of error rates between the two groups <sup>26</sup> No baseline measurements of error rates were made	Risk that comparisons will be confounded by differences between departments or organisations <i>Example:</i> National Evaluation of Sure Start <sup>27</sup> Due to the timing of the Sure Start evaluation, baseline measurements of key outcomes were not possible. Comparisons of the effect of the intervention are being made between Sure Start and “Sure Start to be” communities, although group assignment was not made randomly
Before and after	Allows for specific comparison of change net of any baseline differences. Enables comparisons to be made between sites that change most or least <i>Example:</i> Evaluation of a targeted risk factor reduction plan to prevent falls in older in-patients <sup>28</sup> A pre–post design proved to be essential as differences in the number of falls and injuries from falls between control and intervention wards were statistically significant in the preintervention period	Controls for baseline difference possible—see text <i>Example:</i> Evaluation of nationally mandated drug use reviews in nursing homes <sup>29</sup> Levels of inappropriate medication use in nursing homes before and after the review policy were compared to assisted living facilities. The results identified that the fall in inappropriate medication use could not be attributed to the national drug use reviews, as there was a reduction in inappropriate use in both nursing homes and assisted living facilities

with particular relevance to the current series of articles, patient safety interventions that have undergone careful preimplementation evaluation (PIE). In such circumstances it might be considered unethical to exclude any participants from receiving the intervention and recruitment might be enhanced when no centre feels excluded.<sup>37</sup> The stepped wedge design therefore enables an RCT to go ahead in circumstances where professional equipoise is absent.<sup>38</sup>

- ▶ Stepped wedge designs are particularly useful when, for logistical, practical or financial reasons, it is not possible to deliver an intervention to all participants (whether individuals or clusters) simultaneously, but when a rigorous evaluation of the effectiveness of the intervention is considered desirable.

The stepped wedge design also has scientific advantages. First, individuals/clusters in the trial act as their own controls and hence provide data points in both control and intervention sections of the wedge, as would be the case in a cross-over design. This feature of the stepped wedge design reduces the risk of bias, which is most important in non-randomised studies. In cluster studies, this implies that ICCs have a minimal effect on power.<sup>39</sup> Second, data analysis in a stepped wedge trial primarily involves comparing all data points in the control section of the wedge with those in the intervention section. In the statistical model the effects of time can be included, hence controlling for temporal changes in the effectiveness of the intervention.

The main disadvantage of the stepped wedge design is the amount of data collection required. Hence the cost of using the design could be prohibitive, unless the study can use routine, or other easily collected data. Care is required in undertaking the statistical analysis, although a guide to such analysis has recently been published.<sup>39</sup>

## MASKING

As with all other evaluations, masking patients, caregivers and observers, as well as those undertaking the statistical analysis<sup>40</sup> is important in minimising information bias. The importance of masking is related to the nature of the end point being assessed, since lack of masking is more likely to lead to bias when the end points are subjective rather than objective.<sup>41</sup> The ability to mask participants can depend on whether a placebo intervention can be used, which is not always possible with patient safety research. For example, although it is possible (although not ethical) to use a placebo to an alcohol hand sanitising gel, a placebo to an educational programme to encourage the use of such gel is less feasible.

Where assessments are being made about the quality of care, reviewers tend to give worse ratings if an adverse outcome occurred—hindsight bias.<sup>42</sup> This can bias the size of a patient safety problem and lead to exaggerated estimates of cost effectiveness. Evaluating all cases of care and looking for all errors, rather than first selecting adverse events and then looking to see whether an error occurred, can reduce such a bias. Hindsight does not bias assessment of relative safety improvement in a comparative study if it is applied equally across comparison groups. Here, bias is a risk if the observer is aware of the group (intervention or control) to which an individual or cluster has been assigned. Observers should therefore be blinded to the “source” of the particular care data. For example, the authors of this paper are involved in a study of patient safety involving case note review. The case notes will be masked as to the site and time period of care and in this way the observers will not know whether a particular case note originated from an intervention or control site or from the preintervention or

postintervention period.<sup>43</sup> Similarly, Landrigan and colleagues<sup>17</sup> randomised clinicians to different on-call duty rotas and the observers who measured their error rates were masked about whether a particular doctor was an intervention or control subject.

Lack of masking of caregivers may result in a Hawthorne effect, where caregivers change their behaviour in response to being studied, rather than as a result of the intervention. For example, in Landrigan’s trial it would not have been possible to blind clinicians to the on-call rota to which they had been assigned. Knowledge of their rota assignment may (subconsciously) have affected clinicians’ behaviour. In such cases, it is important that subjects (in this case the clinicians) do not know what end points are being measured during the study. Lack of masking of caregivers is a particular risk with many safety interventions since these are often educational and behavioural interventions. In these circumstances randomisation is a less fail-safe guard against bias than in much clinical research where such masking is often possible.<sup>44</sup>

## GENERALISING BEYOND STUDY SITES: THE IMPORTANCE OF CONTEXT

A distinction is sometimes made between internal and external validity. Bowling states that<sup>45</sup>:

A [measuring] instrument is assigned [internal] validity after it has been satisfactorily tested in the populations for which it was designed ... external validity ... refers to the generalisability of the research findings to the wider population of interest.

Hence a study has internal validity if it avoids bias in the population studied. It has external validity if the results hold good when “generalised” across time and place. This extrapolation may be undermined because of systematic differences between study and non-study sites, in terms of service users, caregivers or the way in which services are organised or funded. In other words, the context within which a study is carried out is important. This is because the results might not apply if the context changes. Therefore it is important to examine the context in which the study is carried out and compare it to that in which the study results may impact. For example, family therapy is found to improve safety for patients with severe mental illness when all available RCTs are combined in a single meta-analysis. However, the effect of family therapy has attenuated progressively over time: the observed effect is much greater in earlier trials.<sup>46</sup> Clearly the context has changed with time and the most plausible explanation is that other effective measures, such as assertive outreach, have all but removed the headroom for further gains.

While racial and cultural differences across countries and health systems may affect responses to a *clinical* intervention, the problem of context may be even greater when we try to generalise from studies of *service delivery* interventions. Thus it is arguably more problematic to generalise across countries and health systems in the case of safety SDO interventions than in the case of typical clinical interventions. Furthermore, important differences in capacity to benefit may exist. For example, there may be a “ceiling” effect whereby non-intervention sites are already very good and have less headroom for further improvement than the intervention sites that demonstrated positive results. Three points can be made about the issue of generalisation:

- ▶ the issue applies to any study that may be done, not just contemporaneous comparative studies;

## Developing research and practice

- ▶ the more consistent results are over time and place (ie, over different contexts), the more confident one can be in making generalisations—larger studies are better than small studies, multicentre studies are better than single centre studies and replicated studies are better than unreplicated studies;
- ▶ the extent to which it is “safe” to generalise over place and time is a matter of judgement.

In order to make judgements regarding generalisability, it is necessary to have some knowledge of any systematic differences between study and non-study sites. In the event that such differences exist, the theory as to why and how an intervention may work will help in predicting the impact of such differences on effectiveness. A theory-free, evidence-based approach is inappropriate whenever context is important or, in other words, the more an intervention is thought (and has been shown) to be sensitive to context, the more important it is to peer into the “black box” and understand the “how” and not just the “what”. For example, Oakley and colleagues<sup>47</sup> highlight how the effectiveness of sex education varied according to interactions between the extent to which the education was “participative” and who provided the education (peers or teachers). Similarly, qualitative or observational ethnographic studies undertaken in parallel with RCTs, as well as retrospective realist reviews or meta-regression models, can give important insights into the why and how that an intervention was effective or not.<sup>48–51</sup>

In addition to contextual differences, studies associated with greater (and significant) effects tend to be those where the intervention was implemented with greater fidelity. It is therefore important that context and the fidelity with which an intervention was implemented are clearly described in evaluations of patient safety interventions, since these will properly inform judgements about the generalisability of the findings. We stressed the importance of a causal chain linking structure and process to outcomes in Part 1 of this series and we will return to the theme of “triangulation” of information in Parts 3 and 4.

### CONCLUSION

This article has considered the types of study design that could be employed in patient safety research. We have defined certain circumstances where before and after studies can provide convincing evidence of effectiveness. However, contemporaneous controls are usually required to distinguish cause and effect. Controlled before and after studies are much preferable to cross-sectional controlled studies, since the potential for bias is reduced by an analysis of “differences in differences”. We advocate more widespread use of stepped wedge designs in the evaluation of patient safety interventions, given that such designs do not mean withholding the intervention from any participants and also allow for an intervention to be rolled-out over time. Organising studies requires collaboration between those who commission services and those who commission research: given the benefits in terms of accuracy and precision associated with before and after studies, communication between service providers and researchers is essential if such designs are to be operationalised. Masking of patients, caregivers, observers and analysts is a crucial principle whatever study design is used, in order to avoid information bias, although this may be difficult to achieve in patient safety research. It is important to incorporate an assessment of the factors affecting the generalisability of the results of patient safety research into the research itself, since the results may be

particularly sensitive to the context in which the research is undertaken.

**Acknowledgements:** We would like to acknowledge the support of the National Coordinating Centre for Research Methodology and the Patient Safety Research Programme. The authors would also like to acknowledge the contributions of attendees at the Network meetings and the helpful comments of the peer reviewers.

**Competing interests:** None.

**Authors' contributions:** RL conceived the Network and formulated the first draft of the report and the current paper with assistance from AJ. CB contributed to subsequent drafts of the report and this paper. BDF, TH, RT and JN contributed to the Research Network and provided comments on drafts of the report and papers in their areas of expertise.

This work forms part of the output of a Cross-Council Research Network in Patient Safety Research funded by the Medical Research Council (Reference G0300370). More details of the Research Network can be found at: <http://www.pcpoh.bham.ac.uk/publichealth/psrp/MRC.htm>

### REFERENCES

1. **Campbell M**, Fitzpatrick R, Haines A, *et al*. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;**321**:694–6.
2. **Campbell MK**, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;**328**:702–8.
3. **Bland JM**. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;**4**:21.
4. **Lilford RJ**, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;**88**:552–9.
5. **Ovretveit J**. *Which interventions are effective for improving patient safety? A review of research evidence*. Stockholm: Karolinska Institutet Medical Management Centre, 2005.
6. **Leape LL**, Berwick DM, Bates DW. What practices will most improve safety? Evidence-based medicine meets patient safety. *JAMA* 2002;**288**:501–7.
7. **Campbell DT**, Stanley JC. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1963.
8. **Espinosa JA**, Nolan TW. Reducing errors made by emergency physicians in interpreting radiographs: longitudinal study. *BMJ* 2000;**320**:737–40.
9. **Lilford RJ**, Braunholtz DA, Greenhalgh R, *et al*. Trials and fast changing technologies: the case for tracker studies. *BMJ* 2000;**320**:43–6.
10. **Rhodes P**, Giles S, Cook GA, *et al*. Evaluation of Patient Safety Alert on correct site surgery. 2007. [http://pcpoh.bham.ac.uk/publichealth/psrp/Pdf/PS\\_044/PS044\\_CSS\\_Wright\\_Fina\\_Report.pdf](http://pcpoh.bham.ac.uk/publichealth/psrp/Pdf/PS_044/PS044_CSS_Wright_Fina_Report.pdf) (accessed 1 Apr 2008).
11. **Pronovost P**, Needham D, Berenholtz S, *et al*. An intervention to decrease catheter-related bloodstream infections in the ICU. *N Engl J Med* 2006;**355**:2725–32.
12. **Edwards SJL**, Braunholtz DA, Lilford RJ, *et al*. Ethical issues in the design and conduct of cluster randomised trials. *BMJ* 1999;**318**:1407–9.
13. **Donner A**. Some aspects of the design and analysis of cluster randomised trials. *Appl Stat* 1998;**47**:95–113.
14. **Lilford RJ**, Kelly M, Baines A, *et al*. Effect of using protocols on medical care: randomised trial of three methods of taking an antenatal history. *BMJ* 1992;**305**:1181–4.
15. **Kaushal R**, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch Intern Med* 2003;**163**:1409–16.
16. **Keogh-Brown MR**, Bachmann MO, Shepstone L, *et al*. Contamination in trials of educational interventions. Birmingham: University of Birmingham, 2005.
17. **Landrigan CP**, Rothschild JM, Cronin JW, *et al*. Effect of reducing interns' work hours on serious medical errors in intensive care units. *N Engl J Med* 2004;**351**:1838–48.
18. **Killip S**, Mahfouz Z, Pearce K. What is an intraclass correlation coefficient? Crucial concepts for primary care researchers. *Ann Fam Med* 2004;**2**:204–8.
19. **Campbell MK**, Thomson S, Ramsay CR, *et al*. Sample size calculator for cluster randomised trials. *Comput Biol Med* 2004;**34**:113–25.
20. **Ukoumunne OC**, Gulliford MC, Chinn S, *et al*. Methods of evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* 1999;**3**:iii–92.
21. **Lilford RJ**, Jecock R, Chard J, *et al*. Commissioning health services research: an iterative method. *J Health Serv Res Policy* 1999;**4**:164–7.
22. **Hillman K**, Chen J, Cretikos M, *et al*. Introduction of the medical emergency team (MET) system: a cluster randomised trial. *Lancet* 2005;**365**:2091–7.
23. **Lilford RJ**, Mohammed M, Spieglerhalter D, *et al*. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;**263**:1147–57.
24. **Saint S**, Hofer TP, Rose JS, *et al*. Use of critical pathways to improve efficiency: a cautionary tale. *Am J Manage Care* 2003;**9**:758–65.
25. **Oliver S**. Comparison of the results of RCTs with other study designs. 2007. [http://pcpoh.bham.ac.uk/publichealth/nccrm/Publication\\_RH03\\_JH09\\_SO.htm](http://pcpoh.bham.ac.uk/publichealth/nccrm/Publication_RH03_JH09_SO.htm) (accessed 1 Apr 2008)

26. **Greengold NL**, Shane R, Schneider P, *et al*. The impact of dedicated medication nurses on the medication administration error rate. *Arch Intern Med* 2003;**163**:2359–67.
27. National Evaluation of Sure Start. <http://www.ness.bbk.ac.uk> (accessed 1 Apr 2008).
28. **Healey F**, Monro A, Cockram A, *et al*. Using targeted risk factor reduction to prevent falls in older in-patients: a randomised controlled trial. *Age Aging* 2004;**33**:390–5.
29. **Briesacher B**, Limcangco R, Simoni-Wastila L, *et al*. Evaluation of nationally mandated drug use reviews to improve patient safety in nursing homes: a natural experiment. *J Am Geriatr Soc* 2005;**53**:991–6.
30. **Murray DM**, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomised trials. *Eval Rev* 2003;**27**:79–103.
31. **Carney PA**, Nierenberg DW, Pipas CF, *et al*. Educational epidemiology: applying population-based design and analytic approaches to study medical education. *JAMA* 2004;**292**:1044–50.
32. **Brown CA**, Lilford RJ. The stepped wedge trial design: systematic review. *BMC Med Res Methodol* 2006;**6**:54.
33. **Cook TD**, Campbell DT. *Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton Mifflin Company, 1979.
34. **The Gambia Hepatitis Study Group**. The Gambia hepatitis intervention study. *Cancer Res* 1987;**47**:5782–7.
35. **Priestly G**, Watson W, Rashidian A, *et al*. Introducing critical care outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Med* 2004;**30**:1398–404.
36. **Smith PG**, Morrow RH. *Field trials of health interventions in developing countries: a toolbox*. London: Macmillan Education Ltd, 1996.
37. **Hutson AD**, Reid ME. The utility of partial cross-over designs in early phase randomised prevention trials. *Control Clin Trials* 2004;**25**:493–501.
38. **Lilford RJ**. Formal measurement of clinical uncertainty: prelude to a trial in perinatal medicine. *BMJ* 1994;**308**:111–12.
39. **Hussey MA**, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28**:182–91.
40. **Schulz KF**, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet* 2002;**359**:696–700.
41. **Morrison B**, Lilford RJ. Organisational research methods: closing the gap. *Lancet* 2000;**355**:71.
42. **Caplan RA**, Posner KL, Cheney FW. Effect of outcome on physician judgements of appropriateness of care. *JAMA* 1991;**265**:1957–60.
43. **Johal AJ**. Protocol for the evaluation of the Health Foundation Safer Patients' Initiative, 2006. <http://www.pcpoh.bham.ac.uk/publichealth/psrp/SPI.htm> (accessed 2 Apr 2008).
44. **Chalmers I**. Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. *Ann Am Acad Pol Soc Sci* 2003;**589**:22–40.
45. **Bowling A**. Techniques of questionnaire design. In: Bowling A, Ebrahim S, eds. *Handbook of health research methods*. Maidenhead: Open University Press, 2005.
46. **Marshall MN**, Lockwood A. Assertive community treatment for people with severe mental disorders. *Cochrane Database Syst Rev* 1998;(2):CD001089.
47. **Oakley A**, Strange V, Bonell C, *et al*. Process evaluation in randomised controlled trials of complex interventions. *BMJ* 2006;**332**:413–16.
48. **Thomson RG**, Eccles MP, Steen IN, *et al*. A patient decision aid to support shared decision-making on anti-thrombotic treatment of patients with atrial fibrillation: randomised controlled trial. *Qual Saf Health Care* 2007;**16**:216–23.
49. **Murtagh MJ**, Thomson RG, May CR, *et al*. Qualitative methods in a randomised controlled trial: the role of an integrated qualitative process evaluation in providing evidence to discontinue the intervention in one arm of a trial of a decision support tool. *Qual Saf Health Care* 2007;**16**:224–9.
50. **Greenhalgh T**, Kristjansson E, Robinson V. Realist review to understand the efficacy of school feeding programmes. *BMJ* 2007;**335**:858–61.
51. **Burns T**, Catty J, Dash M, *et al*. Use of intensive case management to reduce time in hospital in people with severe mental illness: systematic review and meta-regression. *BMJ* 2007;**335**:336.