An epistemology of patient safety research: A framework for study design and interpretation

Report prepared for the MRC UK, March 2006

Celia Brown PhD*
Research Fellow, Research Methodology Programme, Department of Public Health and Epidemiology, The University of Birmingham, Edgbaston, Birmingham B15 2TT
Tel: 0121 414 8466
Fax: 0121 414 6216
Email: c.a.brown.1@bham.ac.uk

Prakash Patil PhD
Reader, Department of Public Health and Epidemiology, The University of Birmingham, Edgbaston, Birmingham B15 2TT
Tel: 0121 414 6191
Fax: 0121 414 6216
Email: p.n.patil@bham.ac.uk

Timothy Hofer MD MSc
Associate Professor of Medicine
University of Michigan Medical School, VA HSR&D Field Program, PO Box 130170, Ann Arbor, MI 48113-0170, USA
Tel: 001 734 761 2399
Fax: 001 734 761 2939
Email: THOFER@UMICH.EDU

Amirta Johal BSc MSc
Research Officer, The Patient Safety Research Portfolio, Department of Public Health and Epidemiology, The University of Birmingham, Edgbaston, Birmingham B15 2TT
Tel: 0121 414 8157
Fax: 0121 414 6216
Email: johala@medgp3.bham.ac.uk

Richard Thomson MD FRCP FFPHM
Director of Epidemiology & Research
National Patient Safety Agency, 4-8 Maple Street, London W1T 5HD
Professor of Epidemiology and Public Health
Newcastle upon Tyne Medical School, Newcastle upon Tyne NE2 4HH
Tel: 027 927 9359
Email: Richard.Thomson@npsa.nhs.uk

An epistemology of patient safety research

March 2006
Jon Nicholl MSc CStat FFPHM
Director of Medical Care Research Unit
University of Sheffield, 30 Regent Road, Sheffield S1 4DA
Tel: 0114 222 5201
Email: j.nicholl@sheffield.ac.uk

Bryony Dean Franklin PhD
Principal Pharmacist, Clinical Services and Director, Academic Pharmacy Unit
London School of Pharmacy, 29/39 Brunswick Square, London WC1N 1AX
Tel: 020 7753 5980
Email: bdean@hhnt.nhs.uk

Richard J Lilford PhD FRCOG FRCP FFPH
Professor of Clinical Epidemiology, Director of the Patient Safety Research Portfolio of Department of Health and Epidemiology, The University of Birmingham, Edgbaston, Birmingham B15 2TT
Tel: 0121 414 2226
Fax: 0121 414 6216
Email: r.j.lilford@bham.ac.uk

*Corresponding author
Contents

List of Abbreviations 5
Abstract 6
Introduction 7
  Our point of departure.........the nature of patient safety interventions 7
  Low and high prevalence incidents: the safety/quality continuum 8
Evaluation before implementation 10
Evaluation of Safety Interventions: Study Design 13
  Overview 13
  Uncontrolled before and after studies 14
  The Unit of Comparison 15
  Controlled Comparative Studies 18
  Stepped Wedge Trial Designs 20
  The issue of masking 23
  Generalising beyond study sites 24
End-points and Measurement 28
  Overview 28
  Outcomes for patients 29
    Signal to Noise Ratio 29
    Consistent Measurement of Outcomes 31
    Biased Denominators 32
Surrogate End-points 33
Fidelity 33
Intervening Variables 35
  Patient Safety Culture 35
Clinical Process Measurements: Error 36
  Denominator-free error detection 37
  Trigger Tools 38
  Error Rates 39
  Methods for the detection of error 40
  Implicit Versus Explicit Review 42
Validity 44
Reliability 45
Surrogate End-Points: Summary 48
  Collateral Damage 49
One size does not fit all 51
  Overview 51
  Logistic/pragmatic constraints 51
  A Priori assessment of benefit and harm 51
  The possibilities and consequences for improvement 53
Target Audience 54
  Which size fits whom? 55
    Mixed method designs 55
Acknowledgements 59
References 71
List of Tables and Figures

Table 1: Examples of clinical events/error 61
Table 2: Controlled study design matrix 64
Table 3: Some examples of safety issues 65
Table 4: Comparison of Implicit and Explicit methods 69
Table 5: Sample size calculations for Inter-rater reliability 70

Figure 1: The Quality/Safety Continuum 60
Figure 2: Basic study designs 62
Figure 3: Clustered sample size calculations for given ICC values 63
Figure 4: Stepped Wedge trial design 66
Figure 5: Generic and Specific interventions and end-points 67
Figure 6: Intellectual framework to classify links between poor care and outcome 68
**List of Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTA</td>
<td>Health Technology Assessment</td>
</tr>
<tr>
<td>ICC</td>
<td>Intra-cluster correlation</td>
</tr>
<tr>
<td>MRC</td>
<td>Medical Research Council</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>NPSA</td>
<td>National Patient Safety Agency</td>
</tr>
<tr>
<td>NRLS</td>
<td>National Reporting and Learning System</td>
</tr>
<tr>
<td>PIE</td>
<td>Pre-implementation Evaluation</td>
</tr>
<tr>
<td>RCT</td>
<td>Randomised Controlled Trial</td>
</tr>
<tr>
<td>SDO</td>
<td>Service Delivery and Organisational</td>
</tr>
</tbody>
</table>
Abstract

This paper provides a detailed epistemology of patient safety research. We begin by considering the nature of patient safety interventions and their evaluations, which are a particular subset of Service delivery and organisational research. The main two sections of this paper consider the two key elements of the epistemology of patient safety research: study design and end-points (what to measure as evidence of safety). The former section on study design considers the effects of randomisation and whether data are collected both before and after the intervention. We also discuss the potential for applying a stepped wedge design, which may be particularly appropriate for evaluating patient safety interventions that are expected to do more good than harm. On end-points, we consider the difficulties associated with measuring patient outcomes and highlight a number of possible surrogate end-points, focusing on the measurement of error through case note review. We conclude by providing some guidance on how epistemological questions can be addressed, although we propose that mixed method study designs should be used if possible, in order to maximise the potential impact of the study results.
Introduction

We have documented a massive rise in patient safety research over the last decade (Maillard et al., 2005). Much of this consists of basic research in cognate disciplines, such as psychology, sociology, organisational studies, ergonomics and education. Improving patient safety requires that the knowledge gleaned from basic science should be taken up in interventions to improve care. This article is concerned with the role of science in the development and evaluation of such interventions, rather than with the role of science to understand why errors occur. Evaluation of safety interventions turns on two main issues:

1) Study design (type of study)

2) What to measure as evidence of safety

These two issues are related, since certain measurements may be more or less valid depending on how a study is designed. Nevertheless, we have to start our discussion somewhere. Hence, in this paper, we describe the basic study designs that may be used in research to evaluate safety interventions and develop a theory to support use of different designs in different circumstances. We also discuss measures of safety in more detail with particular emphasis on measurement of error.

Our point of departure,........the nature of patient safety interventions
Safer care can be achieved by replacing unsafe technologies (such as medical devices or surgical techniques) with safer alternatives. Assessment of the efficacy of new treatments (broadly defined) is called Health Technology Assessment (HTA). Care can also be made safer by better or more appropriate use of existing treatments and we focus on such improvements in this paper. Assessment of ways to maximise the delivery of (existing models of) care through improvements in the system is often called
service delivery and organisational research (SDO). We use SDO research rather than health services research because of the similarity of the latter to HTA. Since the patient safety movement is usually conceptualised in terms of seeking to improve the systems within which people labour (Reason, 2000), most patient safety research would seem to fit under the broad heading of SDO research (Shojania et al., 2001).

SDO research is designed to close the gap between efficacy (the theoretical maximal beneficial effect of a therapy) and effectiveness (the effect of the therapy in real life), hence improving the quality of the service provided. This invites the question: is there anything special about patient safety research as opposed to SDO research more generally? In order to shed light on this matter, we start by asking whether there is any conceptual clarity on the putative distinction between efforts to improve safety and those to improve quality generally.

Low and high prevalence incidents: the safety/quality continuum
The provision of high quality care necessarily requires that the care is safe: “first, do no harm” is a common re-phrasing of part of the Hippocratic Oath. Nevertheless, the safety incidents which make the news usually involve dramatic events such as wrong site surgery, lethal dose miscalculation, deaths following inadvertent intravenous potassium chloride concentrate and inadvertent intrathecal administration of vincristine.

The link between error and outcome in these high profile cases is:

1) Immediate or very rapid

2) Certain or highly likely

For these reasons, such errors are sometimes given the sobriquet ‘egregious’: the fact that an error has produced the poor outcome is undisputable.
The cataclysmic nature of these errors means that they have a third feature in common:

3) They are all very rare – in some cases a country the size of England may experience less than one case per year.

Ironically some of the most inherently dangerous activities such as aviation, induction of anaesthesia and blood transfusion are among the safest, precisely because everyone involved is aware of ‘clear and present danger’. We give examples of different types of error on the above dimensions of immediacy and causality (Table 1). As a general rule, very rare errors with high immediacy and causality generate concerns over safety. Errors with low immediacy and causality, such as failure to vaccinate, to wash hands or prescribe beta-blockers after heart attack are much more common and are often conceptualised as quality rather than safety issues. It is against targets such as vaccination rates that the performance of health care providers is assessed: here we could consider performance as a synonym for quality. Could safety be distinguished from quality by how ‘egregious’ (immediate and certain) the link between error and outcome? We believe not. Like Hofer and colleagues (2000), we distinguish a safety/quality continuum based on a vector of ‘egregiousness’ (Figure 1), but like them define no clear point on this continuum at which safety topples over into quality. This discussion illustrates that there is no clear water between safety research and SDO research more generally.

Similarly, while it is also possible to postulate that certain ‘egregious’ errors are examples of errors of commission while many errors of omission have lower immediacy and causality, we agree with Yu and colleagues (2005) that there is no consensus on this definition. The presence of a distinction would make it possible to define safety in terms of errors of commission and quality in terms of errors of omission, but there are
many examples of safety issues that are errors of omission. A good example of a potentially fatal patient safety issue arising from an error of omission is the failure to prescribe or administer prophylactic heparin in a bed-bound surgical patient.

We are thus left with the conclusion that there is no clear cut qualitative research distinction between patient safety and SDO research, just as there is no clear divide between improving safety and quality. As a result, safety science is located across the quality/safety continuum: from evaluating a ‘Clean your Hands’ campaign at one end, to removal of potassium chloride concentrate from wards at the other. It thus follows that there is no clear water between SDO research and safety research. Our analysis therefore applies to both safety research in particular and SDO research in general.

There is one particular challenge for safety research however: the extreme rarity of the most egregious incidents. In this article we lay out a framework for SDO research with special emphasis on rare events. Interventions to improve patient safety do not just appear – they have to be conceived, designed and selected. We therefore start our discussion of evaluation at this pre-implementation phase.

**Evaluation before implementation**

The MRC (MRC, 2000) has specified a framework for the evaluation of complex interventions which starts before the intervention is introduced in practice - a pre-implementation evaluation (PIE). We conceptualise PIE in three broad stages. First, any attempt to improve practice should build on theory; an understanding of why things go wrong and what types of action may ameliorate the problem. Second, an intervention should be designed. Third, the consequences of intervening in a certain way should be modelled. This modelling can involve group discussion (thinking it through, with special reference to possible unintended consequences), formal modelling (with or without probability estimates, value weightings and formal mathematical modelling), simulations.
(mock-ups of the real world and role play) or any combination of these. Results of this third stage can be fed back into the design phase on an iterative basis until an intervention is judged fit for roll out into practice.

For example, development of a method to enhance team working might start with a review of the relevant educational theories: activity theory, social cognition and so on. Further work might include in-depth studies of current practice – for example ethnographic studies of how different health professionals work in teams, qualitative interviews or focus groups – to uncover psychological obstacles to team work. Such studies may show how different professional groups relate to and communicate with each other while undertaking various tasks. A training intervention could be designed in the light of such study. Modelling could begin by asking stakeholders (managers, clinicians, patients) to comment on the proposed intervention in prototype form. If the proposed interventions seemed worthwhile, simulations could be created. These simulations could then be analysed by both qualitative and quantitative methods in order to refine the intervention. In short, theory is used as the basis for a systematic approach to developing an intervention.

The methodology for PIE will resonate with safety experts who also use systematic approaches to detect threats to safety such as prospective hazard analysis and failure modes and effects analysis (Marx and Slonim, 2003; Vincent et al., 1998). Safety interventions are notoriously prone to back-fire, particularly when they are not subject to appropriate risk-assessment; many features put in place after the first Manchester air crash impeded rescue in the second air crash, for example. We will therefore take it as self evident that proposed interventions to improve safety should all be screened through such a systematic process. The screening methods that should be used for different types of intervention is a topic of research currently commissioned by the Patient Safety
Research Portfolio in England (http://pcpoh.bham.ac.uk/publichealth/psrp/), but it is clear that there is no one ‘tool-kit’ suitable for all proposals. Nevertheless, the general principle that PIE should be used to select and refine interventions for roll-out is not controversial. Later in this article we will soften the distinction between PIE and formal evaluation in use, by involving Bayesian reasoning. In the meantime, we turn our attention to the evaluation of interventions when the time comes to implement them in practice.

**Introduction: Summary**

The evaluation of patient safety intervention turns on two key issues: study design and what to measure as evidence of safety.

Our focus is evaluating improvements to existing systems rather than new technologies.

There is no clear distinction between patient safety and SDO (quality) research.

Pre-implementation evaluation of any patient safety intervention is essential to identify possible benefits and costs of the intervention.
Evaluation of Safety Interventions: Study Design

Overview
We concluded earlier that there is no qualitative difference between evaluation of safety interventions and evaluation of any other SDO intervention. The repertoire of methods for primary empirical evaluation of safety interventions must therefore be the same as that for any other type of service development. However our discussion of study design is important given the poor quality of research banded as patient safety evaluation (Ovretveit, 2005). Within the dichotomy of controlled/uncontrolled evaluations used in this section, studies may be retrospective or prospective and cross-sectional or longitudinal. Controlled studies either make comparisons between selected intervention and non-intervention sites (sometimes termed natural experiments or “quasi-experimental” studies (Campbell and Stanley, 1963)) or involve an experimental design where intervention and control sites are chosen at random – randomized controlled trials (RCTs). The range of study designs is summarised in Figure 2. Of course, studies must all measure something – one or more ‘end-points’ must be identified - and we will come to this issue shortly.

In this section we discuss both uncontrolled and controlled studies and consider the need for employing clustered designs. An innovative design, known as a ‘stepped wedge’ trial is considered, since this may be particularly appropriate for evaluating patient safety interventions. We then raise two other methodological difficulties associated with study design: blinding or masking and the ability of the design to allow generalisation beyond study sites. We discuss these difficulties as they arise in patient safety research.
Uncontrolled 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, the NPSA directive to remove potassium chloride concentrate from hospital wards and their correct site surgery initiative were both implemented nationwide for good reason. 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. To avoid the problem of regression to the mean, such studies should involve a large number of sites and, if possible, observations should be taken at a number of points in time, both before and after the intervention – time series analysis.

Uncontrolled before and after studies can also provide a basis for measurement of the extent to which end-points vary both within and between organisations (such variances enable calculation of the intra-cluster correlation which is considered in more detail below). While before and after studies may be the only practical method of evaluation in many cases, they suffer from the weakness that they cannot distinguish cause and effect reliably. This is important if any change observed could plausibly be attributed to other developments in the service apart from the intervention of interest.¹ In such cases, comparative studies between sites exposed to the intervention and controls 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: individual people and ‘clusters’ of people and we therefore digress at this stage to consider this point.

¹ Later in this paper we discuss the possibilities for ‘triangulating’ evidence from different sources in a mixed methods study to help solve the problem of attribution. For example, in an uncontrolled study, effectiveness data could be combined with data from interviews with key staff which might provide an insight into the contribution of the intervention.
The Unit of Comparison

We have argued earlier that service delivery interventions are 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. 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. In other instances, contamination – where an intervention intended for members of the trial arm of a study is received by members of the control arm - may be less than 100%, but still a potential concern. For example, it may be possible to toggle a decision aid facility off and on in a randomised sequence but clinicians may be influenced by previous exposure to the system even when the aid is disabled. An empirical example is how the results of assertive outreach for severe mental illness differ according to the degree of separation between intervention and control sites (Marshall and Lockwood, 1998). Note that the bias due to contamination occurs in one direction only: it dilutes the measured intervention effect and therefore leads to a systematic under-estimate of effectiveness. The extent of the under-estimate depends on the degree of contamination.

Despite the risk of contamination, the unit of comparison can be the individual in some cases. For example, most evaluations of computer-based interventions to improve safety have been based on individual randomisation but have nevertheless shown positive results in most cases, implying either that contamination was not widespread or that the under-estimated effect size was still significant (Lilford et al., 1992; Kaushal et al., 2003). However most safety interventions are so deeply embedded in the system that individual randomisation is not possible and clusters must be used. Clusters are typically different sites (hospitals, general practices etc.) but other types of cluster may also be suitable. Clusters may be comprised of patients treated by doctors exposed to
different interventions: for example, Landrigan and colleagues (2004) randomised clinicians to different types of on-call duty rota.

The problem of contamination does not end at the organisation’s door: it is also possible for contamination to occur across organisations. In a sense, this is what we would like to see happen in society — good practice should disseminate from ‘benchmark’ to other clusters. While this tendency is convivial in a general sense, it is a nuisance in science, where we would like to isolate the intervention of interest from other influences that may be termed ‘nuisance variables’ in this context.

The extent to which contamination really is a problem within service development and educational interventions of various types has been investigated in a study commissioned by the NHS Research and Development Methodology programme (Keogh-Brown et al., 2005). There is a trade-off between loss of precision by using a cluster design and a gain in power due to reduced contamination. As a rule of thumb, when contamination affects more than 30% of control subjects, a cluster design is preferable. Insofar as contamination is of this order or is believed to be a potential problem, there are arguments for cluster studies rather than individual evaluation.

Undertaking an evaluation using clusters raises additional ethical issues discussed by Edwards and colleagues (1999) and also introduces statistical challenges associated with sample size estimation and analysis. These are discussed in some detail by Donner (1998), although we now briefly consider the sample size (number of individuals) needed to show the desired effect with accuracy and precision. The number of clusters required depends on the size of the effect sought and the risk of a false negative study result that can be accepted. It also depends on the extent to which end-points tend to
cluster within organisations; the intra-cluster correlation (ICC). The ICC is also useful to planners since it provides a measure of practice uniformity across a service. A typical comparative study would require 20 clusters and 40 observations in each cluster – such a study would be sufficient to detect a reduction in error rate from 50% to 35% assuming an ICC of 0.01 (α =0.05; β=0.2 and σ²=0.5). Figure 3 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 graph is that the effect of increasing cluster size beyond 40 on the power of the study is negligible compared to increasing the number of clusters (Campbell et al., 2004).

The ICC is defined as the proportion of the total variation in outcomes that can be attributed to differences between clusters. The ICC will be zero if individuals in one cluster are no more likely to have similar outcomes than individuals in the other clusters. The ICC will be one if all individuals within a cluster are identical with respect to outcome. The ICC can be used to calculate the ‘design effect’ for the study:

Design effect = 1+(n-1)ICC

Where: n is the average cluster size

The sample size (number of individuals) in the cluster study can be calculated by multiplying the standard sample size (calculated as if an individual-level RCT was being undertaken) by the design effect. Methods for calculating sample sizes for individual-level studies are detailed in a number of medical statistics texts, such as Altman, 1991.

This revised sample size is then divided by the average cluster size (n) to give the number of clusters required. Hence:

\[ c = \frac{s[1+(n-1)ICC]}{n} \]

Where: c is the number of clusters required
s is the original sample size for an unclustered study

Since the ICC will not be known before the trial is underway, researchers will need to make an estimate to use in the sample size calculation. Estimates are generally taken from the results of similar studies. To this end, The Health Services Research Unit at the University of Aberdeen have produced a database of ICCs calculated across a range of interventions and settings (http://www.abdn.ac.uk/hsru/epp/cluster.shtml), while Ukoumunne et al. (1999) also present examples of ICCs and design effects.
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 (Lilford et al., 1999). Nevertheless, some shining examples of such studies exist. For example, Kenneth Hillman and colleagues (2005) randomised Australian hospitals to have or not have an educational intervention to promote the early recognition and treatment of the deteriorating patient.

Ukoumunne et al. (1999) identify two types of cluster design with two or more data collection points in time: cohort designs, where measures are taken from the same subjects in each cluster at each measurement occasion; and repeated cross-sectional designs, where 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, while in safety/SDO research using cluster (cross-sectional) designs terminal end-points can be measured repeatedly.

**Controlled Comparative Studies**

In addition to determining whether clustering is required, there are two further design variables that distinguish controlled comparative studies:

1) before and after observations versus observations made only after an intervention has been put in place

2) randomised comparisons versus natural experiment

Clearly non-randomised post-intervention comparisons are the least reliable design 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 (Lilford et al., 2004b). Randomised studies with pre and post intervention measures are arguably the strongest design (see Table 2), although non-randomised comparative studies with before and after measurements have particular strengths.

One reason to suppose that comparative before and after studies may yield valid results stems from the notion that differences in rates of change are less often confounded than differences in end-points. This is because the confounding factor would need to affect the propensity to change in response to an intervention, net of any base-line differences. Whether randomised or not, a 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 a proper model one can then estimate both an immediate one-time change within the institution of an intervention along with any increment to the background rate of change (Saint et al., 2003).

Further evidence on the need for randomisation in before and after studies is awaited from a study commissioned by the NHS Research and Development Methodology programme where the results of natural experiments and randomised trials of the same interventions are being compared. In the meantime, the idea that change is 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 a RCT is politically inappropriate or practically unobtainable (Carney et al., 2004; Lilford et al., 1999).

There is reason to believe that the statistical power of a cluster study can be enhanced by making a comparison of the rates of change between intervention and control sites,
which is only possible with a pre-post design. The enhanced power arises because such comparisons do not need to take account of the ICC of the end-points, but of the ICC of propensity to change net of base-line differences. Murray and Blitstein (2003) 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 post-intervention stage. However while databases of empirical evidence of ICCs exist, even where these are focused on studies aimed at changing professional practice (e.g. Health Services Research Unit, University of Aberdeen), ICCs are not given for pre-post differences (with most ICCs calculated using pre-intervention data). Hence the point that comparing changes over time requires the ICC of change is insufficiently appreciated – and did not arise, for example, in a systematic review of cluster studies (Ukoumunne et al., 1999).

**Stepped Wedge Trial Designs**

In a stepped wedge design, an intervention is rolled-out sequentially to the trial participants (either as individuals or clusters of individuals) over a number of (equally spaced) time periods. The order in which the different individuals or clusters receive the intervention is determined at random and, by the end of the random 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. An example of the logistics of a stepped wedge trial design is shown in Figure 4.

Stepped wedge designs bear some similarities to the across subjects multiple baseline or ‘time-lagged control’ design, which is particularly common in evaluations of
behavioural interventions (Barlow and Hersen, 1984). The key differences between multiple baseline and stepped wedge designs are that multiple baseline designs are generally applied to analyse the behaviour of single subjects, with analysis undertaken separately for each individual; and the order in which individuals receive the intervention may not be determined at random, with treatment often delayed until a stable baseline has been achieved (Koehler and Levin, 1998).

Cook and Campbell are possibly the first authors to consider the potential for experimentally staged introduction in a situation when an innovation cannot be delivered concurrently to all units (Cook and Campbell, 1979). The first empirical example of this design (although called a stepped wedge design) being employed is in 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 (Gambia Hepatitis Study Group, 1987).

As noted earlier, 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. 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 (Smith and Morrow, 1996). This may often be the case in evaluations of public health and epidemiological policies such as vaccination, screening and training or, with particular relevance to the current study, patient safety interventions that have undergone careful PIE. Alternatively the design may be useful for undertaking a cost-effectiveness analysis to compare different interventions that have already been found to be effective. In such circumstances it might be considered unethical to exclude any participants from receiving the intervention and non-exclusion
can also be beneficial for recruitment (Hutson and Reid, 2004). Hence the stepped wedge design enables trials to go ahead in circumstances where professional equipoise is absent (Lilford, 1994).

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 groups) at the same time, but when an evaluation of the effectiveness of the intervention is considered desirable. In some situations (particularly when resources are scarce), allocating participants to the intervention randomly would be ethically advantageous.

The stepped wedge design also has scientific advantages. Firstly, where randomisation is undertaken at individual level or where a cohort cluster design is used (as opposed to a repeated cross-section cluster design), the participants in the trial act as their own controls due to the one-way cross-over effect in the trial. Hence participants provide data points in both control and intervention groups. This feature of the stepped wedge design helps guard against the bias arising from unforeseen non-equivalence between randomised groups, although such analysis can only be undertaken on non-terminal end-points. Secondly, even with repeated cross-section cluster designs, the cluster acts as its own control. This implies that sample size calculations do not have to be adjusted to take into account the ICC as is the case with parallel cluster RCTs. Thirdly, 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 (and cluster, if appropriate) are also included, hence controlling for temporal changes not attributable to the intervention (and response differences between clusters).
It is essential to note that stepped wedge designs are not a panacea. In particular, such designs have less statistical power and this loss in power needs to be considered against the ethical and logistical motivations for employing a stepped wedge design. In two subsequent papers, we will report on a literature review of stepped wedge trials (including both randomised and non-randomised studies) and provide a scientific and statistical comparison of parallel and stepped wedge trial designs.

The issue of masking
Like other service delivery and organisational interventions, patient safety interventions cannot easily be ‘blinded’: it is difficult or impossible to mask the participants to the intervention. Masking is important for two reasons.

First, lack of masking of observers may lead to biased assessments of outcome. However it may be possible to guard against this danger in two ways. Observers can be blinded to the end-point, depending of the time frame between care being observed and the end-point occurring. For example, observers may be asked to judge whether care was ‘safe’ or appropriate, without knowing whether the care led to a patient safety incident. Alternatively, observers can 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 site or from the pre or post intervention period. Similarly, Landrigan and colleagues (2004) randomized clinicians to different on-call duty rotas and the observers who measured their error rates were masked as to whether a particular doctor was an intervention or control subject.
Second, lack of masking of care-givers may result in a Hawthorne Effect, where service providers 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 they had been assigned. Knowledge of their rota assignment may (subconsciously) affect clinicians’ behaviour. In such cases, it is important that subjects (in this case the clinicians) do not know what outcomes are being measured during the study, assuming they need to be told that a study is taking place. Lack of masking of care-givers is a risk with many safety interventions – since these are generally educational and behavioural interventions – and in these circumstances randomisation is a less failsafe guard against bias than in typical clinical research where such blinding is often possible (Chalmers, 2003). Hence it cannot automatically be assumed that cluster RCTs are superior to natural experiments with comparative cluster studies involving before and after measurements.

**Generalising beyond study sites**

A distinction is sometimes made between internal and external validity. “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” (Bowling, 2005, p. 398). Similarly a study has internal validity if it avoids bias in populations similar to those studied. It has external validity if the results will hold good if generalised across time and place. Extrapolation to other sites may be undermined because there is a systematic difference between study and non-study sites, in terms of service users, care providers or the way in which services are organised or funded. Such differences arise from the context in which a study is undertaken. 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 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 intervention sites are already very good and have less headroom for further improvement than sites that may benefit from positive study results. Three points can be made about the issue of generalisation:-

1) It applies to any study that may be done, not just quantitative comparative studies.

2) The more consistent results are over time and place, the more confident we can be in generalising — larger studies are better than small, multi-centre than single centre, replicated than un-replicated.

3) 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 and, in the event that such differences exist, some theory as to why and how an intervention may work. This is necessary to inform judgement about how safe it is to extrapolate from one time period or place to another: a solely theoretical evidence based practice is inappropriate whenever context is important. For example, Oakley and colleagues (2006) highlight how the effectiveness of sex education varied according to interactions between the extent to which the education was participative and skills based and who provided the education (peers or teachers).
The extent to which study results can be generalised across time needs to be considered since time, as a variable, can sometimes affect study results. For example, a systematic review finds family therapy effective in reducing admissions in people with severe mental illness. However, if the studies are arranged chronologically, the results become gradually less impressive, until they disappear altogether in the most recent study (Marshall and Lockwood, 1998). Knowing more about how family therapy is likely to work – by improving regular medication usage – allows us to develop a plausible theory to explain these temporal effects: as base-line services have improved, there is less headroom for family-mediated effects to show significant changes.

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 consider further the value of incorporating additional sources of data later in this paper.
Study Design: Summary

Possible study designs for patient safety research are the same as for other types of evaluative research: uncontrolled studies with before and after measurements, controlled studies with post-intervention data collection and controlled studies with before and after measurements.

Measurements of changes occurring in response to an intervention are less subject to bias than static measurements. Hence before and after studies:

- May reduce the need for randomisation;
- Reduce sample size requirements in cluster studies if the ICC is based on the effect size; and
- Imply that evaluation should be planned prior to implementation.

Stepped wedge study designs may be appropriate when an intervention is likely to do more good than harm or where classic randomisation of individuals or clusters is impractical or unethical.

Participants, care-givers and observers should be masked/blinded as much as possible to minimise bias and Hawthorne Effects.

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 sub-group effects.
End-points and Measurement

Overview

To determine whether or not a patient safety intervention has been effective, measurements of the attributes or characteristics of key objects, events or abstract constructs must be undertaken. The process of measurement raises two inter-related methodological questions: about what end-point(s) information should be collected and how this information should be collected? This section discusses the methodological consequences of decisions regarding these what and how questions.

To help conceptualise between different end-points, we distinguish between generic management interventions (such as changing skill mix or improved human resource policies) that may strengthen an institution overall by reducing ‘latent errors’ and specific interventions targeted at a particular safety threat or ‘active error’ (such as a delay in operating on people with fractured neck of femur, reducing the risk of falls or avoiding incompatible blood transfusions). This distinction between latent and active errors is discussed by Reason (2000) as part of his Swiss cheese model of system accidents. Clearly the wider the intervention’s net is cast the greater the number of individual safety practices that can be affected, but we may expect lots of things to be improved a little, rather than a few things to be improved a lot. Information technology occupies a somewhat intermediate position, because it allows many specific practices to be targeted by specific algorithms nested in a diffuse enabling technology.

Figure 5 provides a conceptual map of the plausible end-points (in the shaded cells) associated with both generic and specific patient safety interventions. The chain of care runs from left to right: from the structure of the organisation to patient outcomes, based on Donabedian’s famous model (1980). We separate management from clinical...
processes, since generic interventions are focused at management processes and specific interventions at clinical processes. Our discussion of different end-points and associated measurement processes starts at the right-hand side of the diagram, with patient outcomes. We then switch to the other end of the diagram to consider end-points at the beginning of the care chain before discussing clinical process measurements. Collectively, end-points occurring prior to the final patient outcome are known as surrogate end-points or intermediate outcomes. Given the amount of available literature, particular attention is paid to clinical process measures. For similar reasons, we also focus on case-note review as a method of data collection.

**Outcomes for patients**
The intention of any safety intervention is to improve outcomes for patients by reducing harm. Patient outcomes can either be clinical (morbidity or mortality) or patient derived (quality of life or patient satisfaction). We focus on clinical outcomes in this paper: measurement issues for both quality of life and patient satisfaction are discussed in detail in HTA reviews (Fitzpatrick et al., 1998; Crow et al., 2002). We report on three difficulties associated with using patient outcomes as an end-point: 1) separating the signal of improvements resulting from an intervention from the noise: stochastic variance in outcome, 2) ensuring a reliable definition of an adverse end-point (i.e. the numerator used in comparing adverse event *rates*) and 3) ensuring a fair denominator across sites, so as to avoid bias.

*Signal to Noise Ratio*

A common problem for evaluators is that the patient outcomes that may plausibly be affected by an intervention are also influenced by many other factors. For example, a computer programme may be designed to reduce harm from medication errors. Such a
programme could be very effective in reducing harms of various sorts, but an evaluative study may fail to detect true benefit because the signal (improved outcome from less medication errors) is lost in the noise (variance in outcomes due to a great many other factors – both treatment and disease related). The causes of poor outcomes are illustrated in Figure 6, with errors forming just one component of the model.

The model illustrates that using patient health outcomes as the end-point in an evaluation of a patient safety intervention will result in a high risk of a beta (or Type II) statistical error – the error of accepting the null hypothesis when it is not true. In short, adverse patient outcomes would be an imprecise end-point. As an example of the imprecision of measurements of outcome as a reflection of quality, Mant and Hicks (1995) model the contribution of adherence to care standards and other factors to documented differences in heart attack survival across hospitals. Even if all appropriate care standards were followed in some sites and none in others, this could account for only half of the observed differences in mortality between hospitals. Since such large differences in safety practice are implausible, very large sample sizes would be needed to show realistic differences in safety in an evaluative/comparative study.

The risk of losing the signal in noise could be reduced by enumerating only those cases of poor outcome caused by poor care (error). This requires deeper (and more expensive) examination of the process of care so as to select instances when the poor outcome was the result of deficiencies in care. Such enrichment has three limitations.

Firstly, judgement about the outcomes that could be prevented is fallible; it is difficult if not impossible to determine whether a particular case of wound infection was preventable, for example. Different reviewers may therefore disagree: for example in the Harvard Medical Practice Study where reviewers were asked to judge whether an
adverse event was due to negligence, a kappa of just 0.24 was achieved (Brennan et al., 1991). We return to the subject of inter-rater agreement later in this section.

Second, judgements may be affected by hindsight bias, where reviewers make different decisions in a review of care than they would have made during a real-time observation of the care. Generally, there is a tendency for reviewers to judge the events leading up to a patient safety incident as errors if it is known that there was an adverse outcome: the outcome is considered more foreseeable and therefore more preventable than the reviewer would have appreciated in real time (AHRQ glossary: http://psnet.ahrq.gov/glossary/aspx). This theory has been tested empirically by Caplan and colleagues (1991), who found that the severity of outcome influenced reviewers’ judgement of the appropriateness of care. We argued earlier that reviewers should be ‘blinded’ to patient outcomes in order to prevent this hindsight bias, although obviously this would not be possible if contingent harm is the end-point for the study.

Third, enrichment may not overcome the problem of low statistical power because the combined event (poor clinical practice followed by poor outcome) may be very rare. Examples of such rare events are wrong site surgery, inadvertent administration of potassium chloride concentrate or intrathecal vincristine injections. The severe limitations of using ‘preventable’ adverse events as a marker for the quality of care have been discussed in detail by Hayward and Hofer (2001).

**Consistent Measurement of Outcomes**

In addition to their rarity, the outcomes themselves may be difficult to measure in a consistent way. Even when mortality is the end-point, there may be ambiguity over whether a particular death should be attributed to a particular analytical unit. For example, a cardiac patient dying following failed resuscitation in the accident and
emergency department may or may not be included as a hospital death if the heart attack occurred prior to admission. Consistency is more of a problem for morbidity. For example, Bruce and colleagues (2001) undertook a systematic review which identified over 40 definitions of two surgical adverse events (wound infection and anastomotic leak). They document considerable variations, not only in definitions, but also in the implementation of the same definition across sites. Of greatest scientific importance is the difference in the way the same definitions are applied across sites. Definitional ‘gaming’ is one possible consequence if performance indicators and ratings are based on clinical end-points (Bird et al., 2005). In the context of research, it is important to ensure that the same observers make measurements across sites, since otherwise it is almost inevitable that the results will be biased due to observational differences across observers, even when they are supposed to be using the same definitions.

**Biased Denominators**

Outcomes may also be affected by case-mix, which may create biases across denominators in comparative studies, although the bias can be minimised by randomisation. In non-randomised studies, it is sometimes argued that statistical adjustment for case-mix can overcome this bias. While agreeing that such adjustment is advisable in non-experimental studies, it cannot be relied upon to eliminate all bias. Patients who are sicker and/or older have more co-morbidities and are at increased risk not only of worse outcomes, but of experiencing more errors, leading to bias in comparative studies. Different methods of case-mix adjustment give different results (Lilford et al., 2004a) and it is only possible to adjust for potential sources of case-mix bias when information about potential confounding variables has been (accurately) collected. Differences in end-points across intervention and control sites are the net result of variances in definitions, measurement and case-mix as well as of quality. It is
inappropriate to assign the residual after case-mix adjustment to quality differences alone, since unadjusted differences in case-mix may remain: the case-mix adjustment fallacy (Lilford et al., 2003).

**Surrogate End-points**

As an alternative to assessing patient outcomes, we now focus on possibilities for measurement occurring at the process level in the systemic pathway shown in Figure 5. Oakley and colleagues (2006) have recently argued that including a process evaluation alongside a more traditional outcomes-based approach would improve the science of many RCTs, particularly of those evaluating complex interventions or in cluster trials, where the intervention is non-standardised. We consider three levels of surrogate end-point, the fidelity with which an intervention is implemented, the effect of the intervention on an intervening variable such as morale and the effect of the intervention on clinical error rates. These different surrogates reflect potential 'holes' in Reason’s Swiss cheese accident model (2000), moving ever closer to the patient outcome.

**Fidelity**

In evaluating a new service level intervention targeted at either managerial or clinical processes, a possible surrogate end-point is the fidelity with which the required action is implemented, rather than the consequences arising from non-implementation. Where the benefit of the intervention is not in doubt, measuring compliance ‘stands in for’ patient outcomes. For example, a forcing function to prevent mis-connection of tubes delivering oxygen and other gases may be regarded as an effective service level intervention. In such circumstances, it might be entirely satisfactory to measure hospital compliance with directives to deploy the intervention. In the more usual situation, where the benefits of a service level intervention are contested, demonstrating high fidelity at least shows that the distal benefits (in terms of patient outcomes) are plausible. If the
proximal measures have not even been put in place, then the distal improvements are implausible. Fidelity is therefore often a necessary, but not necessarily a sufficient, condition to prove an intervention has improved patient outcomes. Evaluation can make use of this asymmetry in two contexts:

1) The context where direct measurements of safety (terminal errors and/or outcome) can be made within high precision (and accuracy). In this case the fidelity measure may help explain a null result.

2) The context where direct measurements are impossible/unaffordable at reasonable precision, in which case showing that the intervention was instituted on the ground at least provides reassurance that the putative benefits are possible.

For example, the NPSA Wrong Site Surgery Patient Alert consists of a series of measures to prevent surgery on the wrong part of the body. Showing that the prescribed measures had in fact been put in place would not prove, beyond all reasonable doubt, that the intervention was effective at the patient level, but the reverse finding – continuing violation of the tenets of correct site surgery - would carry the implication that little patient benefit is likely be forthcoming. The evaluation of the Wrong Site Surgery Patient Alert funded by the Patient Safety Research Portfolio being undertaken by John Wright and colleagues at Bradford Royal Infirmary goes beyond assessing compliance, using qualitative data to describe factors that hindered or helped implementation (http://www.pcpoh.bham.ac.uk/publichealth/psrp/). Such observations may buttress fidelity measurements by providing further evidence for or against the plausibility of a beneficial outcome: we return to the idea of using multiple end-points in the final section of this paper.
Intervening Variables
We return to the distinction made earlier between safety interventions targeted at a specific patient safety threat and more diffuse interventions aimed at strengthening an organisation in a more generic sense. The importance of organisational characteristics in determining the quality of care is highlighted by Wagner (2001) in his model of care for chronic illnesses, yet there is insufficient evidence about the organisational factors that lead to more effective care (Shortell et al., 2000). Diffuse interventions are more difficult to study since many aspects of patient care may be expected to change a little, so that they cannot be measured with sufficient statistical precision to avoid a Type II error. As one solution to this problem, we introduce the concept of an intermediate or intervening variable which can be used as a surrogate end-point in evaluations of diffuse interventions. For example, improved human resources (HR) policies (such as staff appraisal) are expected to impact on errors by means of effects on staff motivation and morale and reduced sickness. It may thus be informative to measure the effects of such HR policies on these intervening variables. The NHS Research Methods programme is currently sponsoring research into the reliability and validity of measurement of ‘intervening’ variables in human resource management (http://www.pcpoh.bham.ac.uk/publichealth/nccrm/).

Patient Safety Culture
One particularly interesting intervening variable is the extent to which an organisation demonstrates a patient safety culture: the NPSA’s Seven steps to patient safety (2004) includes “Build a Safety Culture” as Step 1. The NPSA define a safety culture whereby “staff have a constant and active awareness of the potential for things to go wrong. It is also a culture that is open and fair and one that encourages people to speak up about mistakes. In organisations with a safety culture people are able to learn about what is
going wrong and then put things right” (p. 18). The NPSA provide examples of assessment tools that could be used to measure safety culture over time. While the metric properties of these tools (reproducibility and sensitivity) have been measured, their construct validity has not been confirmed, as noted in a recent review by Scott and colleagues (2003). These authors therefore note that the research on assessment tools is inadequate. In the meantime Cooper (2000) argues that safety culture is made up of many dimensions that probably do not change in harmony (one dimension may change leaving the others unaffected). Furthermore, different dimensions of the ‘safety culture’ may impact on different clinical practices and hence different end-points. It is certainly true that improvements in one aspect of care do not always correlate well with improvements in other aspects (Jha et al., 2005; Wilson et al., 2002). There are therefore good arguments for measuring the dimensions of the patient safety culture independently rather than relying on a single ‘global’ measure. The NHS Research Methodology Programme will be sponsoring a research programme to measure the construct validity of culture tools.

Clinical Process Measurements: Error
Our final surrogate end-point is clinical error, which we discuss in particular detail. Clinical error, which causes or may cause the adverse outcome, is the closest surrogate to clinical outcome itself. Error can be defined as the failure to apply the correct standard of care. Sometimes the error and the outcome are one and the same – wrong site surgery for example. In such cases the error and the poor outcome occur with the same (or almost the same) frequency and so any problem of rarity is not ameliorated by measuring error. In most cases however, the error is more common than the corresponding adverse event, because many errors, even if carried through to the patient, will not turn out to be harmful. Medication error is an example. Serious medication errors are rather uncommon (Dean et al., 2002; Taxis and Barber, 2003), but
they are very much more common than contingent instances of iatrogenic harm. Where errors are more common than adverse events, their relative frequency affords greater precision in measurement. However, measurement of error is not a panacea and in this section we discuss some of the issues that should be considered in the measurement of error. We begin with enumeration of error alone and progress to the measurement of error rates. We consider case-note review in some detail, although the ideas relating to type of data extracted, validity and reliability can be generalised across methods.

Denominator-free error detection

As a method of error detection, reporting systems (both statutory and voluntary) are denominator free. There are two major problems associated with reporting systems:

1. It has been shown that reporting is highly selective and very incomplete (NAO, 2005).

2. Any change in rate may reflect propensity to report rather than a true change in the underlying rate. Indeed, as an institution improves in the care it delivers, more problems may be reported, since open reporting is a tenet of safe practice.

This does not mean that reported rates are of no scientific value whatsoever. While moderate differences in reported error are not reliable or valid measures of performance, massive proportional changes may be indicative of improvement, especially in the case of some of the most egregious errors such as intrathecal injections of neurotoxic pharmaceuticals, which are hard to ‘hide’. Furthermore, national-level data from reporting systems are useful for identifying priorities for patient safety interventions. Following the introduction of the National Reporting and Learning System (NRLS), the Patient Safety Observatory at the NPSA received 85,342 reports of incidents and near misses from 230 Trusts between November 2003 and March 2005, which are
documented in their first report, *Building a memory: Preventing harm, reducing risks and improving patient safety*, published in July 2005. By the end of January 2006, the total number of reports had risen to over 400,000, suggesting that a habit of reporting via NRLS is beginning to develop.

**Trigger Tools**

“A trigger tool is a collection of indicators such as abnormal laboratory values and drugs that may be prescribed as antidotes, used to trigger more extensive investigation into whether medication-related harm has occurred” (Barber et al., 2005). Like reporting systems, trigger tools can be denominator-free, but will have a denominator if the number of patients to which the tool has been applied is known. While Rozich et al. (2003) report how a manual trigger tool identified a useful proportion of medication errors in the US, an adaptation of the method identified less than 1% of prescribing errors in a UK pilot study in secondary care, although it did identify 50% of those errors that resulted in harm (Barber et al., 2005). Not only do trigger tools have low sensitivity, but specificity is also low; Barber and colleagues report that just 1.6% of positive triggers were true positives. The use of such tools may mean that investigations into false positives may not be cost-effective for researchers. Most importantly, from the research point of view, different trigger tools may yield highly biased comparative data across organizations, since they will identify different errors with different sensitivities, depending on the database concerned and the precise details of algorithms used to interrogate the databases. It could be argued that trigger tools may be more effective in primary care, due to the availability of rather more uniform IT systems that include more detailed information on medication and other aspects of care such as investigation results. However, as we approach the full electronic record, the less we need rely on particular types of numerator and the more we can attempt a full and consistent
enumeration of errors as a proportion of cases where the error might have occurred. This takes us to our next topic, measurement of error rates.

**Error Rates**

Error rates require measurement of both numerator and denominator data. Errors are often denominated at the level of the patient (i.e. errors per patient). However error rates enumerated on a per patient basis may yield biased results. Two important sources of bias must be considered. Firstly, case-mix bias affects comparisons based on error rates as it does comparisons based on outcomes. This is because different case mixes may yield different opportunities for error. A partial way out of this conundrum is to denominate errors on the opportunity for error, rather than on the number of patients, as suggested by the first author elsewhere (Lilford et al, 2003). Using these opportunities for error provides a method to reduce, but not eliminate, bias due to systematic differences in case-mix across comparators. However, case-mix bias is not likely to be completely avoided for two key reasons.

First, it may be easier to detect opportunities for error in one place than in another, for example due to differences in note-keeping. This applies particularly to pre-recorded materials, such as case notes. For example, consider two hospitals, one of which keeps exemplary case notes and the other in which note keeping is perfunctory. Clearly, the observed opportunity for error will be greater in the former case and this may mask differences between institutions, or worse, result in more adverse measurements for better performing sites. Even if the case notes were exactly the same and completed with equal diligence, bias could still occur if different observers made observations in different hospitals. If this difference between observers was random, then this would widen the variance, diluting any findings and increasing the risk of a Type I error. If the difference was non-random, the effects would be worse, since the results would be liable
to inaccuracy as well as imprecision. Using the same observers across all sites reduces random error and masking observers to the origin of the notes reduces the risk of observer bias, as discussed above.

Second, equally performing clinicians may simply find some errors more difficult to avoid than others and this confounds comparisons when the opportunities for error differ from place to place. That said, we believe that the use of opportunities for error provides at least some protection against case-mix bias. Moreover, it provides an elegant method to deal with contingent errors, where the opportunity for error arises only if certain pre-existing conditions are fulfilled — for example, giving insulin therapy in intensive care when the glucose level crosses a certain threshold. It is necessary to make an *a priori* error (failure to make glucose measurements) in order to expose the risk of contingent error (failure to institute insulin therapy when glucose exceeds a certain threshold). Clearly, it would be very biasing to make comparisons of a contingent error, without first allowing for the primary error. Opportunity for error methodology provides a solution to this problem as we demonstrate elsewhere (Barratt et al., 2005).

Methods for the detection of error

There are four main methods for the direct detection of error rates (where both numerator and denominator information are collected from the same source):

1) Case-note review - retrospective

2) Prospective data collection by clinical staff

3) Prospective data collection by independent observers

4) Prospective data collection by a participant observer – simulated or real patients
By a long margin, the greatest amount of methodological research in patient safety has been undertaken using case-note review and we therefore focus on this method. There is over ten years of evidence of case-note review in the literature, starting with a study evaluating changes in the quality of care in the US following a change in Medicare’s payment process (Kahn et al., 1990a). The case-note review literature is generally focused on the quality of care, including, but not specific to, patient safety. There is evidence to suggest that the four different methods above will not result in identification of the same set of errors. Michel et al. (2004) show that prospective data collection by clinical staff produced a higher error rate than retrospective evaluation of case-notes in the context of accident and emergency care. Divergence between prospective pharmacist review and retrospective case-note review for identifying prescribing errors is also identified by Barber and colleagues (2005).

The relative strengths and weaknesses of these (and other) methods of data collection are reviewed by Thomas and Petersen (2003), Murff et al. (2003) with a focus on IT detection systems and the NPSA (2005) with a focus on UK-systems. Thomas and Petersen suggest that the relative utility of different methods depends of the type of error or incident being investigated. Direct observation requires skilled observers and is expensive although there are many errors such as those arising from drug administration, which are only assessable by direct observation. Participant observation by patients is more reliable and valid for detection of failures in communication and empathy, than for the technical aspects of care (Safran et al., 2006). Observation by trained observers or experts simulating patients can be used to describe both the technical and humane aspects of care, but are applicable only to non-acute settings, for obvious reasons. While the use of ‘mystery shoppers’ is now common within retail
settings, most people think that, ethically, clinicians should be forewarned that some of their patients will be providing assessments of their care.

Implicit Versus Explicit Review

Data on error (whether in case notes or observations) can be collected using implicit and explicit methods. The implicit method involves experts making their own judgements about the quality of care provided and can be either structured or unstructured. In an implicit structured review, experts are presented with a series of preparatory questions designed to elicit a complete review of the important facets of care. Such reviews are generally used for research purposes and to establish a quality level for groups of patients or providers. In contrast, experts are given little guidance for unstructured reviews, which are more often used for credentialing, administrative or litigation purposes. Unstructured reviews are focused on outliers and as would be expected, there is more emphasis on the narrative since there are no summary quantitative scores. A structured review may however also include an unstructured element. For example, reviewers may be asked to rate the overall quality of care. One approach to this type of summary question is to ask whether the expert would take their mother for treatment to the organisation in question. A combination of structured and unstructured methods used in the Kahn Medicare pre-post study is a seminal approach to implicit review (Rubenstein et al., 1990).

Explicit reviews involve the objective application of pre-determined standards, for example to facets of care recorded in the case-notes. The standards are developed using expert groups and/or national care protocols. Explicit reviews can be focused (using a limited set of supported and feasible measures) or global (using a broader set of quality measures for a large number of conditions). Jha and colleagues (2005) apply a focused approach to assess care in 3,558 US hospitals. Data were collected on 10
indicators of the quality of care across three conditions: myocardial infarction, congestive heart failure and pneumonia. The quality of care in the US has been studied using a global approach, applying 439 indicators for 30 conditions and preventative care (McGlynn et al., 2003). These global indicators have been adapted for use in the UK by Kirk and colleagues (2003), who use a limited set of 200 indicators across 23 conditions and preventative care. More recently, the American Medical Association has agreed to requests from the US government to develop by consensus 140 evidence-based measures of doctor performance in 34 clinical areas (Hopkins Tanne, 2006). However the global approach is highly resource intensive which may pose logistical constraints on studies of this nature (Kirk et al., 2003).

Criteria or indicators are applied in an explicit review if they are considered relevant to the patient in question: for example if the patient was likely to benefit from the process being considered (Khan et al., 1990b). Such conditionality means that a different number of criteria will apply to each patient. Where prevalence of some criteria is low, measurements will be imprecise unless very large (expensive) samples are used (Kirk et al., 2003).

One important methodological consideration is the analytical approach to multiple criteria and this is considered in some detail by Kirk and colleagues (2003). Different criteria can be analysed separately, but a more common approach is to combine the criteria. This could simply mean summing the number of criteria judged to have been met or applying a method of weighting. These weightings could be based on judgement of clinical importance, perceived strength of empirical evidence or the extent to which they are shown to discriminate between good and bad organisations as judged by other criteria. Not surprisingly, different methods of combining the same data produce very different rankings (Jacobs et al., 2005). While this might affect research conclusions, the
key in comparative studies is to use the same method across the units of comparison and to ensure that whatever method is used is pre-determined rather than data driven.

Both implicit and explicit methods of data collection have advantages and disadvantages, which are compared in Table 4. The Patient Safety Research Portfolio is currently funding a study comparing implicit and explicit methods using approximately 3,000 case-notes in patients with COPD and heart failure (http://pcpoh.bham.ac.uk/publichealth/psrp/index.htm).

The increasing use of patient electronic records will afford greater opportunities for the detection of error. The records contain coded information (typically vital signs, drugs given, allergies, test results and disease codes) and un-coded information (free text). Clearly this should facilitate the detection of some errors which will be measurable in an unbiased and precise way under program control, but all the other caveats about error detection will continue to apply to the free text component of the notes, which is where much of the clinical richness lies.

Validity

Any method of data collection needs to demonstrate construct validity: the results need to be an accurate reflection of the underlying concept intended for the data. A surrogate end-point has construct validity if it has been shown empirically to correlate with outcome, or if this is judged to be self-evident. Thus, if a particular clinical action is associated with worse outcome, then failure to implement the action is an error which has high validity (Lilford et al., 2004b). For example, failure to administer anti-D prophylaxis to a rhesus negative woman after birth of a rhesus positive baby is an error which is a good surrogate for harm: sensitisation of the mother.
However many actions (or failure to act) may be more ambiguous and hence their construct validity is less clear cut. For example, it is commonly believed in patient safety circles that minor prescription errors are a good surrogate for major prescription errors. It would be convenient if this was so, since the relative frequency of minor errors means that measurements can be more precise, increasing the power of statistical conclusions. However for this to hold good there must be a correlation between minor and major errors. Heinrich’s seminal work in 1959 outlined a ratio in which there were 29 minor injuries and 300 no-injury accidents for every major injury (Heinrich, 1959). However the empirical basis for this conclusion is very poorly described and a review of studies on prescribing error found insufficient evidence to support a single ratio that could be used to validate the use of minor errors as a surrogate for major errors (Dean et al., 2002; Barber et al., 2005). Intuitively, we would expect (any) relationship between minor and major errors to be context dependent and hence we think that this is an important topic for future enquiry. In the meantime we think that minor errors are only a surrogate for major errors if they are clearly on the direct causative pathway. Failure to test blood group is on the causal path to failure to give prophylaxis which is on the causal path to Rhesus sensitisation. However a minor dosage error is not on the direct causative path to a potentially fatal drug interaction error. They are indirectly linked through the idea that both are a sign of lack of vigilance or some general failing: the idea behind Heinrich’s ratio. While this remains unproven we urge caution in concluding that real improvement has occurred when indirectly linked minor errors have been reduced.

Reliability

Any measure of patient safety needs to be reliable (repeatable). Inter-observer reliability implies that two (or more) observers/reviewers would come to the same conclusion when evaluating a single care process. It is therefore important to assess inter-rater
reliability in scientific studies where end-points involve an element of judgement, as is the case for error rates. The traditional method of assessing the extent of inter-rater reliability between independent reviewers is Cohen’s Kappa (Cohen, 1960), although the sensitivity of kappa to the prevalence of error (Edwards et al., 2005) has resulted in other methods, such as tetrachoric or polychoric correlations, being advocated where possible (Hutchinson, 1993). Furthermore, when the prevalence of the phenomenon in question (error rate) becomes more extreme, the possibility of agreement above chance becomes small and it is very difficult to achieve even moderate values of Kappa. In such instances, the Phi statistic can be used to calculate chance-independent agreement (Guyatt and Ciliska 2003). Since duplicating observations is expensive, it is important to select a sample size for the measurement of reliability that will be statistically precise. Calculations for sample size can be derived from formulae given by Altman (1991) or other statistics texts. Table 5 provides examples of the sample sizes required to produce a Kappa of greater than 0.6 (which Altman (1991) considers ‘good’ agreement) for given proportions of actual agreements and expected agreements, with a statistical significance level of p<0.05.

Edwards and colleagues (2005) build on a previous review of inter-rater reliability of case-note review by Goldman (1992). A total of 66 sub-studies (taken from 25 original papers) were included in the review, with kappas ranging from 0.10-0.83. Edwards and colleagues consider factors that may effect the kappa obtained and report that the use of explicit, rather than implicit criteria for review raises mean kappa from 0.39 to 0.62; assessing outcomes, rather than outcomes-from-process or processes increases kappa (mean kappas were 0.59, 0.41 and 0.37 respectively); and that, as noted earlier, kappa is positively correlated with the prevalence of error (r=0.44, p<0.01).
Reliability in implicit reviews can be enhanced through reviewer training. However it is important to consider the need to train reviewers, rather than homogenise them. For example, Rubenstein et al. (1990) used reviewer training to encourage use of a uniform set of rating terms rather than to change opinions about what should have been done. Alternatively, discussion between reviewers can be used to share knowledge and (theoretically) increase reliability. However when Hofer and colleagues (2000) compared the effects of discussing case-notes on inter-rater reliability they found more agreement within, but not between, pairs of reviewers.

A discussion of a study within our group (Barratt et al., 2005) can be used to illustrate reasons for relatively poor inter-rater reliability. The authors examined care for severe sepsis and septic shock in the Intensive Care Unit. Two key factors make agreement on key end-points problematic:

- There are difficulties associated with identifying the exact time of onset of the conditions (in this case severe sepsis or septicaemia) that bring the standards into play. Yet the clinical criteria are time dependent (treatment must be given within a certain time from diagnosis).

- Complications and positive test results arise in some people and not others and some standards are contingent on these complications/positive test results.

We have already highlighted evidence suggesting that different errors or decisions based on preventability/negligence are identified when different reviewers extract data from the same chart or when different methods of data collection are used (e.g. Brennan et al., 1991; Michel et al., 2004; Barber et al., 2005). One approach to this conundrum is to sum every error regardless of whether it was identified or considered an error by every source. This begs the question of how many reviewers or methods of data
collection are necessary before the number of errors identified becomes sufficiently asymptotic to the actual number of errors to make additional sources cost-ineffective. A further problem is that the distribution of errors detected by different reviewers is positively skewed, so that the results are over-influenced by ‘hawks’ who find the greatest number and the most doubtful errors (Haywood and Hofer, 2001).

**Surrogate End-Points: Summary**
While we have focused our discussion in this section on the use of case-note review to measure error, much of this discussion would apply to any method of data collection and to the full range of possible end-points. Since no method or end-point is infallible, we consider the possibilities for triangulation of research methods and/or end-points in some detail below.

We illustrated in Figures 4 and 6 that poor outcomes for patients may or not be the result of clinical error, which in turn may or may not be the result of management-level, or *contextual*, errors or failures. However, the relevance of such management-level error is often a matter of dispute because contextual errors are often associated with failure to implement protocols and procedures whose value in improving patient safety (i.e. their validity as a surrogate end-point) may initially be contested. Contextual errors may therefore be the subject of an investigation as the explanatory, not the outcome variable in the statistical model. However, once the value of the protocol or procedure is established beyond reasonable doubt, then its status can change from an explanatory to an outcome variable and it can also be used as a performance monitoring tool. Arguably, computer decision support, staff appraisal and having a clinical pharmacist available at all times have made much of the transition from research to performance management, but valid questions are still asked about how, when, where. In other
words, such standards occupy something of a middle ground between topics for research and standards for implementation.

**Collateral Damage**

Safety interventions, like clinical interventions, often have side-effects. In promulgating a safety intervention the hope is that one will substitute a less serious for a more serious set of problems. The effects of an intervention may lie outside the immediate target. An intervention to reduce delay in treatment for one set of patients may increase delay for another set, for example. It is therefore important to think through the possible collateral implications of safety interventions – indeed pre-implementation evaluation and prospective hazards analysis are designed, in part, to discern what these collateral costs might be. The corollary of this thinking is that when safety interventions are evaluated, the range of measures should be as wide as possible to include all the knock-on effects which should have been unmasked by the PIE.

In Table 3 we provide some examples of different kinds of patient safety intervention and the kinds of end-points that may be studied in connection with each.
End-points and Measurement: Summary

We make a distinction between generic interventions to reduce latent errors in the system and specific interventions to reduce active errors in care.

There are problems with using patient outcomes as an end-point:
  The signal is often lost in noise;
  It is difficult to apply a reliable definition of an adverse end-point; and
  Denominators used in the calculation of event rates are often biased.

Instead of enumerating patient outcomes, it is possible to use surrogate end-points of fidelity, intervening variables and clinical errors, although such end-points are not panaceas.

Case-note review is the most frequently applied method of error detection, although different methods result in different errors being identified. Further research in this area is underway to improve error detection methodology.

Researchers need to ensure the validity and reliability of any method of data collection and above all consistency and masking are important for avoiding bias in comparative studies.
One size does not fit all

Overview
We propose that selection of study design should be influenced by four considerations:

1. Logistic/pragmatic constraints
2. A priori assessment of benefit and harm: the cost of making the wrong decision
3. The possibilities and consequences for improvement
4. The target audience for the results: who is the study intended to influence?

We discuss each of these considerations in turn.

Logistic/pragmatic constraints
It is immediately apparent that comparative quantitative designs are feasible for the evaluation of interventions aimed at common, rather than rare incidents. It is simply logistically very difficult to organise such studies to evaluate the effects of interventions to prevent events such as inadvertent intrathecal injections of vincristine, which occur only once every few years. Hence, RCTs and other prospective designs are eminently feasible at the high frequency end of our spectrum; techniques to improve adherence with guidelines, for example, have been extensively evaluated by this method (Gordon et al., 2003; Grimshaw and Russell, 1993; Lilford et al., 1992).

A Priori assessment of benefit and harm
It is also important to consider the purpose underlying the research when designing evaluative studies. Specifically, this implies the extent of uncertainty surrounding the outcomes or consequences arising from an intervention, both clinical and financial. If it is anticipated, following careful PIE, that an intervention is unlikely to do no material harm then an evaluation across contemporaneous control and experimental groups may
be a counsel of perfection. A study of the fidelity of uptake followed by a before and after comparison of error rates may be sufficient to quantify effectiveness and identify factors that may impact on effectiveness. Alternatively, a stepped wedge design may be ethically or logistically desirable. For interventions where the expected benefit is unclear or where potential harms may reduce net benefit, a more robust evaluation involving intervention and control groups will be necessary.

If there are likely to be financial consequences arising from the intervention or if two or more interventions are being compared, then an economic evaluation may be required. Methods for the economic evaluation of health care are detailed in Drummond (1997). Such economic evidence could be used to justify capital expenditure on an intervention, given the savings that could be made in the medium and long term. Ovretveit (2004) provides examples of how research can be used to predict savings from interventions to improve the quality and safety of health care. We will be addressing the economic perspective on the evaluation of the relationship between science and decisions in a subsequent article. However cost-effectiveness calculations can only be built around an estimate of effectiveness and it is with obtaining an accurate estimate that this article is concerned.

The potential effectiveness of an intervention (preferably one that has undergone PIE) can be captured explicitly and quantitatively in a Bayesian prior (Lilford and Braunholtz, 1996). Such a Bayesian prior can be used to model the potential value of an intervention using expected value theory/decision analysis (Lilford et al., 1998). The Bayesian method provides not only an opportunity to combine data of the same type (in a meta-analysis) but also to integrate different types of evidence when head to head comparisons are not available or when they are not conclusive (Lilford and Braunholtz, 1996). For example, Roberts and colleagues (2002) undertake a Bayesian synthesis of
An epistemology of patient safety research

qualitative and quantitative evidence to identify factors affecting the uptake of childhood immunisations. We have argued that Bayesian methods are particularly useful in the world of policy and service delivery where conclusive, generalised, comparative studies are hard to come by (Lilford and Braunholtz, 2003).

The possibilities and consequences for improvement

When we consider the costs of nationwide interventions to tackle rare incidents we are typically looking for large effect sizes in relative terms. For example, a proposal to reduce by 25% the one or two cases of death from inadvertent bolus injection of potassium chloride concentrate that previously occurred each year in England would seem rather nugatory and in population health terms. Indeed, the National Patient Safety Agency (NPSA) has directed that the concentrated form of potassium chloride will no longer be available on the wards and this should all but eliminate the problem. Hence, not only are contemporaneous comparative studies less feasible (they would inevitably be under powered as argued above) there is also less need for them. A simple study of fidelity of uptake along with a before and after census of reported incidents would be fit for purpose (and such a study has been sponsored by the Patient Safety Research Portfolio).

On the other hand, when common errors are the target of a focused (specific) intervention, then even small improvements in relative risk may be cost-effective and hence it is important that improvements are accurately quantified. For example, improving by 25% the use of influenza vaccine among the elderly or detection of incipient deterioration in patients with pneumonia would result in important gains that have eluded many. While the effects of interventions in the context of high frequency incidents are often modest in terms of relative risk reduction, such practices may have large benefits at the population level; they are worthwhile because the total effect (e.g.
number of lives saved) is large even if the relative effect (e.g. 25% improved uptake) is rather small. Indeed it is precisely in the area of these high frequency incidents that the greatest health gains lie. Furthermore it is also in such circumstances – those where relative effects are modest – that the principles of randomisation and blinding are most important. This follows from the disarmingly simple principle that results will be most distorted when the plausible magnitude of effect is small relative to potential bias (Collins et al., 1996).

In conclusion, prospective comparative studies are most feasible when incidents are frequent and it is also in the context of such incidents that more modest relative risk reductions may be expected and hence where minimisation of bias is most crucial.

**Target Audience**

The level of evidence required to convince managers who have experienced the intervention that it should be continued may well be less than the evidence required to convince a manager in another part of the country or a different service (Lilford et al., 2003). In other words, a hospital manager may feel compelled to act and make a change in a service. It is appropriate for the manager to collate data and make such qualitative observations that can be afforded. The results may be sufficient to prioritise or adjust a course of action, but may be quite insufficient to persuade an external audience who, quite properly, may require a higher standard of proof. This distinction between evaluation to improve versus scientific evaluation to convince has been made in more detail elsewhere (Solberg et al., 1997; Lilford et al., 2003). In particular, an external audience will also want to see how intervention sites compared to control sites and this brings to the fore all the issues regarding precision and accuracy discussed in this article.
Which size fits whom?
An important tenet of the above argument is that there is ‘no one size fits all’ methodology for evaluation of patient safety interventions. Decisions regarding study design should reflect the four considerations outlined above. Our article has been focused on formal research aimed at producing generalisable findings to influence practice in many settings (sometimes called summative research), not just to help decision making within an organisation or continuous quality improvement (sometimes called formative research).

Here, we summarise our recommendations regarding the process involved with choosing a study design for the summative evaluation of a patient safety intervention. Firstly, proposed interventions should be ‘well-found’, having undergone PIE so that the most propitious possibilities are selected for implementation and formal evaluation. Secondly, RCTs should serve as the gold standard for studies of common incidents, although natural experiments incorporating before and after measurements are a viable option in many circumstances. Thirdly, as we move to the low prevalence end of the frequency spectrum, so the need for concurrent controls becomes both less feasible and less important; observational studies (including before and after measurements) become more acceptable as prevalence declines. Concurrent controlled studies become less feasible because the end-points of error and patient harm are rare. They become less important because we are looking for big improvements that are unlikely to be obscured by bias. Fourthly, it is usually very informative to measure the fidelity with which a proposed intervention was actually implemented on the ground.

Mixed method designs

Given the methodological difficulties associated with evaluating a patient safety intervention, we propose that mixed method designs should be considered. Such a call
for the use of mixed methods is not new, originating from the claims of Cook and Reichardt (1979). In particular, Stock (1985) highlights how mixed method designs help to ensure objectivity, enabling the virtues of one method to compensate for the flaws and biases of another. The use of mixed methods allows one end-point to be measured by two (or more) methods or for the measurement of the effects of an intervention across multiple end-points across the care chain of Figure 5. The use of mixed method designs therefore illustrates how study design and end-points interact, which in turn suggests the importance of selecting a study design appropriate for the end-point being considered.

There are three key purposes of mixed method designs:

1) Triangulation: to achieve or ensure corroboration of data

2) Complementarity: to elaborate results

3) Development: to guide additional sampling.

The concept of triangulation is borrowed from sociology and the history of science. Sociologists feel more comfortable with research findings when one type of result seems to be corroborated by another, for example if observational studies and questionnaire surveys produce similar conclusions or if carers and patients provide similar accounts of care processes. In basic science a hypothesis gains credence if different types of experiment add evidence to the same underlying theory. This additional credence will add to the strength and generalisability of the findings. Alternatively, where the results from different methods conflict, triangulation may prevent inappropriate generalisations that would have been made had results been obtained from a single research method (Brown, 2004).

In systems research if a diffuse intervention is implemented with fidelity, affects an intervening variable in a positive direction, reduces error and improves outcomes, then
we will be more confident that we have identified a cause and effect relationship, than if some of the end-points were not in agreement. Even if some of those end-points fail to reach statistical significance, the overall picture may suggest a positive effect. The risk of a false negative study result is high for many safety interventions, particularly diffuse interventions where end-points are considered separately using frequentist (conventional) statistics. We therefore advocate a Bayesian approach, where all end-points and plausibility based on PIE can all be integrated in an explicit quantitative framework.

Oakley and colleagues (2006) consider how the integration of process and outcome data can also help in interpreting and elaborating results. Such analysis enables researchers to consider why an intervention has or has not been effective and hence propose recommendations for wide-scale implementation. However this integration requires that certain methodological standards are met, such as the analysis of process data before outcome data in order to avoid bias in interpretation (Oakley et al., 2006).

In terms of development, the results arising from one research method can be used to guide research questions or designs in subsequent studies. For example, interviews could be used to identify the key issues to be investigated using quantitative methods, or alternatively the results of a quantitative analysis may highlight areas requiring in-depth investigation using qualitative approaches (Brown, 2004).

A good example of a mixed methods approach is reported by Wilson and colleagues (2002). The authors assess the change in compliance with four evidence-based recommendations for perinatal care in two years, before and after the publication of randomised trial evidence to support the recommendations. The quantitative audit was
supported by interviews with 88 members of staff across 20 hospitals. Data from the interviews are used to explore differences in final compliance rates.

Finally, methods and instruments for assessing the study’s end-point(s) must be determined. Research to identify valid and reliable methods of data collection is currently underway to add to the extant evidence reviewed in this paper. The use of an appropriate study design which applies valid and reliable measures of a meaningful end-point is essential if the results of the study are to be assimilated into health care policy and hence have a significant impact on patient safety.

One size does not fit all: Summary

The study design chosen for an evaluation of a patient safety intervention will depend on a number of factors:

- Logistical/pragmatic constraints
- *A priori* assessment of benefit and harm
- The possibilities and consequences for improvement
- The target audience for the results

The strength of any conclusions drawn could be increased if different end-points concur and we therefore advocate the use of mixed methods and an assessment of the effect of the intervention on different end-points.
Acknowledgements

MRC Patient Safety Grant

NHS Research Methodology Programme

Evaluation of the uptake of NPSA guidance - NPSA grant

NHS Patient Safety Research Portofolio
Figure 1: The Quality/Safety Continuum

Note: The letters A-D refer to examples of clinical errors with different degrees of causality and immediacy provided in Table 1.
Table 1: Examples of clinical events/error that have differing degrees of causality and immediacy and thus lie at different points of the quality/safety continuum

<table>
<thead>
<tr>
<th>Error</th>
<th>Causality</th>
<th>Immediacy</th>
<th>Point on Figure 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrathecal vincristine/inadvertent intravenous admin</td>
<td>High</td>
<td>High</td>
<td>A</td>
</tr>
<tr>
<td>istration of potassium chloride concentrate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Failure to vaccinate</td>
<td>High</td>
<td>Low</td>
<td>B</td>
</tr>
<tr>
<td>Failure to use thrombolytics in myocardial infarction</td>
<td>Low</td>
<td>High</td>
<td>C</td>
</tr>
<tr>
<td>Failure to use beta-blockers post myocardial infarction</td>
<td>Low</td>
<td>Low</td>
<td>D</td>
</tr>
</tbody>
</table>

By causality we mean the confidence with which a bad outcome, if it occurs, can be attributed to the error. So if someone who should have been vaccinated against influenza gets the disease, then it is quite likely that this could have been prevented. On the other hand, recurrent myocardial infarction probably would not have been prevented by beta-blockers.
Figure 2: Basic Study Designs

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Time period of Data Collection</th>
<th>Control</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncontrolled</td>
<td>Pre-intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post-intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controlled, Cross-sectional</td>
<td>Post-intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controlled, Longitudinal</td>
<td>Pre-intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post-intervention</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Control/Intervention determined by convenience (natural experiment) or randomisation

Key

- Aggregated assessments for individuals within a study group (there may be multiple groups) or individuals within clusters within a study group (only for controlled studies – there will be multiple clusters within a study group). If there are multiple groups, comparisons of the effects of the intervention across groups are also possible.

- Comparison of study end-points.
Figure 3: Clustered Sample Size Calculation for given ICC values: alpha = 5%, power = 80%. Detecting a difference in error rate between control (0.1) and intervention (0.05).
Table 2: Controlled study design matrix

<table>
<thead>
<tr>
<th>Phasing</th>
<th>Allocation: Randomised</th>
<th>Natural experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post intervention</td>
<td>Large variance due to base-line difference will yield an imprecise result unless the number of clusters is large.</td>
<td>Huge risk that comparison confounded by differences between organisations.</td>
</tr>
<tr>
<td>Before and after</td>
<td>Allows for specific comparison of change net of any base-line differences. Enables comparisons to be made between sites that change most or least.</td>
<td>Controls for base-line difference possible—see text.</td>
</tr>
</tbody>
</table>
Table 3: Some examples of Safety Issues and how they may be studied in the context of evaluation

<table>
<thead>
<tr>
<th>Example</th>
<th>Issue</th>
<th>Suggested Approach to Evaluating an Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong site surgery</td>
<td>Error and outcome are one and the same, but very rare</td>
<td>Develop an intervention through diligent PIE. Monitor fidelity with which instituted. If possible monitor any national statistics—e.g., litigation rates.</td>
</tr>
<tr>
<td>Prescription error</td>
<td>Outcomes for patient have large signal to noise ratio, so they are of little use. Trigger tools (use of antidotes, suddenly discontinued medication) can indicate where to look, but they are likely to provide biased estimates in comparative studies, unless precisely the trigger tools used across comparator sites all have precisely the same sensitivity and specificity.</td>
<td>Measure errors from case notes including prescription charts. Can measure errors trapped by pharmacy, but they may not be a good surrogate for error carried through to patient.</td>
</tr>
<tr>
<td>Hospital acquired infections</td>
<td>Comparisons between hospitals highly biased by case mix. Will also be biased if each hospital measures its own rate, therefore because of demonstrated observer bias, even when using guidelines/protocols</td>
<td>If outcome is end-point use independent observers each observing many institutions. Monitor safety practices—e.g., hand washing/use of antibiotic impregnated lines. Cluster RCT design feasible.</td>
</tr>
<tr>
<td>Failure to detect deteriorating patient</td>
<td>Collapse and death are common occurrences, but the signal to noise ratio is very high. However, quality of care/clinical errors can be discerned from case notes using the ‘opportunity for error’ principle</td>
<td>Large cluster RCT or natural experiment with before and after measurement. If enough clusters could be obtained then mortality and collapse rates could be measured. If not, error rates should be measured. Even if outcomes used as an end-point, we would advocate measurement of error rates on a sample of cases.</td>
</tr>
</tbody>
</table>
Figure 4: Stepped Wedge Trial Design

<table>
<thead>
<tr>
<th>Participants/Clusters</th>
<th>Time periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Shaded cells represent intervention periods
Blank cells represent control periods
Figure 5: General and Specific Interventions across the system and Evaluation end-points
Figure 6: Intellectual framework to classify links between care and poor outcome

- Unalterable features of disease → poor outcome
- Treatment properly given → poor outcome
- Latent factors → Error (of commission or omission) → poor outcome

Adverse Event
Table 4: Comparison of Implicit and Explicit methods of data extraction in case-note review

<table>
<thead>
<tr>
<th>Implicit</th>
<th>Explicit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy to develop and administer</td>
<td>Require training of reviewers</td>
</tr>
<tr>
<td>High face validity</td>
<td></td>
</tr>
<tr>
<td>Self-updating through use of experts</td>
<td>Need to be updated constantly</td>
</tr>
<tr>
<td>Reflects the full scope of clinical decisions that apply to a particular patient</td>
<td>Limited scope in terms of content and context (relevant populations)</td>
</tr>
<tr>
<td></td>
<td>Does not capture the subtleties of health care (e.g. contraindications)</td>
</tr>
<tr>
<td>Involves physicians in the quality of care process</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expensive</td>
</tr>
<tr>
<td></td>
<td>Potential for gaming</td>
</tr>
<tr>
<td></td>
<td>Need to decide how to analyse multiple criteria</td>
</tr>
<tr>
<td></td>
<td>Possible bias if different numbers of criteria apply to patients between comparative sites, particularly if some criteria are harder to meet than others</td>
</tr>
<tr>
<td>More arbitrary than evidence based</td>
<td>Explicit criteria</td>
</tr>
<tr>
<td>Developed principally for inpatient care</td>
<td>Reproducible</td>
</tr>
<tr>
<td>Poor reproducibility of judgements</td>
<td>Easy to explain low score in terms of criteria – which may narrow score of improvement efforts</td>
</tr>
</tbody>
</table>

Advantages are shown in the shaded cells.
The actual sample size needed will depend on values of $p_o$ (observed proportion of agreement) and $p_e$ (expected proportion of agreement). The table below shows kappa values (K) and the number of case-notes needed (N) to show $K>0.6$ with 95% confidence. All sample sizes have been rounded to the next integer.

<table>
<thead>
<tr>
<th>$p_e$</th>
<th>$p_o = 0.7$</th>
<th>$p_o = 0.8$</th>
<th>$p_o = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>K=0.67</td>
<td>K=0.78</td>
<td>K=0.89</td>
</tr>
<tr>
<td></td>
<td>N=157</td>
<td>N=17</td>
<td>N=4</td>
</tr>
<tr>
<td>0.2</td>
<td>K=0.63</td>
<td>K=0.75</td>
<td>K=0.88</td>
</tr>
<tr>
<td></td>
<td>N=1412</td>
<td>N=30</td>
<td>N=5</td>
</tr>
<tr>
<td>0.3</td>
<td>K=0.57</td>
<td>K=0.71</td>
<td>K=0.85</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>N=68</td>
<td>N=8</td>
</tr>
<tr>
<td>0.4</td>
<td>K=0.50</td>
<td>K=0.67</td>
<td>K=0.83</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>N=269</td>
<td>N=13</td>
</tr>
<tr>
<td>0.5</td>
<td>K=0.40</td>
<td>K=0.6</td>
<td>K=0.8</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N=24</td>
</tr>
</tbody>
</table>
References


Barber, N, B Dean Franklin, T Carnford, E Kleun, I Savage, 2005, Piloting technology interventions to reduce medication errors, Report to the Patient Safety Research Portfolio.


Cook, TD, DT Campbell, 1979, Quasi-experimentation: Design and analysis issues for field settings, Boston, MA: Houghton Mifflin Company.

Cook, TD CS Reicgadt (Eds.), 1979, Qualitative and quantitative methods in evaluation research, London: Sage Publications.


Gordon, R, J Grimshaw, M Eccles, R Rowe, J Wyatt, 2003, On-screen computer reminders; effects on professional practice and health care outcomes (protocol for a cochrane review), The Cochrane Library: Chichester, UK, John Wiley & Sons Ltd.


Lilford, RJ, D Braunholtz, J Harris, T Gill, 2004b, Trials in surgery, British Jnl Surgery, v. 91, p. 6-16.


An epistemology of patient safety research

March 2006


MRC., 2000, A framework for development and evaluation of RCTs for complex interventions to improve health.


NPSA, 2004, Seven steps to patient safety, London: NPSA.

NPSA, 2005, Building a memory: Preventing harm, reducing risks and improving patient safety, London: NPSA.


Ovretveit,J, 2005, Which interventions are effective for improving patient safety? A review of research evidence: Karolinska Institutet Medical management Centre.


Sandelowski, M, 2000, Combining qualitative and quantitative sampling, data collection and analysis techniques in mixed method studies, Research in Nursing and Health, v. 23, p. 246-255.


Shortell, SM, RH Jones, AW Rademaker et al., 2000, Assessing the impact of total quality management and organisational culture on multiple outcomes of care for coronary artery bypass surgery patients, Medical Care, v. 38, p. 207-17.


University of Aberdeen Health Services Research Unit Database of Intra-cluster Correlation Coefficients. http://www.abdn.ac.uk/hsru/epp/cluster.shtml


