Evaluating Complex Social Interventions
Volume 1: Challenges and Choices

John Øvretveit, jovret@aol.com
Director of Research, and
Professor of Health Innovation, Implementation, and Evaluation
The Medical Management Centre,
The Karolinska Institutet, Stockholm.

Reference citation: Ovretveit, J 2013a Evaluating Complex Social Interventions: Volume 1: challenges and choices, CIPRS, Veterans Health Administration, Sepulveda, Ca.

Reference citation: Ovretveit, J 2013b Evaluating Complex Social Interventions: Volume 2: Guidance, tools and resources for researchers, CIPRS, Veterans Health Administration, Sepulveda, Ca.

Contents

Example evaluation studies and their strengths and weaknesses are given in a separate appendix document. Practical guidance is in separate and third document.

Abstract

Most health services researchers know about experimental evaluation designs and quantitative research methods. Some have carried out a controlled trial. Fewer are aware of strategies for using these methods to evaluate complex social interventions (CSIs).

This "challenges and choices" document describes different research designs for providing the information about complex social interventions which practitioners at different levels of the health system need to provide more effective services.

It describes methods designed to take account of the social nature of such interventions – methods more often used in knowledge domains other than health services research.

The purpose is to enable researchers to choose a design, within the constraints for the evaluation, with a greater awareness of the strengths and limitations of the design for the research-user's question.

Acknowledgements: My thanks to many colleagues at CIPRS and the Center for the Study of Healthcare Provider Behavior in Sepulveda and Los Angeles for their support and comments on earlier drafts, especially Brian Mittman, David Ganz and Lisa Rubenstein.

Controlled trials are the best of designs,
and the worst of designs, depending on the question.

Volume 1 Summary

A wider range of designs and methods than many medical and health service researchers are familiar with are useful for evaluating complex social interventions (CSIs) and for the first research phase of question-formulation. Some CSIs can be prescribed, implemented with fidelity to the prescription, and evaluated using experimental designs and controls. Conventional and modified designs of this type are described in this volume for answering “does it work” effectiveness questions, when these designs are feasible. However, such designs may not be practical or appropriate for other CSIs or for other questions. Other observational and action evaluation study designs are described as alternatives, with their strengths and weaknesses noted.

This volume gives examples of “complex social interventions” (CSIs), describes the challenges in evaluating them and discusses the research designs which can be used to answer different questions.

The accompanying Volume 2 gives practical resources, guidance and tools for researchers and implementers for carrying out these eight steps in evaluating a CSI:

1. User goal-specification
2. Reviewing relevant research
3. Defining practical and scientific questions
4. Listing and choosing designs
5. Preparing for the evaluation
6. Data gathering
7. Data analysis
8. Reporting, publishing and dissemination activities

For implementers, these two volumes are useful for planning implementation of a change to a healthcare service.

Why is guidance needed to evaluate CSIs?
For three reasons:

- Because of the challenges in applying familiar designs,
- To choose a design which best answers the evaluation question, researchers need to know about a wide range of designs,
- Some designs are unfamiliar but could be applied by researchers using this and other guidance.

Evaluation is systematically gathering information and making comparisons to enable a customer/user to make a better-informed decision about a change they are considering. There is useful guidance for evaluating complex interventions (CIs), but this guidance has limitations for answering many questions posed to evaluations of small- and large-scale changes or programs. Complex social interventions (CSIs) are often not standardizable, may involve adaptation and dynamic iteration, may be multi-level, and are also subject to psychological, social and political influences, which are often unpredictable.

Evaluation users typically need to know whether an intervention or change is effective according to some criterion of effectiveness, such as whether the intervention reduces infections or improves patient experience. Traditional experimental evaluation designs can be used to answer some effectiveness questions about some CSIs, but many CSIs require other designs to answer the "does it work?" question.

In addition, there are other questions which users of evaluation research have about interventions including: Would it be effective in our service at this time? How should we effectively implement it locally? How much does it cost and save for different parties, and will it give a return on investment? Other designs can be used to answer these questions, as well as to give answers to the "does it work" question when controlled trials are not feasible.

The discussion in these two volumes emphasizes that the CI vs. CSI distinction is less about the intervention and more about how it is conceptualized. Many interventions – such as a care improvement "bundle" to reduce infections - can be thought of as complex interventions (CIs) or as complex social interventions (CSIs). The term "social" emphasizes, where appropriate, consideration in the evaluation of different humanly-created influences by implementers which may affect the recipients of the intervention and others. It points to evaluation methods and questions that are more commonly used and proven in scientific fields outside of medical and health services research.

Evaluation challenges and questions

All evaluations face five challenges, phrased here as questions. Each becomes more difficult to address with the more complex and more social interventions,

- Aims: How to formulate the questions which the evaluation will be designed to answer and decide what information is needed to answer them?
- Description: Which data to collect to describe the intervention, implementation and the context, how each may change, and which data to collect to assess effectiveness?
- Attribution: How to assess the extent to which the intervention caused the outcomes and not something else?
- Generalization: How to assess whether the intervention can be implemented in another setting and whether it would show similar effects?
- Usefulness and Use: How to make the evaluation useful for practical action and how to enable others to make use of the findings?

Different designs and their strengths and weaknesses

For the purposes of this "challenges and choices" document, evaluation designs suitable for different questions about CSIs can be categorized as:

1. Experimental and quasi-experimental,

2. Observational, and,
3. Action evaluation.

The different designs within each category, and their strengths and weaknesses for answering different user questions are described in more detail in the accompanying Volume 2 "guidance and tools" document.

In this volume, examples of studies using some of the common designs are given in the appendix:

- Randomised controlled trial (RCT) of a care transitions intervention
- Pragmatic cluster RCT of a multifaceted intervention in ICUs
- Interrupted time series evaluation of a hand hygiene intervention
- Prospective observational evaluation of an intervention to change primary care physicians' behavior
- Prospective observational evaluation of a large scale safety program
- Process observational evaluation of an intervention to promote smoking cessation
- Case evaluation of a program for electronic summaries of patients' medical records
- Realist evaluation of large-scale "transformation" of health services in London

The researcher-evaluator has choices about which design to use depending on the questions the evaluation is to inform and on the practical constraints to the evaluation, such as time, resources and whether the intervention has already started.

Summary of CSI evaluation challenges

Challenges exist for evaluators seeking to answer users' questions, which include:

- Is this intervention or change effective?
- Would it be effective in my local setting?
- What is the most effective way to implement it?
- In which conditions do we need to be successful in implementing it?
- How much does it cost and will it give a return on investment?

Evaluations of CSIs face similar challenges and issues to those facing evaluations of standardizable complex clinical interventions; notably: identifying which of the component interventions are necessary and which may be discarded or reduced, identifying which outcome measures are to be used, and assessing how much different contexts are likely to influence outcomes. Experimental designs can give high-certainty answers to the effectiveness question when they are able to control effects other than the intervention, and where effectiveness can be assessed through a few specific measures.

These designs are less useful for answering other questions and for evaluating some social interventions where human choice and changing social conditions may be powerful influences which make outcomes less predictable. In these interventions, discovering how people interpret the influences exerted to enable them to change may help answer other questions, but also can help answering the “is it effective?” question.

Many observational designs for answering both effectiveness and implementation questions are less familiar to both researchers and practical users of evaluations and face similar difficulties in generalizing the findings to other settings. This presents challenges in communicating the findings, and for users in assessing their significance and local applicability. These challenges are particularly acute with action evaluations, where evaluators need to describe in detail how they assisted the development of the CSI so that others can assess whether they can or need to provide similar assistance.

Conclusions

Many interventions to healthcare, patients and populations are not only complex but social interventions: the “social” emphasises specific dynamics and an unpredictability which need to be considered by evaluators. This makes CSIs more challenging to evaluate that some interventions. Yet the cost and time increasingly invested in CSIs also calls for more resources to be allocated to their evaluation and to improving evaluation methods and evaluator skills.

Experimental designs are suited to assessing associations between the presence of the intervention and measured outcomes for providers and/or patients. Observational designs provide less certainty about these associations because there is less control for alternative explanations for outcomes. Some observational designs are, however, able to develop explanations for outcomes which can assist in implementing the CSI in settings other than the evaluation setting. In action evaluations, researchers help improve the CSI as it is implemented and gain insights, which further contributes to explanations about outcomes. However, the evaluator’s participatory role, if significant, may be difficult for others to reproduce, which can reduce the external validity of the findings.

Awareness of the range of designs, and of ways better to formulate answerable evaluation questions, makes it more likely evaluators will better match design to the purpose and constraints of the evaluation. Other developments are needed such as more attention to pre-data-gathering program theory, reporting standards and assigning costs.

There are other methods which can be grouped as “pre-implementation evaluation approaches” which include those described by Craig et al (2008) and Brown et al (2008), as well as simulation modeling, or policy analysis which extrapolates from existing evidence to estimate costs and impact of spreading an intervention. Neither these nor multi-level modeling, Bayesian and adaptive trials are described, but these are noted in Volume 2. Volume 2 provides guidance, tools and resources for researchers to improve their evaluations of CSIs. The rest of Volume 1 draws on the literature to discuss the issues noted in this summary and to show the choices of design, with examples.

1. Introduction

This is the first of two documents to help researchers to choose an evaluation research design and to plan a study which answers the study user's questions. Implementers will also find these two volumes useful to plan a change to a healthcare service, and to carry out and perform an evaluation of a change which may be the introduction of a new service.

Volume 1 gives examples of "complex social interventions" (CSIs), describes the challenges in evaluating them, the research designs which can be used, with illustrative studies and their strengths and weakness for answering different questions, and considerations in choosing design. Volume 2 gives practical resources, guidance and tools for researchers and implementers for carrying out the different steps in evaluating a CSI:

- User goal-specification
- Reviewing relevant research
- Defining practical and scientific questions
- Listing and choosing designs
- Preparing for the evaluation
- Data gathering
- Data analysis
- Reporting, publishing and dissemination activities

The two volumes draw on an overview of research and on experience conducting evaluations and implementation research in the VA and other health systems. This overview was carried out by the author, who is a Swedish medical university researcher with experience and knowledge of the field, and who has worked with the Veterans Health Administration (VA) on evaluating complex social interventions for the last 4 years.

The short section below explains the broad conception of evaluation used in this document. Then the rest of the introduction considers why we need to evaluate CSIs and features of such interventions which need to be considered when planning and carrying out an evaluation.

Broad definition of evaluation

The document is limited to research for evaluating CSIs, defined as judging the value of a complex social intervention by gathering information about it in a systematic way and by making a comparison, for the purpose of making better informed decisions (Øvretveit, 1999, 2002). This approach views evaluation as providing the "customer" of the evaluation ("evaluation user") with systematically collected data and comparisons to enable them to make better informed decisions. This is accomplished by assessing one or more of:

- the intermediate effects of the CSI on health care provider perceptions, behavior and organization,

- the ultimate effects on patient experience, clinical outcomes and resource use,
- the “implementablity” of the CSI, by showing the number of actions and steps needed to achieve similar results,
- the conditions needed for successful implementation (the CSI is of no value to some users if it requires resources or other conditions which cannot be created locally),
- whether the CSI has to be copied exactly to get similar results, or if the implementer could get similar or better results if they adapt parts.

Evaluation research is thus viewed more broadly than only assessing the difference the intervention makes to one or a few defined outcome variables. It also includes research to gather other information which users may need to judge the value of the intervention locally, such as whether it is implementable in or adaptable to their setting.

The document also draws on ways of thinking about “interventions”, “complexity” and “social” which are explained below, and which may be unfamiliar to some medical and health service researchers, but which underlie some of the observational research approaches described later.

1.1. Background

Why do we need to evaluate complex social interventions?
Across the world, substantial amounts of time and money are being spent on large scale intentional change programs: schemes to introduce digital health technologies, national, regional or system-based quality and safety programs, and programs to influence performance with financial incentives or indicator-reporting. In the USA, the progress of the Affordable Care Act requires that demonstration projects document effective methods to improve quality and control costs (PPACA 2010). The US Center for Medicare and Medicaid Innovation (CMMI) requires timely but also rigorous evaluations of the complex interventions it is funding (Gold 2011).

Smaller scale changes to one hospital or unit are also complex social interventions, such as,

- “Care bundles” introduced in to one intensive care unit to reduce ventilator associated pneumonia (VAP) or central line associated blood stream infections (CLABSI),
- A multiple intervention to a long term care facility to prevent patient falls, or to prevent pressure ulcers, or to prevent infections with hand-hygiene programmes.
- Introducing a computer decision support system to assist physicians in diagnosis and treatment decisions.

By spending a percentage of the project on evaluation, management or policy makers can gain information which can be used to improve the next phases or to decide whether or how to spread the intervention to other hospitals or units. Evaluations can contribute to science if they provide not only knowledge of efficacy, but also insight and theory into how or why the intervention has its effects, which may be none, or harmful effects.

Evaluation of these changes can also provide accountability. Process or formative evaluation can help to modify and improve a program as it develops (Stetler et al 2006). Later, summative evaluation of outcomes can help to design more effective programs in the future, as well as decide whether to extend the program (Landefeld et al 2008).

What are "complex interventions"?

Authoritative guidance for medical researchers was published in 2000 and described a complex intervention as being "built up from a number of components, which may act both independently and inter-dependently." (MRC 2000).

Craig et al (2008) updated this guidance and describe the following as features of an intervention which make it complex:

- The number of interacting components within the experimental and control interventions,
- The number and difficulty of behaviours required by those delivering or receiving the intervention,
- The number of groups or organizational levels targeted by the intervention,
- The number and variability of outcomes,
- The degree of flexibility or tailoring of the intervention permitted.

The table below shows increasing degrees of complexity of interventions and their implementation strategies, and the level from which the intervention is initiated. Evaluations are needed of each type of intervention, and will require different evaluation methods to answer different user questions.

Different complexity of interventions and level of intervention

Level of intervention	Simple intervention	Complex intervention & progressive implementation (CPI)	Complex intervention & adaptive implementation (CAI)
Region, National (policy, public health)	Intervention: directive to withdraw a pharmaceutical	CLABSI bundle implemented in some ICUs, then more ICUs added	Pilot an Accountable Care Organizations, then revise after government funding model is changed, then make further changes.
Health system	Change the wound care kit we use.	Same as above	Infection control program, continually revised
Health facility	Change the hospital signage to help patients find different services	Same as above	Infection control program, continually revised
Team or department	Install more hand washing gel dispensers.	ICU use CLABSI bundle with some and then all patients	Lean manufacturing principles applied to make initial changes, then more changes after experience with testing first set of changes and with changes in reimbursement

What are “complex social interventions”?

The guidance by Craig et al (2008) and others is useful also for evaluating complex social interventions but also has its limitations. There are features of the “social” part of many interventions which are not fully considered in the above guidance, as well as other information which decision makers need, and which call for methods which are discussed in more detail in this document below.

Examples of CSIs are:

- Care improvement “bundles” such the combination of actions which need to be take in an ICU to reduce ventilator associated pneumonia (VAP) or central-line associated blood stream infections (CLABSI). These can be evaluated as complex interventions to answer questions about efficacy using the designs described in the above guidance (Craig et al 2008). Or they can be evaluated as complex social interventions,
- Quality and safety improvement initiatives: local small projects or large-scale programmes such as the IHI saving 100,000 lives programme (IHI McCannon et al 2006),
- Introducing a digital health technology to support providers and or patients,
- Systems redesign projects such as changing primary care organization to improve access (Lukas et al 2007),
- Different implementation programs, typically where one or more services applies a practice or model tested elsewhere, such as a model for chronic care management (Wagner et al 2001),
- Large scale reforms, such as parts, or all of the Affordable Care Act (PPACA 2010)

The “CSI” term refers both to the intervention being complex and social, but also to the unit of change being complex and social, such as a group or organization. These interventions are often not standardizable, as are pharmaceutical interventions, and the effects on the unit of change are not as predictable effects on the unit of change are not as predictable effects on the unit of change are not as predictable. Evaluation designs which are suited to assessing the effects of changing interventions to social groups are often necessary to answer the questions of the customers of the evaluation. This document describes these designs and when they may be more appropriate than a design whose purpose only is to identify the effects of an intervention on defined outcome variables.

Many interventions – such as a care improvement “bundle” to reduce infections - can be thought of as complex interventions (CIs) or as complex social interventions (CSIs).

The distinction is not so much about the intervention, as it is about the way it is conceptualized: the term “social” points to evaluation methods and questions less familiar in medical and health services research.

1.2. Complex

Complexity is both in the nature of the intervention, and in how it is conceptualized: any intervention can be conceptualized as simple (e.g., one action causes results) or as complex (e.g,. many parts interacting). The features of complexity which are important for designing an evaluation and data gathering are:

- **Content** of the intervention: complex, if it involves a number of changes or is "multiple component" (e.g., a CLABSI bundle).
- I**mplementation** may involve an infrastructure of groups, a strategy of actions, and information technology or other systems to support care providers or patients.
- **Change** may be made to the content or implementation at different times. The intervention may be adapted to fit the organization or adapted in response to changes in financing. Many programs which are the subject of CSI evaluations are reviewed and revised at certain times, and these changes need to be taken into account in the evaluation. This is especially so when evaluating quality improvement projects where testing and iteration are part of the change methods.
- **Levels** of the health system at which the intervention is directed. The components of the intervention may be directed at a single level, for example all directed at individual care provider behavior (e.g., training, reminders and feedback). But some interventions also involve actions to influence more than one level (e.g., in addition, directives or financial incentives to influence management to support the care provider's behavior in different ways) and to create organizational- or wider- conditions to enable lower-level change (e.g., state financial incentives, or regulations).

1.3. Social

"Social", added to "complex intervention", highlights the human and group aspects of the change. It draws attention to the behavioral content of the change (e.g., "wash your hands") and the psychological (choice, motivation) and social nature of the implementation (e.g., project groups, training programs, negotiating union agreements, negotiating changes to physicians' schedules and work practices).

It implies that the intervention is not acting through predictable physical processes following natural laws, but according to psychological, social and political patterns and influences. These have different "causal processes", because people individually and in groups have a choice about how they act and influence each other. This does not mean that there are not discoverable regularities and patterns in human behavior. It does mean that prediction is less precise than it is for many natural physical and biological processes where human interpretation and choice has no or little influence. How one group behaves in one place may be different to how another group behaves elsewhere. Knowing why groups do what they do may help to predict or design an intervention elsewhere.

1.4. Intervention

A third set of concepts relate to the meaning of “intervention” when applied to changes of this type rather than to discrete pharmaceutical or surgical interventions to patients.

- **Scale**: Changes falling within the category of “CSI” can be large scale, covering many organizations or units or people (e.g., national safety program), or small scale covering one unit (e.g., a multiple-component “CLABSI care bundle” in one ICU).
- **Unit of Change**: the people, organizations, facilities or communities which are intended be different, as a result of the intervention. (e.g., a CLABSI intervention aims to change ICU care providers’ behavior and ultimately patient outcomes)
- **Content**: the way in which the units of change are intended to be different after the intervention (e.g., care providers carry out the components of the CLABSI bundle: 1 Wash your hands, 2 Clean skin with chlorhexidine. 3. Use maximal barrier precautions. 4. Avoid the femoral site. 5. Ask daily whether the benefits of the line exceed the risks)
- **Implementation**: what is done to enable the units of change to adopt the change content (project group established, education, reminders, observation, performance feedback).

Is the intervention-change itself complex, or is complexity in how we see it?

Enabling one physician to use a medication more selectively is usually a less complex endeavor than enabling all physicians in a health system to do so. Some changes are more complex, when compared to other changes. But complexity also is a property of how we perceive of a change: we can choose to understand a change in a more or less complex way. It is common in medical and health services research to understand a change as a cause, and to study effects using experimental research methods. Arguably a more complex way of viewing the change, is to view the change as one of many influences affecting physicians, and operating in a system. A systems view expects this change to have delayed effects on physicians, and that physicians may take actions to modify the change itself. There are other ways to view the change and possible effects which offer more or less complex ways of understanding the change.

This document regards “complex social interventions” as changes which are objectively more complex, when compared to other changes. But also, that complexity resides in how we choose to view a change – we can understand it in a simple way for example as a cause and effect, or in a complex way, for example as one influence in a dynamic system. The approach taken is that no one way of viewing a change is better than another, but that one way may be more useful for answering a specific question. Viewing a change as if it were a simple natural process, with the change as the cause and the outcome as an effect, may be useful for some purposes. Viewing it as a process subject to psychological, social and political processes may be useful for other purposes.

1.5. Lessons for evaluating CSIs from the field of informatics

Within informatics and the sub-field of digital health technologies (DHTs) some researchers take the view controlled experimental trials should be the design of choice for CSIs. This is proposed by some especially for those DHTs where the stakes are high if the DHT were rapidly adopted without rigorous testing – where the potential risks of harm for patients and costs of the intervention are high (Liu and Wyatt 2011). But it is also widely accepted within this field that a range of designs are appropriate for answering different questions, and for evaluating where informatics CSIs which do not expose patients, providers or others to risk of harm and are inexpensive.

Illustration: Digital Health Technology (DHT) systems as complex social interventions: evaluation issues and choices of method

Some of the issues in evaluating CSIs and the use of different methods are exemplified in the field of informatics (UKIHI 2001, Stoop et al 2004, Brender 2006). Lehmann & Ohno-Machado (2011) note that informatics systems involve multiple interventions, and that "adherence, in terms of adoption, varies greatly within an institution; and training for HIT systems is notoriously variegated." They note that the practical challenges of using some controlled trial experimental designs prevent their being used, and that the technology of the intervention is often changed during the trial and that,

> *"JAMIA publishes qualitative evaluations and reports of studies that employ methodologies designed to model behavioral and social systems and does not restrict its publications to quantitative evaluations or studies that utilize methodologies designed to model physical systems. Therefore, JAMIA ignores the perceived dichotomy between "soft" versus "hard" sciences. The field of health and biomedical informatics is diverse and each study is unique; thus, it is important to understand what is most appropriate for a particular investigation and to avoid a priori rule-in or rule-out of particular methodologies"*
>
> Lehmannm & Ohno-Machado (2011)
> Journal of the American Medical Informatics Association.

Liu and Wyatt (2011) note some of the criticisms of RCTs in the field of medical informatics:

- Trials are unethical.
- Clinical information systems are too complex to be evaluated by RCTs.
- There is mixed evidence that RCTs can be successfully applied to medical informatics.
- Other study designs can also provide evidence that is just as reliable as RCTs.
- Trials are too expensive.
- Theory or case studies can reliably predict what works.
- Clinical information systems can do no harm.
- RCTs take too long, and technology in medical informatics moves too quickly.
- Trials answer questions that are not of interest to medical informatics.

They answer each criticism and present the case for RCTs, arguing that it has been underused in medical informatics, could be used with modifications, and needs to be used when the risk/cost stakes are high. They describe quasi-experimental designs which can be used more (discussed later in this volume and in

Friedman & Wyatt 2006), and note how a multi-arm randomized trial can be used to quantify the contributions of different components of a complex intervention.

For evaluating DHTs, Freeman et al. (2006) propose that evaluation assesses:

- Extent of fit between the innovation and context
- Stakeholder perceptions and experiences of the innovation
- Extent of change to services and outcomes
- Extent to which new practices have become embedded
- The effects (and unintended consequences) of the innovation on services, services users, and the wider system
- Learning that can be transferred to other settings and how this relates to the broader literature on innovation.

Williams (2011) proposes that:

> *"…when measuring inputs and outputs of innovations, the former are likely to include quantifiable financial, human and physical resources alongside the more difficult to measure tacit knowledge (Adams et al 2006). Issues to bear in mind when drawing up a list of outcome measures include not just benefits to the organization and patients, but also the distribution of positive net benefits, for example, between organizations, functions and user groups."*

Coyte & Holms (2007)

Overall, evaluators of DHTs have been readier to use a range of evaluation designs and action evaluation than medical and health service researchers. There is much to be learned from the debate about methods in this and other fields, and from the way methods have been developed which would help the much needed innovation in methods for evaluating complex social health interventions: one of the aims of this document is to draw on these discussions to contribute to such innovation in methods.

1.6. Summary

Evaluation is systematically gathering information and making comparisons to enable a user to make a better informed decision about a change they are considering. There is useful guidance for evaluating complex interventions, but this guidance has limitations for answering many questions posed to evaluations of small and large scale changes or programs. CSIs are often not standardizable, may be multi-level, and are also subject to psychological, social and political influences which are often unpredictable. Although traditional experimental evaluation designs can be used to answer some questions about some CSIs, many CSIs require other designs to inform a variety of evaluation users' decisions. The distinction between complex interventions and complex social interventions does not distinguish different interventions. The distinction is one between different ways of conceptualizing interventions and points to different ways of answering evaluation questions.

2. Challenges for describing and evaluating complex social interventions

Some challenges are more extreme versions of those faced by any evaluation, but some are unique to CSIs. This part of the document notes shortcomings described by previous research, and discusses the challenges for evaluators.

2.1. Shortcomings of evaluations of CSIs

The first set of limitations described in the literature refers to inadequate design for certainty about effects. Many systematic reviews of evaluations of complex interventions assume the question is only, "did the CSI have an effect on important patient outcomes?", and assess an evaluation according to whether it used designs and methods thought to be best for identifying any such effects, such an RCT. Thus Cochrane reviews often note the limitations of many evaluations in relation to review criteria such as randomization, blinding, and control of confounders (EPOC 2009).

Practical: time, costs, randomization, controls and static standardization
Controlled trials take time and are costly to carry out. When the intervention is a change to provider units, randomization of these in sufficient numbers may be difficult to achieve. It may be difficult to agree beforehand on a set of controls not getting the intervention. The intervention may be difficult to standardize or hold constant, and some interventions may need to be adapted during implementation in ways which cannot be pre-defined. Proponents of RCTs suggest that RCTs can accommodate tailoring to the situation, as in adjusting the dose to the patient, and have suggested different modifications to design which are described later in this document.

In considering shortcomings, Kessler & Glasgow (2011) suggest about RCTs:

> *"When applied to the other major issues facing health care today, such trials are limited in their ability to address the complex populations and problems we face."*

…and further, that,

> *"A moratorium is proposed on such research for the next decade, and pragmatic, transparent, contextual, and multilevel designs that include replication, rapid learning systems and networks, mixed methods, and simulation and economic analyses to produce actionable, generalizable findings that can be implemented in real-world settings is suggested"*

Descriptions: intervention and context is not described
One shortcoming noted by Campbell at al (2000) is a lack of description about exactly what the intervention was and the setting in which it was implemented. This has also been noted in many other reviews (Michie et al., 2009).

Descriptions are needed to allow others to reproduce the CSI. One example is evaluations of dedicated stroke units, where there are often great variations in staff characteristics, clinical practices, management protocols, and infrastructure – another example is evaluations of rapid response teams, where there are similar variations. Without good descriptions of the intervention and the context, it is difficult to assess if the intervention was implemented fully and as intended, or implemented in a context for which it was not designed.

This makes it difficult to judge generalizability: whether the intervention could be implemented elsewhere and if the results would be similar. Campbell et al's (2000) conclusions are not just regarding the need for better description, but also that "unless the trials illuminate processes and mechanisms they often fail to provide useful information".

Outcomes: not explained by theory

One criticism is that many evaluations do not assess whether all the components of a CSI are needed or the contribution which each individually and together make to the outcomes. RCT designs in particular have been criticized for this, but a multi-arm randomized trial can be used to quantify the contributions of different components of a complex intervention (Friedman & Wyatt 2006, Altman 1991). Each arm of the trial exposes a group to an intervention component, or refers to the control group. If components of the intervention are not known, an expert panel can be used to define the potentially important components as a basis for the design (Friedman et al. 1998).

The above description-limitations are echoed in Glasziou et al. 2008) and in a review of 72 evaluations of quality improvement breakthrough programs (Schouten 2008). These involve project teams from different services meeting to learn and apply quality methods and to share experiences and results (Kilo 1998). This review first notes shortcomings arising from researchers not being able to apply adequate experimental design features, such as possible differences in baseline measurement between sites, limited data about the characteristics of control sites, no specification of blinded assessment, and possible contamination. But in addition, Schouten et al. ask for more understanding of why there were variations of performance between the sites:

> *"The data collected in the included studies did not provide the information needed to understand and explain the findings. To understand how and why quality improvement collaboratives work it is necessary to look into the "black box" of the intervention and to study the determinants of success or failure"*
>
> Schouten (2008)

Proponents of RCTs and many quasi-experimental designs argue that all you need to know is that it works, or that it works in a number of settings. This is countered by the argument that, for CSIs, as opposed to discrete prescribed interventions like medications, you may need to know why it works, and how it depends on or interacts with its environment, in order to reproduce it. These points are discussed in a later section on "the causality debate".

Figure 1: Common shortcomings of evaluations of complex social interventions

Implementation assessment failure—The study does not examine the extent to which the program was actually carried out. Was the intervention implemented fully, in all areas and to the required “depth”, and for how long?

Pre-study theory failure—The study does not adequately review previous empirical or theoretical research to make explicit its theoretical framework, questions, or hypotheses.

Outcome assessment failure—The study does not assess a causal chain of outcomes or a sufficiently wide set, such as short and long term impact on providers, organizations, patients, and resources.

Outcome attribution failure—The study does not establish whether the outcomes can unambiguously be attributed to the intervention or list and assess other possible influences on outcomes.

Explanation theory failure—There is no theory or model that explains how the intervention caused the outcomes and which factors and conditions were critical.

Measurement variability—Different researchers use very different data to describe or measure the complex social intervention process, structure, and outcome. It is therefore difficult to use the results of one study to question or support another or to build up knowledge systematically.

Øvretveit & Gustaffson (2003)

Questions beyond “does it work?” effectiveness questions

The above and other discussions of experimental evaluations of complex interventions emphasize the need for better descriptions, but also a better understanding of why there may be differences in outcomes between sites, and of the influences through which the intervention may have its effects (Schulz et al. 2010, Glasgow et al. 2003, Fan et al. 2010, Glasziou et al. 2008). The “explanation criticism” is unfair because experimental designs are not often designed to answer questions other than the “does it work?” effectiveness question. However, these discussions indicate a need to answer other questions beyond the effectiveness, questions such as:

<u>Description: what are the details of intervention and context?</u>

- What are the details of both the content of the intervention-change and the steps and actions to implement it, so that we could reproduce it in our setting?
- Are there factors apart from the intervention which were helpful or necessary to carry out the intervention-change and which we need to know about to reproduce it in our setting?

<u>Implementation: how was it implemented?</u>

Adaptation methods, facilitation, researcher role, and progress monitoring feedback

- Are the methods to adapt the intervention to the setting described and reproducible in our setting, if adaptation is needed?
- What was the role of researchers or others in assisting implementation and are additional specialists needed to reproduce the intervention in our setting?
- Is monitoring data on implementation and results necessary to adapt and check the intervention impact, and if so which data, and can they be collected and reported easily and at low cost in our setting?

<u>Attribution: how certain can we be that the intervention caused the outcomes reported?</u>

These relate to the "does it work?" effectiveness question but ask further questions about benefits to patients, if these are not reported, and to whether other possible influences could have caused outcomes even if there were controls:

- Are there controls for ensuring implementation is carried out as described, and for excluding influences other than the intervention on the outcomes?
- How certain can we be of positive outcomes for patients if the outcomes are process of care outcomes which are "thought to ensure improved patient outcomes".

<u>Generalization: can we copy it and get similar results?</u>

- Do we have enough details to copy it and do we have the resources to do so?
- If we did copy it could we expect similar outcomes or are there context influences which would affect the outcomes in our setting?

<u>Usefulness: are the intervention and implementation feasible?</u>

- Are the costs to implement, operate and maintain the intervention-change estimated, and is it economically feasible for us?
- Can we reproduce the resources and necessary conditions for implementation and sustainability in our setting?

In summary, criticisms have been of failure to apply rigorous experimental design, limited descriptions, and no explanations for variations in outcomes between units receiving the intervention. Innovation in research methods is needed to improve descriptions, explanation, and prediction of results in different settings.

2.2. Challenges for experimental evaluations of complex interventions

Experimental evaluations focus on answering the, "is it effective?" question. Usually these designs concentrate on a few measures of effectiveness and collect these data "before" and "after" the intervention to assess if the intervention made a difference to these measures. The designs use different methods to assess how certain we can be that difference between the "before" and "after" data are caused by the intervention rather than by something else. These designs include the randomized controlled trial design, matched comparison trials, pragmatic trials, "before" and "after" evaluations with no comparisons, and other designs (MRC 2000, Craig et al 2008, Tunis et al 2003, Mercer et al 2007, Fan et al 2010). Details of each are given later in this document.

Researchers using these designs to answer the “is it effective?” question face the following challenges, which are not unique to complex interventions but may arise in a more extreme form:

- **Describing the intervention** and the different components, including how much of the implementation actions to include in the definition of the intervention:
    - For an intervention to an ICU to prevent central line blood stream infections (CLABSI), is the intervention only the changes to clinical practice, or does it also include changes previously made to the ICU which make it easier consistently to follow the new clinical practice, such as a clinical unit safety program (Pronovost et al 2008, AHRQ 2011)?
    - A description is needed for others to understand what the intervention was and to reproduce it if they choose to. Trials of simple clinical interventions have also been criticised for not giving sufficient descriptions (Begg et al 1996; Schulz et al 2010).
    - If it is essential for implementers to adapt an intervention it in an undefined way, how to document the adaptation? Physicians adjust dose according to trial protocol, but implementers have far less precise guidance and greater latitude in how to adapt some CSIs.
- **Deciding which outcomes** would indicate effectiveness, and how to capture any indications of unexpected and unwanted outcomes:

    Which possible effects of a CLABSI intervention to an ICU can be measured or documented, and which are most relevant to the evaluation user’s decisions? Is there evidence that intermediate or process data, such as orders for medications for treating bloodstream infections, are reliable and valid indicators of effects of the intervention? Should the evaluation collect data about costs and possible savings? If the intervention is at multiple levels, which outcome data should be collected? Should data on many outcomes be collected or should researcher-resources be concentrated on collecting fewer outcomes with more validity and reliability?

    The “causal chain” from intervention to patient or cost outcomes may be through a number of intermediate stages. This requires that outcomes are collected at each stage of the causal chain to trace final impact, as shown in the example study 4 described in the appendix (Nazareth et al 2002). Or can the evaluation rely only on evidence from other research that achieving an intermediate outcome is predictive of likely impact on later outcomes?
- **Randomization** of subjects to receive the intervention and its comparison may be difficult when these are individual providers, service teams or units or health systems.

    If the number of possible providers or units is small (e.g., 30 or fewer), randomization does not distribute sufficient numbers to “experimental” and “control” groups to allow sufficient certainty about whether outcome differences between the groups are due to the intervention or something else.

- Alternatively, **matching** intervention and comparison groups to reduce confounding influences also may be difficult. This is especially so if there is no evidence or theory about which characteristics of providers or services may make them more or less susceptible to influence by the intervention.

Generally, an evaluation study or an evaluation design can only answer one or a few questions, and its validity is to be judged in terms of how well it answers these questions. A randomized controlled trial (RCT) is often the best of designs for answering the question, "what were the effects of this intervention on the recipient in terms of these measured variables…?" It is the worst of designs for answering questions such as, "how did the intervention evolve over time and which factors influenced this evolution?" Thus the shortcomings of an evaluation are in relation to the question it set out to answer.

2.3. Challenges for evaluations of complex social interventions

Critics of the experimental approach to evaluation have proposed that, in addition to the challenges noted above, there are further challenges which arise because the "social" nature of the intervention is such a central part to it (Greenhalgh et al. 2004, Dixon-Woods et al. 2011, Ferlie et al. 2005, Øvretveit 2011a). These criticisms arise from conceptualizing CSIs as,

- a series of actions carried out by human beings, with varying intensity and quality,
- to influence other human beings, who have choice, and are exposed to competing influences,
- in a changing social, economic and political situation which affects people's actions and responses.

The actions of a pharmaceutical or surgical treatment can be described as a "mechanism" which operates according to natural laws of physiology, which are largely independent of what a person thinks about it. Actions such as education, feedback, financial incentives, or directives depend for their effects on how individuals interpret the meaning of the action. A directive to nurses to wear a sign "do not interrupt while preparing and dispensing medications" was not used by some, who interpreted this as an insult to their intelligence, and another action by management to be resisted because it disrespected their professionalism.

There is a regularity in how people, groups and organizations interpret the intervention actions to influence them, and in the outcomes of the intervention because individuals share group norms and cultures. However, the outcomes are still less predictable than those from a mechanical or physiological action on a human body.

In addition, the social world of which the individuals and groups are part of changes in a way in which the human body does not, and in ways which are significant for how the implementation actions are interpreted. A series of articles in a nursing journal about "do not interrupt" signs, and a recommendation by a safety association persuaded more nurses that this was good nursing practice and not an insult to their intelligence and professionalism.

There are often many outcomes which are of interest to one evaluation user group or to different stakeholders. Often qualitative data about outcomes is the only feasible data collection strategy, such as documenting informed observers' assessments of whether the outcome was achieved and the evidence they give for their assessments.

Complex social interventions often comprise of a series of actions, as for example in a health promotion campaign for "healthy lifestyle" (e.g., HealthyPeople.gov 2011). The desired effects depend on the intervention actions being carried out, as well as the social conditions which support the actions: the degree to which the actions can be fully and consistently carried out depends on the conditions surrounding the actors, such as higher level support, resources, time, knowledge, and skills. The effects of these actions also depend on conditions which help and hinder the actions to have the desired results (e.g., media or cultural influences which contradict or complement the actions and which interact with the actions). The sustainability of such interventions and their results depends even more on the surrounding conditions – short-term evaluations are less sensitive to the role of surrounding conditions and how they change. To understand which conditions are necessary for implementation and desired effects, it does not always help to "control these out" as "confounders". It may be better to understand how they have their influence.

As a result of these considerations, some evaluation researchers seek to answer questions other than the, "is it effective?" question, and use methods which seek to understand how the intervention "causes" its effects. These methods give more recognition to the social nature of the intervention.
This alternative way to conceptualize CSIs leads some evaluators who hold these ideas to be reluctant to use the term "intervention", which to them suggests a discrete isolatable force, aimed at "targets", viewed as objects in their response. Instead subjects are viewed as "taking up" the "intervention" resources or ideas offered, if they choose to do so, and creatively applying it in their setting. "Intervention" is regarded more as "facilitated uptake" by active subjects who may then make innovations to the intervention which can make it more or less effective than it would otherwise be in their setting. Indeed "evaluation" is viewed by some using this idea as a less useful term than enabling continuous learning for improvement.

2.4. Challenges summary

Some CSIs can be prescribed, implemented with fidelity to the prescription and evaluated using experimental designs and controls. However, such designs may not be appropriate for other CSIs and for other questions because:

- Replication of a complex intervention may be more difficult elsewhere: requirements for standardization and control needed in the study design may be difficult to achieve in routine practice.
- Prescription of the intervention components may not be appropriate. Implementation often requires implementers to revise the intervention in ways which cannot be prescribed, but guidance can be given for methods for making adaptations. Description of the changes and

methods used in practice is thus necessary, which may call for refocusing the evaluation on other outcomes than those originally planned.

- The time taken for the study may be too long for practice and policy, especially for evaluations of changes which involve rapidly-developing technologies.
- To inform their actions, stakeholders have a need for information in addition to "is the intervention effective for producing these selected outcomes?"
- For these and other reasons, observational and action evaluation designs are needed, but these have their shortcomings and are not as well understood as experimental designs.

3. Example Evaluations

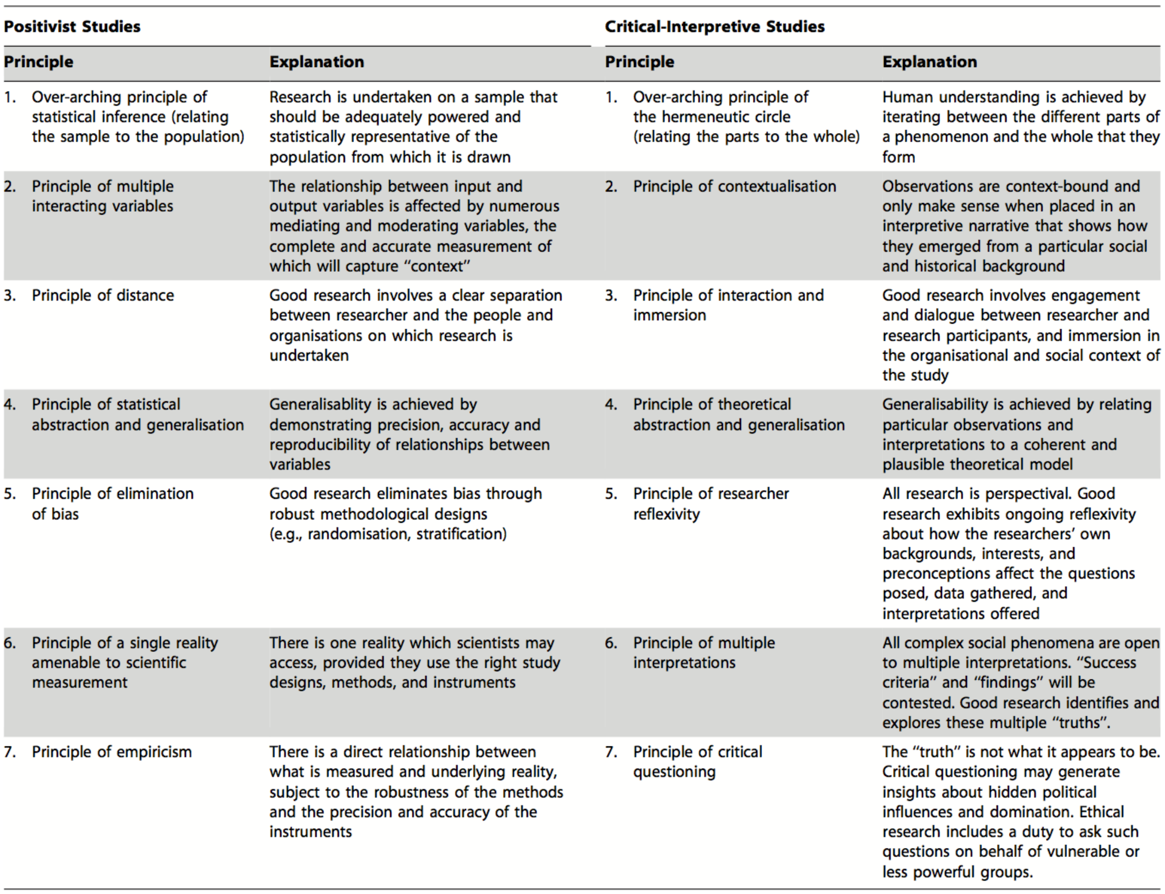
Appendix 2 of this document presents evaluations of CSIs to illustrate a number of contrasting designs for answering different questions. The summaries of these evaluations present the question addressed, the intervention evaluated, the design, the outcomes measured and discovered, the certainty about and question addressed, the intervention evaluated, the design, the outcomes measured and discovered, the certainty about and question addressed, the intervention evaluated, the design, the outcomes measured and discovered, the certainty about and The eight example studies are:

1. RCT of a care transitions intervention
2. Pragmatic Cluster RCT of a multifaceted intervention in ICUs
3. Interrupted time series evaluation of a hand hygiene intervention
4. Prospective observational evaluation of an intervention to change primary care physicians' behavior (mixed methods)
5. Prospective observational evaluation of large scale safety program (mixed methods)
6. Process evaluation of an intervention to promote smoking cessation
7. Case evaluation of a program for electronic summaries of patients' medical records (mixed-methods)
8. Realist evaluation of large scale "transformation" of health services in London

3.1. The eight example studies in appendix 2 reflect different perspectives and paradigms in evaluation

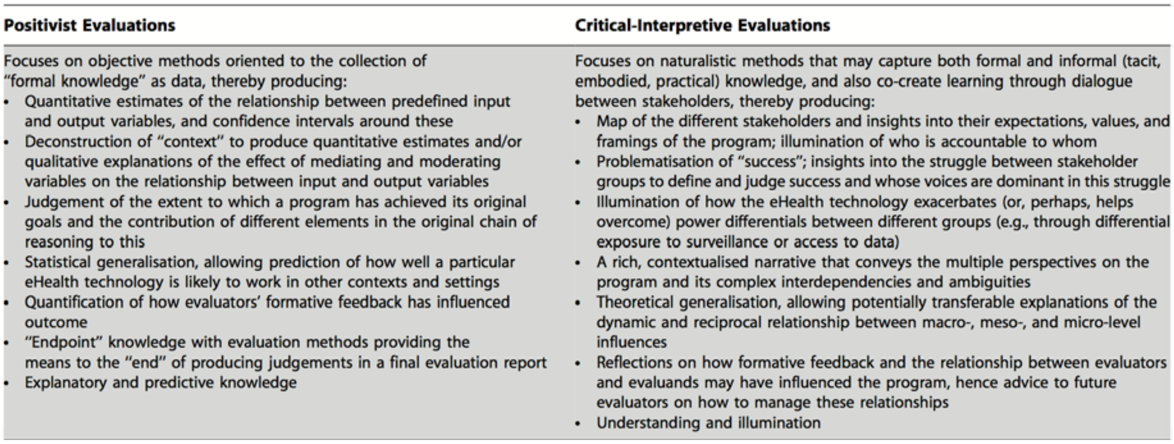
Differences between designs to some extent reflect different perspectives and paradigms in evaluation, most notably a contrast between positivist and interpretivist perspectives. This is summarised by Greenhalgh & Russell 2010, following Klein & Myers 1999, in the following way:

Table 1. Comparison of Key Quality Principles in Positivist versus Critical-Interpretivist Studies.



Positivist Studies		**Critical-Interpretive Studies**	
Principle	**Explanation**	**Principle**	**Explanation**
1. Over-arching principle of statistical inference (relating the sample to the population)	Research is undertaken on a sample that should be adequately powered and statistically representative of the population from which it is drawn	1. Over-arching principle of the hermeneutic circle (relating the parts to the whole)	Human understanding is achieved by iterating between the different parts of a phenomenon and the whole that they form
2. Principle of multiple interacting variables	The relationship between input and output variables is affected by numerous mediating and moderating variables, the complete and accurate measurement of which will capture "context"	2. Principle of contextualisation	Observations are context-bound and only make sense when placed in an interpretive narrative that shows how they emerged from a particular social and historical background
3. Principle of distance	Good research involves a clear separation between researcher and the people and organisations on which research is undertaken	3. Principle of interaction and immersion	Good research involves engagement and dialogue between researcher and research participants, and immersion in the organisational and social context of the study
4. Principle of statistical abstraction and generalisation	Generalisablity is achieved by demonstrating precision, accuracy and reproducibility of relationships between variables	4. Principle of theoretical abstraction and generalisation	Generalisability is achieved by relating particular observations and interpretations to a coherent and plausible theoretical model
5. Principle of elimination of bias	Good research eliminates bias through robust methodological designs (e.g., randomisation, stratification)	5. Principle of researcher reflexivity	All research is perspectival. Good research exhibits ongoing reflexivity about how the researchers' own backgrounds, interests, and preconceptions affect the questions posed, data gathered, and interpretations offered
6. Principle of a single reality amenable to scientific measurement	There is one reality which scientists may access, provided they use the right study designs, methods, and instruments	6. Principle of multiple interpretations	All complex social phenomena are open to multiple interpretations. "Success criteria" and "findings" will be contested. Good research identifies and explores these multiple "truths".
7. Principle of empiricism	There is a direct relationship between what is measured and underlying reality, subject to the robustness of the methods and the precision and accuracy of the instruments	7. Principle of critical questioning	The "truth" is not what it appears to be. Critical questioning may generate insights about hidden political influences and domination. Ethical research includes a duty to ask such questions on behalf of vulnerable or less powerful groups.

Table 2. Different Kinds of Knowledge Generated by Different Kinds of Evaluation.



Positivist Evaluations	**Critical-Interpretive Evaluations**
Focuses on objective methods oriented to the collection of "formal knowledge" as data, thereby producing: • Quantitative estimates of the relationship between predefined input and output variables, and confidence intervals around these • Deconstruction of "context" to produce quantitative estimates and/or qualitative explanations of the effect of mediating and moderating variables on the relationship between input and output variables • Judgement of the extent to which a program has achieved its original goals and the contribution of different elements in the original chain of reasoning to this • Statistical generalisation, allowing prediction of how well a particular eHealth technology is likely to work in other contexts and settings • Quantification of how evaluators' formative feedback has influenced outcome • "Endpoint" knowledge with evaluation methods providing the means to the "end" of producing judgements in a final evaluation report • Explanatory and predictive knowledge	Focuses on naturalistic methods that may capture both formal and informal (tacit, embodied, practical) knowledge, and also co-create learning through dialogue between stakeholders, thereby producing: • Map of the different stakeholders and insights into their expectations, values, and framings of the program; illumination of who is accountable to whom • Problematisation of "success"; insights into the struggle between stakeholder groups to define and judge success and whose voices are dominant in this struggle • Illumination of how the eHealth technology exacerbates (or, perhaps, helps overcome) power differentials between different groups (e.g., through differential exposure to surveillance or access to data) • A rich, contextualised narrative that conveys the multiple perspectives on the program and its complex interdependencies and ambiguities • Theoretical generalisation, allowing potentially transferable explanations of the dynamic and reciprocal relationship between macro-, meso-, and micro-level influences • Reflections on how formative feedback and the relationship between evaluators and evaluands may have influenced the program, hence advice to future evaluators on how to manage these relationships • Understanding and illumination

4. Choices for research design for evaluating CSIs

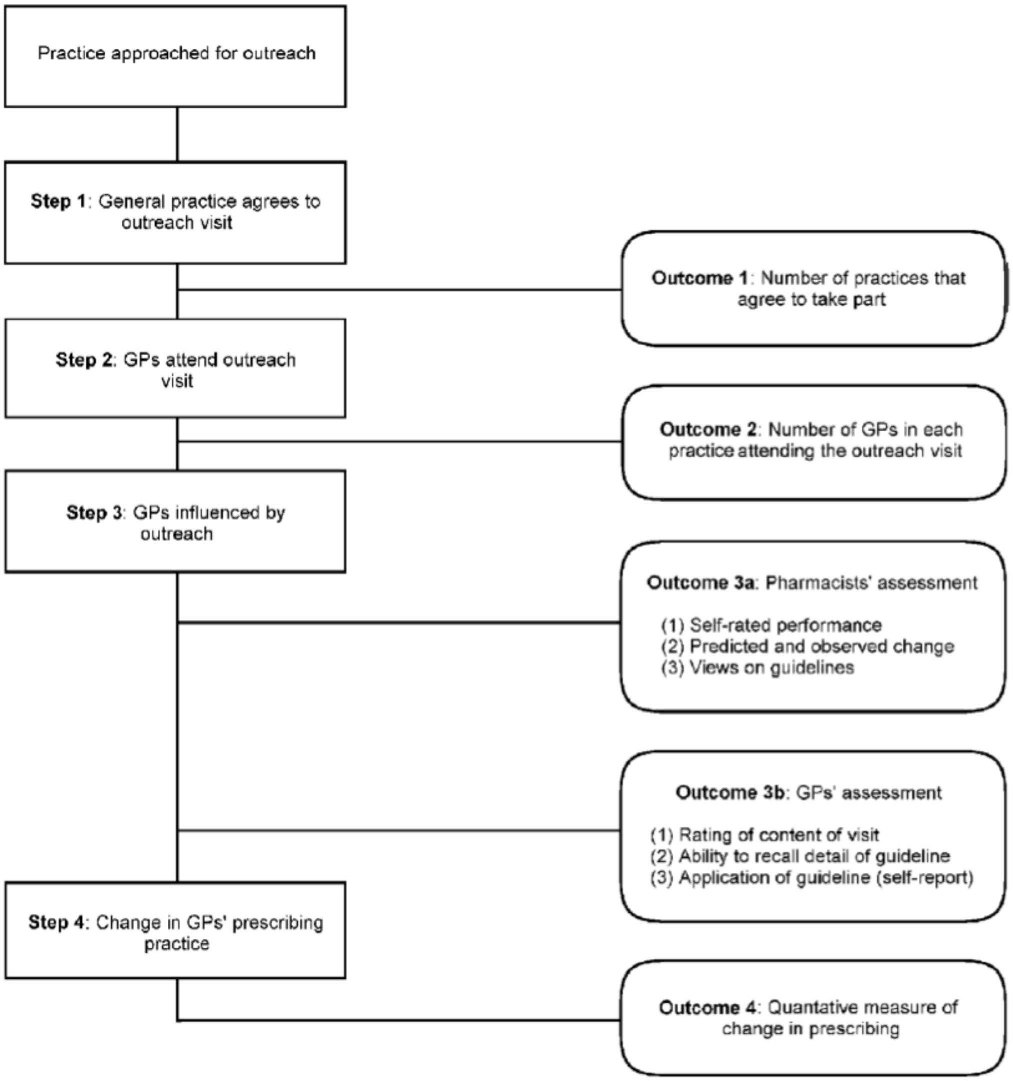
This section describes different designs for answering different questions for evaluating different types of intervention-change or different services or policies. It describes issues in deciding which design to use. The guidance which follows from this, about practical steps for carrying out an evaluation, are given in the second volume companion document, which also give tools and resources for carrying out each design. Appendix 2 summarises eight example studies and gives a commentary of the strengths and weakness of each.

4.1. What is "evaluation design"?

The purpose of an evaluation design is to show which data will be collected and when, and to show the comparisons to be made to answer the evaluation question. A design shows the logic of the evaluation, not the logistics, which are the more detailed practical timetabling of who does what, and when. The design does not show details of data collection methods, and the comparison can be simple, such as before and after the intervention – it does not always involve a comparison group.

The evaluation design shows in outline the people or organizations exposed to the change or the service which is to be evaluated, which data will be collected and when, and the time perspective of the study: whether it is a prospective-planned, contemporaneous, or a retrospective evaluation.

Designs are best represented as a time diagram picture: for example the picture below shows time vertically in one of the study designs summarised in the appendix, of an evaluation of an intervention by pharmacists visiting primary care centers to improve physicians' guideline compliance. The picture shows the timeline starting at the top and going down the page, and shows the outcome data to be collected – one thing missing is the times at which these data will be collected.


Practice approached for outreach
Step 1: General practice agrees to outreach visit
Step 2: GPs attend outreach visit
Step 3: GPs influenced by outreach
Step 4: Change in GPs' prescribing practice
Outcome 1: Number of practices that agree to take part
Outcome 2: Number of GPs in each practice attending the outreach visit
Outcome 3a: Pharmacists' assessment
(1) Self-rated performance
(2) Predicted and observed change
(3) Views on guidelines
Outcome 3b: GPs' assessment
(1) Rating of content of visit
(2) Ability to recall detail of guideline
(3) Application of guideline (self-report)
Outcome 4: Quantative measure of change in prescribing

Pathway of change in general practitioners' (GPs') prescribing practice.

Other ways to draw a picture of design are shown in the Volume 2 guidance.

4.2. Approaches to evaluating CSIs

Three main approaches

An overview of the literature on evaluations of CSI shows that there are three broad approaches, which are each discussed in more detail after this overview:

1) Experimental and quasi-experimental

In this family of designs, the change or service evaluated is thought of as an experiment, and the change is planned before it is introduced. Data gathering methods are piloted and then "baseline" data collected before the intervention-change and then later. Usually the data is quantitative data, from measures carefully selected to indicate any outcomes which are related to the evaluation user's questions and decisions (Campbell et al. 1966).

In this document, prospective time-series designs are classified as quasi-experimental because the intervention and data gathering are planned ahead as if the intervention were an experiment. Some other discussions classify time-series designs as "observational designs" when they do not have a comparison group that is not exposed to the intervention. Also, many researchers do not consider PDSA quality improvement testing as a research evaluation design, but this document here includes these as quasi-experimental designs (see also Speroff & O'Connor 2004, and Morrison et al 1996).

These and other designs are discussed in more detail below and in the Volume 2 guidance.

2) Observational

The design and intervention The design and intervention The design and intervention The design and intervention The design and intervention This approach is used in many evaluations termed "formative", "process" "case" and "program" evaluation. Some designs in this category are called cohort, case control and cross-sectional, usually if the data collected are quantitative in nature.

The fact that a pre-planned intervention experiment is not used does not mean that studies using this approach do not use other scientific techniques, like formulating and testing hypotheses about the possible effects of the intervention or service on the subjects about which data are gathered. Data can be collected through qualitative or quantitative methods or both: mixed methods approaches are increasingly used for evaluating CSIs.

3) Action approaches to evaluation

The main difference from the other two approaches is that the researcher-evaluator does not try to minimize the effect of the research on the intervention or service studied (which would increase generalizability of the findings). Rather the researcher-evaluator tries to help improve the service or intervention as it is implemented. Typically the researcher will present early findings to implementers and sometimes help them to design and redesign the intervention (Øvretveit 2002, 1998). In some respects, formative evaluations are action evaluations, but in the latter the evaluator continues to help improve the intervention using the findings from research and has a stronger continuing collaborative- and participatory- role with the implementers.

Action evaluations are not distinguished by the data collection methods they use – all three approaches can use data gathered quantitatively, qualitatively and mixed data collection methods.

> *Combining many design elements into a single study can produce evaluations with high validity and strength of causal inference as good as, and sometimes better than, randomized trials.*
>
> Wagenaar & Komro (2011), referring to "natural experiments".

4.3. What are the different designs for a CSI evaluation?

The research review for these two volumes found a range of designs are used to answer different evaluation questions. The most common designs are briefly summarised below. More details about each, and when and how to use them are given in the Volume 2 guidance. Appendix 2 summarises the following eight example studies and gives a commentary of the strengths and weakness of each.

1) Experimental and quasi-experimental

Comparative experimental

These designs plan and implement a defined intervention-change to "intervention units" which could be to patients (e.g., the CSI is a combination of services to patients with a chronic illness), or to providers (e.g., a combination of training and reminders to physicians to encourage compliance with clinical guidelines), or to service units (e.g., financial incentives and education to reduce hospital acquired infections) or a combination of interventions to patients, providers and service units. The defining feature is that some units receive the intervention change and some do not (the "comparison" or "control" group). Outcomes data are from measures selected to show the effects of the intervention. These data are collected before exposure ("baseline"), and then at one or more times after exposure, or "later" if the intervention is sustained for some time.

Planning of, and controls in implementation (such as a trial protocol which is "enforced" in different ways) try to ensure each unit is "exposed" to a standard intervention which is as similar as possible across the units. The aim is to reduce possible variations between units in the "dose" or "frequency" of the intervention, so as to maximize the certainty of attribution to the intervention of any of the before/later differences in the outcome data which are documented by the study. It also makes it possible to describe the intervention clearly and provide others with details of how to reproduce the intervention.

The varieties of this design use different strategies to design out other possible influences, apart from the CSI, on the data collected to assess outcomes:

> Randomized controlled trial (RCT): Units are randomly allocated to be exposed to the change-intervention or some comparison intervention (or to none at all). Randomization in sufficient numbers reduces the possibility of there being differences between characteristics of the "intervention group" and the "comparison group," which would make conclusions less certain about whether the outcomes were caused by the intervention rather than by a characteristic of the exposed units. The example studies 1 and 2 in the appendix illustrate how this design was used to evaluate two different CSIs).

Non-randomized comparative trial: Units receiving the intervention and comparison units are "matched" as far as possible for characteristics which might affect how they respond to the intervention. For example, if the patient is the unit being intervened upon, the evaluators may try to assign patients to both groups which are similar in age, sex, education and income. However, the group receiving and not receiving the intervention may be different in other ways, which affects outcomes which are not matched, and this adds an influence other than the intervention which confounds (confuses) the findings (giving a "confounding to the findings (Concato et al. 2010 and Concato & Horwitz 2004).

Cross-over comparative trial

In this variation, one group gets the intervention, while the other group gets none or another intervention, then the intervention is stopped for this group and started for the other group (e.g., Devon et al 2005). This may be suitable for CSIs where it is thought the effects of the CSI on the unit being intervened upon decays rapidly after the intervention is stopped.

Stepped wedge trial

In this, the units are often in the same organization (e.g., hospital units or wards). First one unit is exposed to the intervention, then, after a period, an additional unit is exposed so that the two units are now receiving the intervention. Then another unit is added, and so on, until all are exposed. Those not receiving it act as controls, not only because they do not receive the intervention, but also because both they and the intervention units are exposed to changes other than the intervention and which may affect outcomes, such as a hospital or national campaign on safety, which might not have been predicted when the trial was started (Brown & Lilford 2006). Mdege et al. (2011) propose this design for interventions that have been shown to be effective in more controlled research settings, or where there is lack of evidence of effectiveness but there is a strong belief that the intervention will do more good than harm; however, the authors emphasize the need for consistent data analysis and reporting across the units and through time.

Non-comparative quasi-experimental

Before-After design (or, more often, Before-"Later")

There is no comparison group: only one group of units is exposed to the intervention. The design collects data about features of the units which the intervention is intended to change, first before exposure, then later or after (e.g., providers' knowledge before and after exposure to education, or patient infection rates before and after patient (or provider) exposure to education. The design involves listing other possible explanations for the before/after (or later) data differences and assessing their likely influence, but without comparison groups these explanations cannot be excluded.

Simple time series, or interrupted time series design:
The simple time series design is like a before-after design but there are many before- and many after- data-collection time points for the outcome of interest. The idea behind the design is that a series of data points before show an average level over the period (e.g., data collected every month about infection rates for 12 months). Then the intervention is introduced and the impact on the outcome data is assessed over the subsequent data points (e.g., there is a significant rise (or none) each month after, for a period of time). The general trends are assessed to see if the intervention changed the outcome of interest.

If enough time data points are collected in the right way, then statistical process control (SPC) methods can be used to define upper and lower control limits and to identify any special causes (such as the intervention) which significantly change the process (Carey & Lloyd 1995, Wheeler 1993).

Although there is no comparison group, the attribution of the change to the intervention is made more certain than for a simple before-after design by the series of data points over time which can show any “significant” change in trend. The time trends are also useful to see if any outcome is sustained. For example, one training event may show a sudden change to the outcome data but reduce to the earlier levels after a few weeks.

The interrupted time series design may be suitable for evaluating some CSIs: in this the CSI is introduced, and then stopped for a series of outcome data points, so as to assess whether the outcome data levels return to the average before the CSI. The interruption can be repeated, to increase certainty about whether the CSI really does affect the outcomes. The study example 3 in the appendix uses this design.

Quality improvement testing (PDSA)
Many quality improvement interventions are complex social interventions to provider behavior, care process or organisation. The plan – do – study – act cycle (PDSA) is an evaluation technique used in quality improvement to study whether a planned change, when implemented, has an effect on the outcomes of interest (Langley et al 1996). If this testing cycle is applied with a certain rigor, and possibly also using a time series design, then some researchers view it as allowing a reasonable degree of certainty about attribution of outcomes to the intervention in real-world settings, even without a control group. The design allows for, indeed requires, repeated changes to the intervention to check if revisions can improve the apparent impact on outcomes (“iteration testing” for “continuous improvement”).

Certainty of attribution for evaluation can be enhanced by using the approach described by Speroff & O’Connor 2004: formation of a hypothesis for improvement (Plan), a study protocol with collection of data (Do), analysis and interpretation of the results (Study), and the iteration for what to do next (Act). Different designs can be used within this framework but the time series design, with or without statistical process control methods, is the most common.

2) Observational

The intervention-change is not planned and introduced as an experiment with careful controls, but it and the outcomes are observed. This is often done retrospectively or concurrently with little time to plan, but sometimes the evaluation is planned and prospective before the intervention-change starts. These designs are used when controlled implementation is difficult or unethical, or when little time and resources are available or for other practical reasons – sometimes these designs are called "naturalistic evaluations" where the intervention or service to be evaluated is studied "in the wild," often as it "evolves" in its "environment."

The first sub-category are observational evaluations which collect quantitative data - the terms cohort, case-control and cross-sectional are usually applied to such evaluations, but these terms can be applied to describe evaluations using qualitative data as the terms describe the design, not the data collection method.

Quantitative: Cohort-, case control- and cross sectional- evaluation designs

Observational evaluation designs using quantitative data are usually termed "cohort" evaluations when they collect data about subjects exposed to the intervention at different times, sometimes prospectively planned. "Cross-sectional" designs are where data is collected at one time, looking across a range of subjects are usually retrospective: subjects showing the outcome of interest are matched with a control group who do not show it, and the researcher looks back to see which subjects were exposed.

<u>Cohort evaluations:</u> units are chosen, either on the basis of who will receive or have received the intervention. An alternative is to choose on the basis of high or low performance on an outcome, where the evaluation is seeking to assess what the effect of a variety of influences on the outcome may be (e.g., "why are these units performing so much better than others?"). The evaluator then measures a variety of variables that might influence the outcome. Over a period of time the units in the sample are observed to see whether or how they develop the outcome of interest (e.g., show high performance on some indicators), and statistical analyses are used to assess which variables are associated with these outcomes.

An advantage of the cohort design is that it can be used where retrospective evaluation is required or where RCTs are unethical or impractical. It is also useful for exploring different hypotheses about the strength of different influences on outcome – if data about these variables are available. The latter points to one of the limitations of retrospective cohort evaluations of CSIs – quantitative data may not be available for some possibly important influences, and informed observers' assessments are often the only data which can be gathered. Another is that if two groups are compared it is difficult to control for all other factors that could influence outcomes.

Cross-sectional studies: this design collects data at one time point across a range of units, and is usually used to assess prevalence (e.g., how many personnel received the training?).

Case-control evaluation: these are usually retrospective. People or units with the outcome of interest are matched with a control group who do not show it. Then the evaluation assesses whether they were exposed to the intervention or other influences which may explain the difference (Mann 2003, see also Dreyer et al. 2010 on observational studies for comparative effectiveness research).

Major difference underlying the designs

Experimental designs focus on whether the intervention causes changes in a few measurable outcomes of interest, and use different methods to exclude "confounders." Naturalistic approaches examine more variables or influences but are not able to establish the same type of certainty about associations, and sometimes conceive of the intervention as part of a system of influences.

Naturalistic approaches do not standardize the intervention but describe how it is implemented and how different local and broader factors affect implementation. They also document a variety of intermediate outcomes and impacts. As such they are suited to questions about impact and implementation (including adoption, spread and sustainability).

Qualitative or mixed methods observational evaluation designs

The second sub-category of observational designs are those which collect qualitative data or use mixed data collection methods, sometimes called "naturalistic approaches" to evaluation. The designs use specific techniques to maximize internal and external validity. This group of designs and these techniques are perhaps less familiar to medical- and health services researchers but have a long history with social scientists, international health researchers and in health promotion/education and public health research, as well as for educational, social work, mental health and welfare program evaluators (WHO 1981, Shadish, et al 1991, Greene 1993, JCSEE 1994, Owen & Rodgers 1999)

The main sub-categories are: single case evaluation, case-comparison evaluation, and the more recent realist evaluation, each with different designs within these sub-categories.

Single case evaluation

The single case usually refers to the "unit" receiving the intervention, which may be defined at different levels: one team, or one micro-system, or one organizational unit or system, or a region, and sometimes a nation, as in "evaluation of X health reform".

Take an example intervention to improve hand hygiene: a combination of online training, assistance to build a measurement system for feedback to providers, and financial incentives. This intervention may only be directed at a specific level (e.g., one nursing unit), and the evaluation considers its impact on the unit and other outcomes, as well as the surrounding "context" influences which may affect implementation of the intervention (e.g., top management support, existence of IT systems which can be used for easier collection and feedback of data (Alexander et al. 2006). Or the same intervention may be directed at more than one level (e.g., a state-wide intervention, run by a state-wide unit) in which case the "context" is national and local factors which may affect "uptake" of the intervention by the lower levels in the state.

Sometimes "the case" is the intervention, such as an intervention to patients, for example a special stroke service with multiple components which differ from "usual care". The single case evaluation describes the data collection times and methods and validity techniques used to evaluate the stroke service, but without a comparison to "usual care".

Some formative-, process-, outcome-, or program evaluations use this design. The appendix gives examples of three types of single case observational designs (Examples 4, 6, 7):

- Prospective observational evaluation of an intervention to change primary care physicians' behaviour (mixed methods)
- Process evaluation of an intervention to promote smoking cessation
- Case evaluation of a program for electronic summaries of patients' medical records (mixed-methods)

Other examples are provided by Walshe & Shortell (2004), Øvretveit & Aslaksen (1999), and Ovretveit et al. (2012).

Keen & Packwood (1995) describe case study evaluation (CSE) as,

> *"valuable where broad, complex questions have to be addressed in complex circumstances…when the question being posed requires an investigation of a real life intervention in detail, where the focus is on how and why the intervention succeeds or fails, where the general context will influence the outcome and where researchers asking the questions will have no control over events. As a result, the number of relevant variables will be far greater than can be controlled for, so that experimental approaches are simply not appropriate."*

Single CSE is useful where an intervention is not well defined or standardizable and cannot easily be distinguished from the general environment, or where it involves a number of components, each of which may change over time. Rather than assessing efficacy, CSE's are more useful for describing and explaining implementation, a variety of intermediate outcomes, and for addressing questions concerning adoption, spread and sustainability. For evidence and for explanations of intermediate and patient outcomes, many CSEs use triangulation of data, which often include informants' estimations about CSI outcomes. The degree of certainty about whether the CSI results in intermediate or final outcomes is less than for experimental designs because confounders and competing explanations are not controlled for and cannot be excluded. However, if other influences apart from the intervention are recognized and their influence understood (e.g., context factors) a case study design can allow the development of knowledge about how both intervention and context contribute to intermediate and final outcomes, as well as knowledge about implementation.

Case evaluation designs vary in the type of pre-data collection work which different designs undertake. Some use minimal pre-planning of which data to collect, and focus on documenting and describing the changes planned and made as the intervention is implemented, before then deciding which data to collect to discover different types of intermediate and later outcomes, or to check if intended outcomes are achieved. More often, and increasingly, case evaluations use theory to define which data to collect about which outcomes (Bickman 2008). This can be a theory of how actions in the program will lead to certain results, which allows the evaluators to identify which data to collect to assess if the actions are carried out, and which indicators or data may show intended or theorized outcomes at different stages (the program theory or logic model of the intervention (Shekelle et al. 2010)). The appendix summarizes the different types of theory which can be used.

Case comparison evaluation

This is like the above-noted "cohort" design, but uses the validity-enhancing strategies for qualitative data and mixed methods of the single case evaluation just mentioned, such as triangulation and program theory. The appendix gives as example 5 an evaluation of a large scale safety program which compared each of four hospitals receiving the program with other hospitals not receiving the program (Benning et al. 2011). Another example is a comparison of two case hospitals, each of which received a CSI to implement an electronic medical record (Øvretveit et al. 2007). This case comparison evaluation used previous research into similar interventions to pre-define which data to collect to test hypotheses from earlier studies about which aspects of the intervention were necessary for effective implementation.

Realist evaluation

These designs identify context-mechanism-outcome (CMO) configurations in complex interventions in different settings, and aim to establish "what works for whom in which settings" (Pawson & Tilley 1997). The assumption, based on some evidence from education and criminal programs, are that, in social interventions, outcomes are a function of both "the mechanism" through which the intervention works and the context in which it is applied: there are different effects in different settings even if the same intervention is used. The appendix gives one example of an evaluation of a large scale "transformation" of health services in London which uses this design, and describes its strengths and weaknesses.

The aim is not only to describe the intervention but to clarify the "generative mechanism": the essential idea or "active ingredient" which is the basis for the intervention (e.g., performance feedback). Using this approach, superficially different interventions can be grouped and compared through their underlying logic. Another aim is to examine how much and how the mechanism depends on, or interacts with, the context to produce different effects.

The design uses a program model or theory (sometimes called a logic model) to select programs and test hypotheses about CMO configuration for one intervention in one setting. Then the design studies programs theorized as "similar" in other settings to examine how the interactions between C, M, and O vary. Interventions are viewed as "theories in practice". Discovering poor or no outcomes for a similar intervention in a different setting is an opportunity to refine the logic model of the CMO configuration.

This approach thus emphasizes studying, in a variety of situations, the mechanism which is thought to generate certain results rather than one intervention at one site. The approach is similar to case study in describing and understanding outcomes as the product of an intervention implementation in context. However it differs from some case study evaluations in emphasizing the logic model testing and the comparison between different implementations, as well as elucidating the essential feature of the mechanism to allow comparison of a variety of superficially different changes.

A possible limitation of the realist approach for studying CSIs are that the concepts of context, mechanism and outcome are not easy to define and only illustrated in a few studies. It is also unclear exactly how "mechanism" is elucidated in such studies: "mechanism" does not just refer to the intervention components or implementing actions, but how this higher level conceptualization of "mechanism" is created is unclear - in terms of how the actions "work" ("generative mechanism"), which is different from their interaction with context.

Examples of studies using these methods to study safety, quality or other interventions to organizations include those by Redfern et al. (2002), Blaise & Kegels (2004), Byng et al. (2005 and 2008). Some limitations are described by Davis (2005).

Evaluating innovations in healthcare delivery and DHT (from Williams 2011)
"The complexity of innovation makes it unsuitable for experimental evaluation design (Booth & Falzon 2001, Berwick 2008). In order to capture the range of individual, group and organizational level processes and outcomes, a combination of approaches might be adopted. For example, qualitative individual reflections, evaluation of group process through action research and quality assurance in relation to organizational processes.

3) Action evaluation

Action evaluations aim to provide early feedback from the evaluation to enable CSI implementers to improve the CSI and its implementation. One assumption is that if the evaluation is useful to different parties during rather than after the evaluation, then evaluators can gain information and insights which they may not otherwise gain. By collaborating and participating in the shaping of the CSI, they may be better able to document how it has changed and why, and be better able to explain later findings. The Greenhalgh 2009 realist evaluation described above was also an action evaluation, and formative to the transformation programme evaluated. Another example of an action evaluation of a continuous quality improvement CSI in a hospital is reported by Potter et al. (1994).

One variant of an action evaluation is being used in the VA to develop and evaluate the VA version of the patient centered medical home. Related to an earlier approach, termed "evidence based quality improvement" (Rubenstein et al. 2006, 2010), this approach involves the researchers assisting primary health care personnel in a number of ways to design and implement changes (providing evidence of effective practices, training in quality improvement methods), and to report findings from the evaluation to the implementers.

The limitations of action evaluation, compared to experimental designs, are to both the internal and external validity of the evaluation, and more-so than the qualitative case evaluations described above, because of the participatory role of the researcher. These limitations are traded in this design by the greater explanatory power gained by researchers from their action researcher role and participation. One challenge is to assess the role and impact of the researchers, in order that others are able to judge whether they need a similar input if they were to implement a similar CSI, either from other researchers or internal or external experts. Sometimes action evaluation is suitable for development phases of the CSI, or for CSIs which are likely to require significant modification to be implemented. The best simple overview of action evaluation is a section by Robson (1993) and a more comprehensive overview is by Waterman et al. (2001) (see also Hart & Bond 1996, and Morton-Cooper 2000 and Øvretveit 2002).

Overlap between designs and different understandings of what the evaluation design terms mean

There is an overlap between the different observational approaches. For example, the Southampton Heart Integrated Care Project (SHIP) gathered qualitative data parallel to an RCT, and was able to explain the negative results found in the RCT and show how different beneficiaries made sense of the intervention (Bradley et al. 1999). This was titled as a case study, but could be described as a qualitative-, process- or program- evaluation and has been cited as an early example of a realist evaluation.

Paradigms underlying the above designs

The "designs" used in some observational and action evaluations are sometimes called "approaches" because "design" implies, for some, a specific standardized set of research methods and procedures. In contrast, "approaches" are frameworks for conceptualizing the intervention and how to gather data. There is more latitude for the evaluator as to which methods and how they are applied than for experimental designs. Also, some of these approaches are newer and the methods have not been discussed and described as extensively as for experimental designs. This has its strengths, and also its weaknesses, namely, it is less easy to follow a procedure to learn and apply the approach, and less easy for others to reproduce the study to see if they get the same findings – an important criterion within the dominant view of science.

The idea of "approach" is illustrated in three different strategies used in program and case evaluation. At one extreme, inductive strategies build theory about how the intervention has its effects from the data collected in the study, usually qualitative data. In contrast, deductive strategies start with theory and test or explore theoretical propositions in the study, using theory to predefine the data needed to test these propositions. There are also "snowball" strategies which draw-in and test different theories during the research about how the intervention has the effects which are being documented, building and using theory in interaction with the data gathering. In addition case evaluation and action approaches can be classified as positivist and objectivist, assuming causal mechanisms and using quantitative measures; or as subjectivist and qualitative, assuming "interventions" work through people who make choices about whether to respond depending on their interpretations, motives and values, and influenced by culture and other social factors. This is not to say there are not procedures and conventions for testing the findings and ensuring validity, just that these are different, and arguably less prescribed and more reliant to certain interpretive skills of the researchers. In my view this means that to do good case and action evaluations requires more researcher skill and ability than some experimental designs, but also that bad case and action evaluations are both more common and less easy for researchers not familiar with these approaches to recognize.

More details

Appendix 3 gives more discussion about these approaches as they are often unfamiliar to medical and health services researchers, but increasingly used and useful for evaluations of CSIs. It includes a summary of how these approaches conceptualize both the CSI and the units responding to it as "moving targets" in a fast-changing environment, and gives notes on the resources needed and on criteria for assessing the quality of observational evaluations which use qualitative or mixed methods observational designs. The Volume 2 guidance also gives more details.

There are also other methods which can be grouped as "pre-implementation evaluation approaches" which include those described by There are also other methods which can be grouped as "pre-implementation evaluation approaches" which include those described by There are also other methods which can be grouped as "pre-implementation evaluation approaches" which include those described by There are also other methods which can be grouped as "pre-implementation evaluation approaches" which include those described by There are also other methods which can be grouped as "pre-implementation evaluation approaches" which include those described by Brown et al. 2008, as well as simulation modeling, or policy analysis which extrapolates from existing evidence to estimate costs and impact of spreading an intervention. Neither are multi-level modeling, Bayesian or adaptive trials described – references and resources giving guidance on these are given in volume 2.

How do I choose which design to use?

There is no one best design to evaluate a CSI. The RCT is the best of designs, but also the worst of designs, depending on the question. But the optimal design also depends on the constraints placed upon the evaluation (i.e., resources and time available, and the setting) and on the type of intervention-change being studied. The guidance in Volume 2 gives more details about the steps in choosing and carrying out an evaluation, but below is a summary of what to consider to get the best match between the design and the question, given the constraints.

4.4. Issues to consider in applying designs to evaluate CSIs

Section 3 of this volume already described some of the shortcomings of previous evaluations of CSIs for answering different questions (Figure 1). In recent years the field has developed a greater understanding of the range of questions which evaluations of CSI sometimes need to address, and that some designs can never satisfactorily answer certain questions. This summary notes some of the issues to consider raised by different commentaries and studies covered in the review of the field carried out for this overview and guidance.

Define the information needed by users of the evaluation

One issue to consider at the start of an evaluation of a CSI is how to identify the information needed and formulate a question in a way which can be answered adequately by any evaluation design, given the constraints.

If the evaluation is to focus on providing practical actors with information to make decisions about the CSI, then the evaluators and the users need jointly to list which information would be needed better to inform which decisions. The evaluator's role may include helping users to list the information needed and consider whether this information would make a difference for practical decisions and actions. The evaluator has in mind different designs and the costs and time-delivery parameters of these designs, and can show the users what is possible by when and for which costs. The process leads to agreeing on information to be collected by the evaluation, the questions it will answer and its purpose.

This implies that evaluations of CSIs always focus on practical users' needs for action-oriented information, which is not the case. Academically-based evaluators may more driven by previous research in formulating their question, as they have a view to publishing in peer-reviewed journals, which requires their research to fill a gap or show it adds to previous research rather than primarily answering a practical question. However, formulating the information needed to do this and the purpose and questions of the evaluation are still a necessary and under-emphasised part of an evaluation of CSIs.

The two aims of answering practical questions in a timely and actionable way and contributing to scientific knowledge can be achieved by one evaluation design. But it is challenging to do so, and it may be that the evaluator has to focus on achieving one of these aims.

Choose the design to answer the questions and provide the information

Whether the question is primarily practice-action driven or research-science driven, the choice of design should follow from the question. Evaluators tend to specialize in particular methods and designs and may be blind to the possibility that another design is more suited to the question. The much needed innovation in designs for evaluating CSIs to answer a range of questions requires all evaluators to consider modifications or hybrid designs which they may not have used before.

Understand and describe the limitations to the design for answering some questions

This is necessary to avoid others being misled by the evaluation findings in the final report. But at the start, defining the limitations can help innovation in design and method by sensitizing evaluators to possible data or strategies that could help the evaluation and provide answers to questions which it may not have otherwise been able to answer.

Recognise and address challenges which readers may have in using the evaluation

The point about researchers being unfamiliar with some designs and methods also applies to readers. Evaluations published in specialist journals can more often assume the average reader of such journals is able to understand the design and judge the significance and applicability of the results. Evaluation reports for a wider academic readership, or primarily for non-researchers, need to provide discussion of the design and methods which is not just more detailed but is also simplified and understandable for readers with knowledge of other designs, or of none at all.

In some cases the evaluator might not use a design if the user may not be able to understand and apply the findings generated by it, or because the evaluator honestly recognizes that they cannot present the design in a usable way. This applies equally to some quantitative analytic methods as to long narratives from qualitative or realist evaluations, which in theory are well suited to CSI evaluation but are difficult for some researchers to communicate or summarize.

Some qualitative observational or action evaluations cannot be summarized in tables or graphs, and only partially in models and diagrams. Long narratives are challenging for some practical decision makers to apply to their situation. Some quantitative techniques which report propensity scores and instrumental variables may be challenging for some researchers and non-researchers alike. Understanding is not widespread of multiple linear regression, or of analysis of covariance for continuous outcomes, logistic regression for binary variable outcomes, proportional hazards analysis or Cox regression, and Poisson regression when outcomes are measured as counts (Feinstein 1996, Concato 2012).

Pay attention to each of the "ADAGU" questions

Overall, it is best when designing the evaluation to note the strategies which can and will be used within the evaluation to address each of these questions below, which were presented in more detail earlier in this document. Stating these strategies will strengthen proposals for funding the evaluation, helps in developing innovations in design, and makes the final report more useful. Guidance on how to address them is given in Volume 2.

- Aims: which information is needed and what are the questions to be addressed?
- Description: what are the details of intervention, implementation and context?
- Attribution: how certain can we be that the intervention caused the outcomes reported?

- Generalization: can we copy it and get similar results?
- Usefulness: in which situations in which situations in which situations

4.5. The causality debate in CSI evaluation: do we need a theory of "mechanisms"?

"An evaluation was done of whether a fly painted on a urinal in Amsterdam airport would reduce spillage in men's toilets. It did, by 96%. We do not need a theory of the male psyche for this evaluation to be useful." (Comment given at a workshop UCLA in 2012)

A final issue to bear in mind when designing a CSI is the different assumptions underlying different designs about causality and about explanations for outcomes.

In the interpretive perspective to evaluating CSIs underlying some observational designs, the "causal mechanism" or explanatory principle is how the people "targeted" by the "intervention" understand it. The approach seeks to discover the meaning and value they give to the change which they see as being implied by the intervention for their behavior. This can help explain variations in success between different sites in one program being evaluated, and can give ideas about what to do to get success elsewhere not by copying the change exactly but by different actions to enable people to value and give meanings to the change which make it more likely they will take up the change in their behavior and organization.

For example, an intervention was made with a lecture about hand hygiene, reminder posters, and feedback about infections in nursing units. The assumption was that any behavior change depends on factors such as whether nurses believe infection rates are related to hand hygiene, whether they feel that they personally could be harming patients, whether they feel better about themselves as professionals and people if they follow hand hygiene and other factors. The intervention is mediated by nurses' interpretations of its implications for their daily work and of its significance for their identity as a person and as a professional.

These causal and explanatory assumptions are different to those which may be appropriate for a drug or surgical treatment. Some proponents of experimental evaluation propose that such designs make no assumptions about the causal processes working through the intervention and its targets: that the experimental design simply discovers whether or not outcomes are significantly associated with the intervention: do nurses change their behavior or not? Put another way, some would argue that in an experimental design, no theory of mechanism is needed to assess if the CSI has the intended effects or not.

To this "defense" of experimental evaluation, proponents of the interpretive perspective argue first, that if the intervention did have the intended effect in the study setting, then there is no certainty it will have similar effects if repeated elsewhere. The nurses elsewhere may respond to it differently.

Secondly, that to reproduce the effects elsewhere it is necessary to understand nurses' interpretations and design an intervention which provokes interpretations conducive to improved hand hygiene.

One counter-response by experimental proponents is that the intervention needs to be repeated in many settings, and if similar effects are observed, then there is then no need for theories about causal mechanisms and understanding interpretations. However, time and money do not usually allow repeated experiments in a range of settings possibly hostile to the intervention (which itself implies a theory to select "hostile settings").

It is likely that different assumptions are appropriate for different questions, and that "opening the black box" of an intervention and theorizing about "causal chains" and "mechanisms" is not necessary to answer some questions, such as "does it have an effect on this outcome measure in this setting?"

5. Summary and future developments

The purpose of these two volumes of guidance is to enable researchers to evaluate complex social interventions more effectively. There is guidance on how to use some experimental research methods for evaluating complex interventions which are more well known to medical and health services researchers, and this is drawn on in volume 2, which gives practical guidance and tools for all designs (Øvretveit 2013b). This volume 1 concentrated on defining and explaining what is special about complex social interventions, and why and when evaluators may need to use designs and methods with which they may be less familiar.

What is special about CSIs is that they are social, which makes them more complex in their workings and possible outcomes. More accurately, many interventions can be viewed as complex, or as complex and social, the latter term emphasising a range of considerations and methods which help to answer certain evaluation questions when evaluating changes to attitudes, behavior and organizations rather than changes to human physiology.

People, individually and in groups, respond to the intervention – such as a combination of training, performance feedback and reminders - in different ways, and this mediates the outcome of the intervention. The outcomes are less predictable than the effects of a pharmaceutical intervention on physiological outcome measures, such as heart rate. They are more dependent on how people interpret the personal meaning of the intervention. Yet people's responses are often surprisingly similar because they interpret interventions in a similar way because they are part of a group with similar norms and values. These are part of a wider "context" which affects their response in a greater way than a person's bodily-context affects their physiological response to a pharmaceutical.

The social nature of CSIs does not mean they cannot be evaluated as if they were a pharmaceutical intervention and using familiar experimental designs - if the aim is to answer "does it work?" effectiveness questions about whether the intervention is associated with a few measurable outcomes. But it does mean the answers may be less generalizable to other people in other settings, unless the experimental trial is repeated in a number of settings - especially in those where theory would suggest the outcomes would not be expected.

Some researchers propose that effectiveness questions about some CSIs can equally or better be answered using observational designs, where theory helps to decide which data is gathered at points on a causal chain of intermediate and later outcomes. For other questions, such as “exactly how should we implement the CSI?” or “what are providers’ or patients’ experiences of being exposed to the CSI?” then qualitative or mixed methods designs give useful information for action and for developing scientific knowledge.

This volume 1 showed the range of designs and how each was suited to different questions, illustrating both common and promising designs in example studies in the appendix. It noted considerations in choosing and planning design and in carrying out an evaluation of a CSI which are further elaborated in the Volume 2 guide.

5.1. Needed developments for evaluations of CSIs

Methods to help formulate answerable evaluation questions

The overview for this document identified an important and unrecognized development which is needed: methods for formulating the information wanted from the evaluation and answerable questions. Evaluators of CSIs tend to be methods- rather than questions-driven, in part because there is no guidance and methods for formulating the wanted information and answerable questions, which would direct attention to the need for certain data and hence methods best for the evaluation with the time and resources available. In part it is also because little attention is given to the importance of this stage of an evaluation in most texts and training, or by funders or others. Such guidance and methods are needed to define: 1) the information which would make a difference to the decisions and actions of users; 2) the questions answerable by evaluations; and 3) the constraints to the evaluation which affect what it can achieve.

Flexibility in matching design to question and constraints

With attention to and methods for this formulating phase, evaluators may then be better able to select a design suited to the information, questions and constraints. Which raises a second development required for better evaluations of CSIs: more flexibility in matching design to question, and to the needed information and constraints. This in turn requires more awareness amongst evaluators and funders of a range of designs, and ways for evaluators to acquire skills for designs with which they may be less familiar.

Innovations to experimental and quasi-experimental designs and improvements to reporting

As regards experimental designs, a number of commentaries and studies have suggested ways in which these could be modified and developed for evaluating CSIs. Some have developed guidance, and standards for reporting imply improvements in data collection, especially about the intervention and its context (Craig 2008, Schulz et al. 2010).

As regard RCTs, Wolff (2001) provides a useful discussion, suggesting that RCT designs, applied to some complex interventions, give findings which do not indicate whether the effectiveness of the intervention is due to the measured features of the intervention or other undocumented influences which are specific to site, staff, protocol or their interactions. Users are thus uncertain about whether the effects can be generalized to other populations or sites.

> *"However, this does not mean that we should disregard RCTs entirely, but rather that they should be modified. There are two possible ways: adding a comprehensive contextual evaluation based on mixed methods to the design, and using multiple sites…..*
> *What the future holds for the randomized trial of socially complex interventions is uncertain. At a minimum, the trials need to be more attentive to issues of selection bias, unmeasured context variables and 'uncontrolled' interaction effects."*
>
> Wolff (2001)

Improving the validity of observational evaluations

For observational evaluations using qualitative and mixed methods, the primary limitations are to internal and external validity, which is where developments in design, data collection and reporting are most needed. New strategies are called for to increase the certainty about outcomes being attributable to the intervention in these designs. This involves:

- choosing outcomes which can be "linked" in theory and by empirical observations to the intervention,
- assessing the influence of the intervention relative to other contextual influences, rather than assuming mono-causal determination.

For quantitative observational evaluations, different statistical techniques including multi-level modeling, as well as more theory-informed data gathering could be used to improve both internal and external validity. These are described in Volume 2.

Some observational evaluation approaches rely on establishing causal chains, by using patterns of evidence from different data sources that point on the direction that certain intermediate or final outcomes are influenced by the intervention and/or other influences. In practice, many studies draw heavily on interviewee perceptions, and a few on expert opinion processes. One view is that stronger evidence of influence can be created through formulating causal chain models, and by more data triangulation and by more rigorously applying these methods.

Using and developing theory about contextual influences

Another view is that strengthening evidence about the intervention effects on intermediate or final outcomes is not where the focus of development should be, but rather on developing ways of constructing better pre-study models to shape data collection about possible contextual influences on implementation, and using methods more rigorously to revise these models in the light of the data (Bickman 2000, Foy et al 2011).

Generic theories of conditions thought to be necessary for many types of change are a useful starting point for deciding which data about context to collect, to then be able to assess the relative influence of these factors on implementation (Helfrich et al. 2009 (ORCA) Damschroder et al. 2009 (CFIR), Stetler et al. 2011 (PARIHS), McCormack et al. 2008)

However, developments are needed so as to rely less on generic context theories or lists, and more on research which has found which contextual influences are most relevant to a defined category of CSI: the context factors which help and hinder implementing a falls prevention intervention in a nursing home are likely to be different to those important for implementing a computer decision support system in a teaching hospital, although both will probably be influenced by some similar factors if we view these at a high level of abstraction, such as leadership, financial incentives and regulations.

Related to this point, where studies do formulate initial models about context and/or causal chains, these are often not based on thorough reviews to search for previous research which has established empirically that certain context factors are important for a specific intervention. To help researchers to find this research and get on with their evaluation, a review is needed to show, by category of type of CSI, the context factors which previous research has found or suspects to influence implementation.

Longer follow-ups, and cost estimates

Another development needed is to cover longer time periods or have more follow up studies to document both how a CSI is or is not sustained, and longer term costs, as well as how changes in the context both help and hinder maintaining the CSI over these periods.

Finally, more priority needs to be given to including simple costing and financial estimations in CSI evaluations. In the current financial climate, an evaluation is largely irrelevant to practical users if does not give some indication of the costs of the CSI, and of the potential savings or ongoing costs of maintaining it, to different parties over different time periods (Øvretveit 2012). These indications can be, at the simplest, range-estimates, or more accurate budget impact assessments -- health economic evaluations are not likely to be feasible or needed for anything but high cost and high risk CSIs.

6. Conclusions

Many interventions to healthcare, to patients and to populations are not only complex but social interventions: the “social” emphasizes specific dynamics and an unpredictability which need to be considered by evaluators. This makes CSIs more challenging to evaluate than some interventions. Yet the cost and time increasingly invested in CSIs also calls for more resources to be allocated to their evaluation. It calls for improving evaluation methods and evaluator skills to answer different evaluation questions so as to help users to make more informed decisions about whether and how to implement an intervention.

Experimental designs are suited to assessing associations between the presence of the intervention and measured outcomes for providers and/or patients. Observational designs provide less certainty about these associations because there is less control for alternative explanations for outcomes. Some observational designs are, however, able to develop explanations for outcomes which can assist in implementing the CSI in settings other than the evaluation setting. In action evaluations, researchers help improve the CSI as it is implemented and gain insights which further contribute to explanations about outcomes. However, the evaluator’s participatory role, if significant, may be difficult for others to reproduce, which reduces the external validity.

Awareness of the range of designs, and of better ways to formulate answerable evaluation questions, makes it more likely evaluators will better match design to the purpose of and constraints on the evaluation. Other developments are needed such as more attention to formulating questions and goals for the evaluation, formulating program theory before data gathering, and attention to context, reporting standards and costing. The accompanying Volume 2 provides guidance, tools and resources for researchers to improve their evaluations of CSIs.

Appendices

Appendix 1: Glossary of terms

(Based on Mann, C 2003 Observational research methods. Research design II: cohort, cross sectional, and case-control studies, Emerg Med J 2003;20:54–60

Designs

Cohort studies: Look forwards in time by following up each subject

- Subjects are selected before the outcome of interest is observed
- They establish the sequence of events and can be used to discover causes of outcomes
- Numerous outcomes can be studied
- If prospective, they are expensive and often take a long time for sufficient outcome events to occur to produce meaningful results

Cross sectional studies: Look at each subject at one point in time only,

- Subjects are selected without regard to the outcome of interest
- Less expensive & quick
- Weaker evidence of causality than cohort studies

Case-control studies: Looks back at what has happened to each subject,

- Subjects are selected specifically on the basis of the outcome of interest
- Only one outcome is studied
- Lower cost
- Efficient (small sample sizes)
- Quantitative studies produce odds ratios that approximate relative risks for each variable studied as long as the outcome of interest is rare.
- Prone to sampling bias and retrospective analysis bias

Other terms

- **Bias**: The inclusion of subjects or methods such that the results obtained are not truly representative of the population from which it is drawn
- **Blinding**: The process by which the researcher and or the subject is ignorant of which intervention or exposure has occurred.
- **Cochrane database**: An international collaborative project collating peer reviewed prospective randomized clinical trials.
- **Cohort**: a group identified so that one or more characteristic can be studied through time.
- **Confounding variable**: A variable that is associated with both the exposure and outcome of interest that is not the variable being studied.
- **Control group**: A group of people without the condition of interest, or unexposed to or not treated with the agent of interest.

- **Incidence**: Is a rate and therefore is always related either explicitly or by implication to a time period. With regard to disease it can be defined as the number of *new* cases that develop during a specified time interval.
- **Latency**: A period of time between exposure and the development of evidence of changes associated with that exposure.
- **Matching**: choosing control subjects, to be similar to the test subjects (e.g. similar age, or size of unit) so that these features can be excluded from the assessment of what caused the difference between controls and intervention subjects in respect of the outcomes measured,
- **Observational study**: A study in which no intervention is made (in contrast with an experimental study). Such studies provide estimates and examine associations of events in their natural settings without recourse to experimental intervention.
- **Prevalence**: Is not defined by a time interval and is therefore not a rate. It may be defined as the number of cases of a disease that exist in a defined population at a specified point in time.
- **Relative risk**: This is the ratio of the probability of developing the condition if exposed to a certain variable compared with the probability if not exposed.
- **Response rate**: The proportion of subjects who respond to either a treatment or a questionnaire.
- **Risk factor**: A variable associated with a specific disease or outcome.
- **Validity—internal**: The rigor with which a study has been designed and executed—that is, can the conclusion be relied upon?
- **Validity—external**: The usefulness of the findings of a study with respect to other populations.
- **Variable**: A value or quality that can vary between subjects and/or over time

Appendix 2: Example studies showing the application of designs to different CSIs

List of the eight example studies

The following evaluation studies of complex social interventions were selected to illustrate the application of each design and to illustrate the strengths and weaknesses of the different designs. The summaries show the question addressed, the intervention evaluated, the design, the outcomes measured and discovered, the certainty about and generalisablity of the outcomes and the strengths and weaknesses the design.

1. RCT of a care transitions intervention
2. Pragmatic Cluster RCT of a multifaceted intervention to ICUs
3. Interrupted time series evaluation of a hand hygiene intervention
4. Prospective observational evaluation of an intervention to change primary care physicians behaviour (mixed methods)
5. Prospective observational evaluation of large scale safety programme (mixed methods)
6. Process evaluation of an intervention to promote smoking cessation
7. Case evaluation of a programme for electronic summaries of patients' medical records (mixed-methods)
8. Realist evaluation of large scale "transformation" of health services in London

Example 1. Randomized control trial of the "Care Transitions Intervention"

Coleman, E Parry, C Chalmers, S Min, S Chalmers 2006 The Care Transitions Intervention Results of a Randomized Controlled Trial Arch Intern Med. 2006;166:1822-1828

Question addressed:

What is the effect on re-hospitalization and hospital costs of a CSI which coaches chronically ill older patients and their caregivers and includes a "transitions coach"?

The intervention evaluated:

1) Tools to promote cross-site communication including a personal health record,
2) encouragement to take a more active role in their care and to assert their preferences, and
3) continuity across settings and guidance from a "transition coach."

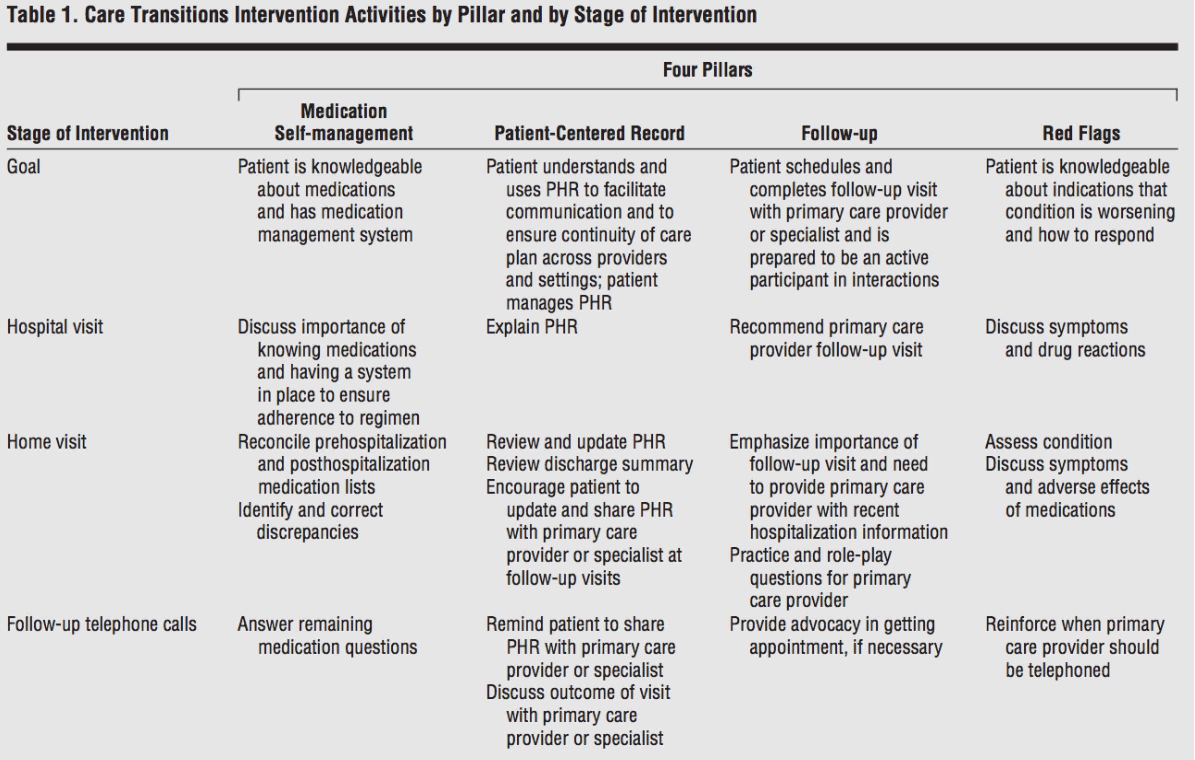
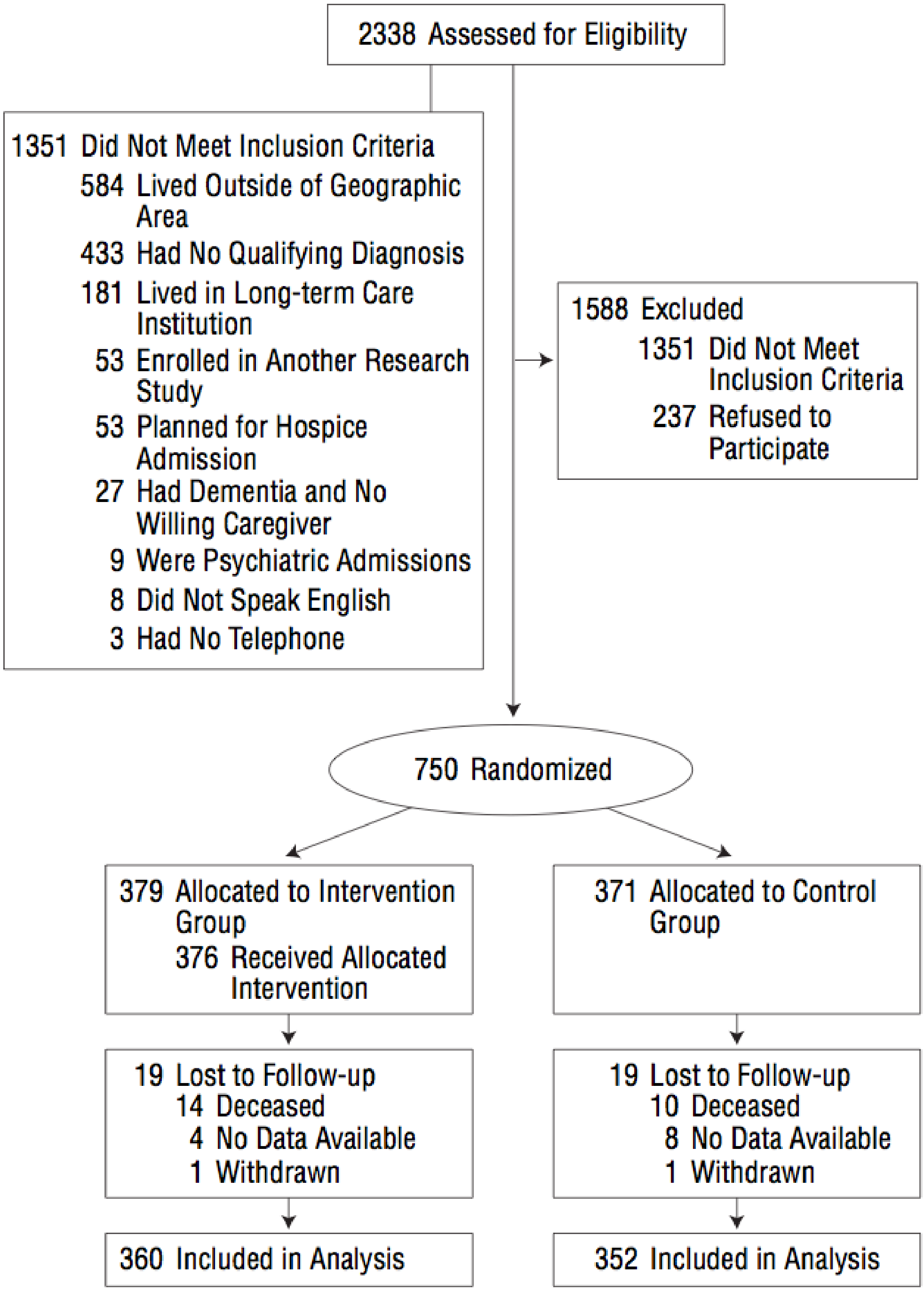


Table 1. Care Transitions Intervention Activities by Pillar and by Stage of Intervention

	Four Pillars			
Stage of Intervention	**Medication Self-management**	**Patient-Centered Record**	**Follow-up**	**Red Flags**
Goal	Patient is knowledgeable about medications and has medication management system	Patient understands and uses PHR to facilitate communication and to ensure continuity of care plan across providers and settings; patient manages PHR	Patient schedules and completes follow-up visit with primary care provider or specialist and is prepared to be an active participant in interactions	Patient is knowledgeable about indications that condition is worsening and how to respond
Hospital visit	Discuss importance of knowing medications and having a system in place to ensure adherence to regimen	Explain PHR	Recommend primary care provider follow-up visit	Discuss symptoms and drug reactions
Home visit	Reconcile prehospitalization and posthospitalization medication lists Identify and correct discrepancies	Review and update PHR Review discharge summary Encourage patient to update and share PHR with primary care provider or specialist at follow-up visits	Emphasize importance of follow-up visit and need to provide primary care provider with recent hospitalization information Practice and role-play questions for primary care provider	Assess condition Discuss symptoms and adverse effects of medications
Follow-up telephone calls	Answer remaining medication questions	Remind patient to share PHR with primary care provider or specialist Discuss outcome of visit with primary care provider or specialist	Provide advocacy in getting appointment, if necessary	Reinforce when primary care provider should be telephoned

The design:

750 older patients were admitted to the study hospital and randomly allocated to the intervention and a control group:



2338 Assessed for Eligibility
1351 Did Not Meet Inclusion Criteria
584 Lived Outside of Geographic Area
433 Had No Qualifying Diagnosis
181 Lived in Long-term Care Institution
53 Enrolled in Another Research Study
53 Planned for Hospice Admission
27 Had Dementia and No Willing Caregiver
9 Were Psychiatric Admissions
8 Did Not Speak English
3 Had No Telephone
1588 Excluded
1351 Did Not Meet Inclusion Criteria
237 Refused to Participate
750 Randomized
379 Allocated to Intervention Group
376 Received Allocated Intervention
371 Allocated to Control Group
19 Lost to Follow-up
14 Deceased
4 No Data Available
1 Withdrawn
19 Lost to Follow-up
10 Deceased
8 No Data Available
1 Withdrawn
360 Included in Analysis
352 Included in Analysis

Outcomes measured and discovered:

- Re-hospitalization at 30, 90, 180 days (data abstracted from the study delivery system's administrative records).
- Non-elective hospital cost outcomes 30, 90, 180 days.
- Intervention group showed lower re-hospitalization rates and lower mean hospital costs ($2058) vs. controls ($2546) at 180 days.

Certainty about and generalisablity of the outcomes:
An analysis showed no significant differences between key characteristics of the intervention group (n =379) and control group (n=371), so it is unlikely that the better results for the intervention group were due to any characteristics of the patients rather than their response to the intervention, all other things being equal. Other explanations could have been a systematic chance difference between the two groups in the primary care and after care services they received: to some extent this was minimized as all were covered by the same delivery system

As regards generalization (reproducibility of the intervention and then the outcomes), the intervention could in principle be reproduced with more details about the different components. The results, though, might be different as in other places the primary care and after care may be different in a way which affects the impact on the intervention.

Strengths and weaknesses of the design in this example:
The strengths of the RCT for some CSIs are illustrated in this design in being able to give a high certainty about the outcome findings as patient-characteristics and some other influences on outcome were controlled for. The design forces a focus on a few measures, and in this case it chooses measures and data gathering methods for these measures which could answer the research question: the data were reasonably accessible at low cost and were valid measures for the question.

The weaknesses are the resources and effort to enforce the trial protocol of carefully selected patients, fidelity of implementation to a prescribed intervention, and measurement are costly and difficult to reproduce in other settings. These implementation steps and the details of the intervention can be described in the trail report to give others guidance to reproduce the intervention, but reproducing it is still not always easy. These are the reasons why, although the RCT gives high "internal validity" (the certainty referred to above), the "external validity" (generalisablity) is low, unless trials are done in many settings, which is usually too expensive and time consuming, unless cluster RCT designs are used – the next example.

Example 2. Pragmatic Cluster RCT of a multifaceted intervention to 15 intensive care units

Scales DC, Dainty K, Hales B, et al. A multifaceted intervention for quality improvement in a network of intensive care units: a cluster randomized trial JAMA. doi:10.1001/jama.2010 .2000 [published online January 19, 2011]

Study funded by the Ontario health care delivery system, rather than a research funding agency.

Question addressed

Does a complex intervention to community based critical care units increase adherence to 6 quality activities that have been documented to improve patient outcomes?

The intervention evaluated

A multifaceted "knowledge transfer" intervention sequentially to improve delivery of 6 practices to personnel in 15 ICUs, covering education, dissemination of algorithms, and audit and feedback, transferred through an interactive telecommunication strategy (video conference based forum).

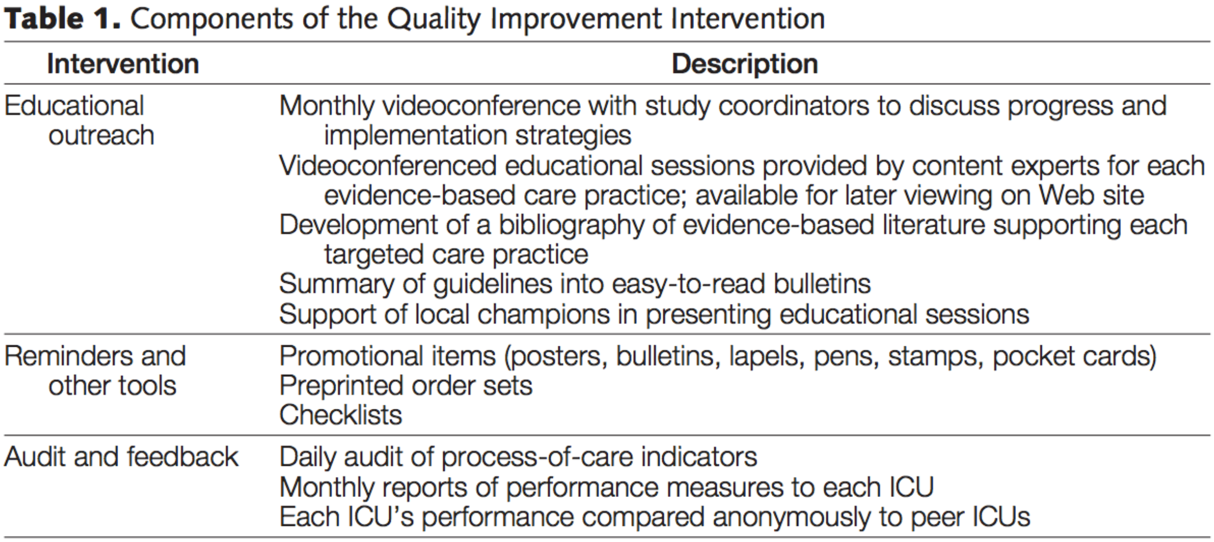
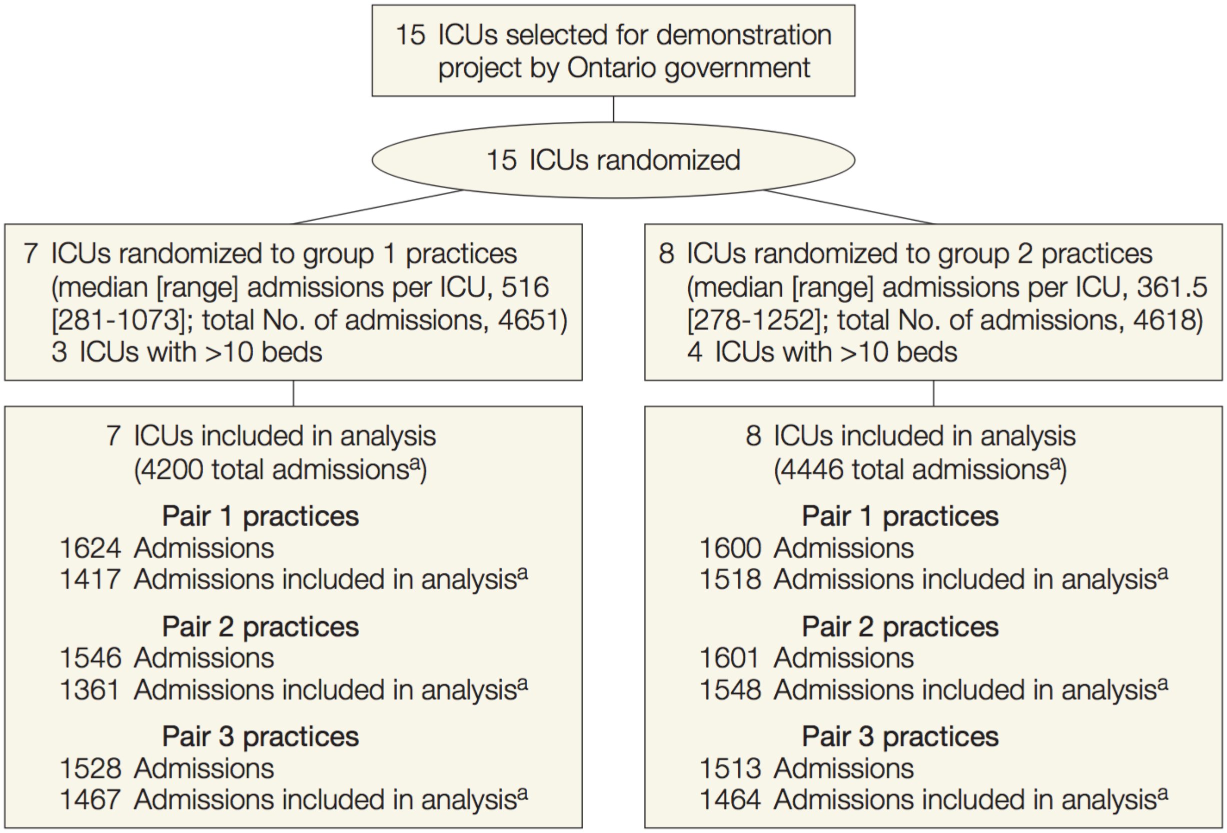


Table 1. Components of the Quality Improvement Intervention

Intervention	Description
Educational outreach	Monthly videoconference with study coordinators to discuss progress and implementation strategies Videoconferenced educational sessions provided by content experts for each evidence-based care practice; available for later viewing on Web site Development of a bibliography of evidence-based literature supporting each targeted care practice Summary of guidelines into easy-to-read bulletins Support of local champions in presenting educational sessions
Reminders and other tools	Promotional items (posters, bulletins, lapels, pens, stamps, pocket cards) Preprinted order sets Checklists
Audit and feedback	Daily audit of process-of-care indicators Monthly reports of performance measures to each ICU Each ICU's performance compared anonymously to peer ICUs

The design

Pragmatic Cluster RCT. In this design both "experimental" group (N=7) and the "control: group (n=6) received the intervention, but sequencing was used so that, over 4 months, a different care practice were targeted in the experimental group to that targeted in the control group. Put another way, this means during each 4-month phase of the trial, each group of ICUs received the active behavior change intervention targeting one care practice and simultaneously acted as a control group for the other group of ICUs that received the active behavior change intervention targeting a different care practice. This avoided randomizing a group of ICUs to no intervention, which, as the researchers say "could have been demoralizing".


15 ICUs selected for demonstration project by Ontario government
15 ICUs randomized
7 ICUs randomized to group 1 practices (median [range] admissions per ICU, 516 [281-1073]; total No. of admissions, 4651) 3 ICUs with >10 beds
8 ICUs randomized to group 2 practices (median [range] admissions per ICU, 361.5 [278-1252]; total No. of admissions, 4618) 4 ICUs with >10 beds
7 ICUs included in analysis (4200 total admissions[a])
Pair 1 practices
1624 Admissions
1417 Admissions included in analysis[a]
Pair 2 practices
1546 Admissions
1361 Admissions included in analysis[a]
Pair 3 practices
1528 Admissions
1467 Admissions included in analysis[a]
8 ICUs included in analysis (4446 total admissions[a])
Pair 1 practices
1600 Admissions
1518 Admissions included in analysis[a]
Pair 2 practices
1601 Admissions
1548 Admissions included in analysis[a]
Pair 3 practices
1513 Admissions
1464 Admissions included in analysis[a]

Outcomes measured and discovered

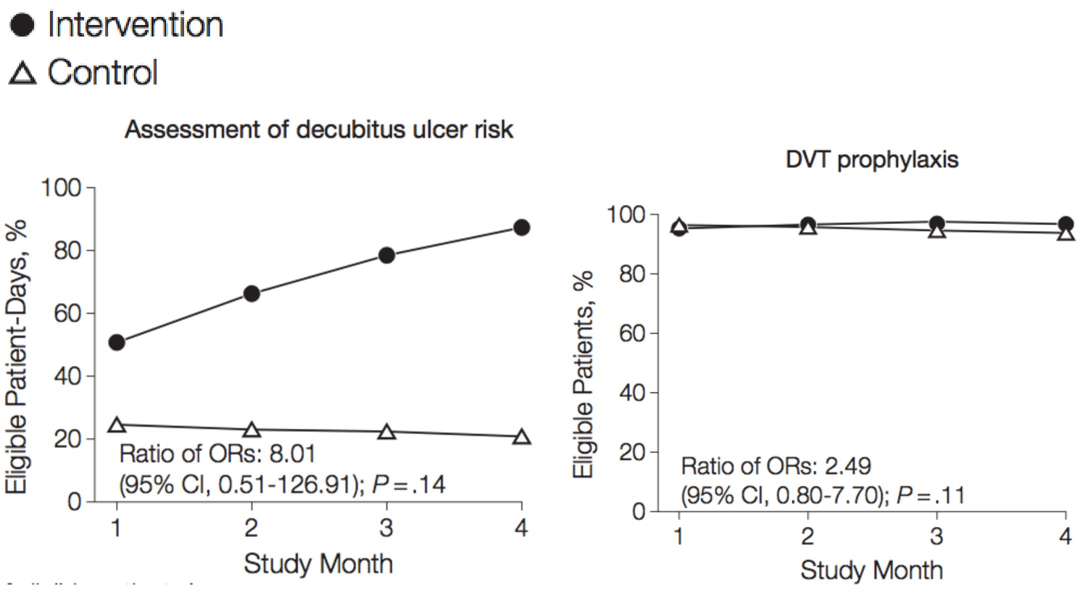
The measures were selected to assess whether activities thought to improve patient outcomes changed as a result of the intervention and by how much. The activities aimed to prevent of ventilator-associated pneumonia (VAP), increase prophylaxis for deep venous thrombosis (DVT), increase daily spontaneous breathing trials, prevent of catheter-related bloodstream infections, increase early enteral feeding, and prevent of pressure ulcers. The "Table 2" from the report below shows a number of process of care indicators and measures which were collected from different data sources using different methods:

Table 2. Process-of-Care Indicators for Each Targeted Care Practice

Care Practice	Process-of-Care Indicators	Main Measurement	Other Measurements
Prevention of ventilator-associated pneumonia	Semirecumbent positioning Orotracheal intubation	No. of eligible patient-days with head of bed ≥30°	No. of eligible patient-days associated with orotracheal (vs nasotracheal) intubation
Prophylaxis against deep vein thrombosis	Administration of anticoagulant prophylaxis Use of antiembolic stockings if anticoagulant prophylaxis contraindicated	No. of eligible patients receiving appropriate anticoagulant prophylaxis within 48 h	No. of eligible patient-days associated with receipt of anticoagulant prophylaxis Ineligible days associated with use of antiembolic stockings
Daily spontaneous breathing trials	Spontaneous breathing trial or extubation within previous 24 h	No. of eligible patient-days on which spontaneous breathing trial (or extubation) was performed	
Prevention of catheter-related bloodstream infections	7-Point checklist for sterile insertion completed Fulfillment of all 7 criteria listed on checklist Anatomical site of catheter insertion	No. of central venous catheters inserted using all 7 criteria on checklist	No. of central venous catheters inserted at the subclavian site (vs jugular or femoral sites)
Early enteral feeding	Initiation of enteral feeds within 48 h of ICU admission	No. of eligible patients receiving early enteral feeding within 48 h of ICU admission	No. of eligible patients achieving 50% of their target caloric goal via the enteral route by 72 h
Decubitus ulcer prevention	Completion of the Braden index[27] at least daily	No. of patient-days with Braden index completed	

Trained data collectors used handheld wireless electronic devices that connected to a central database via a local server. Each participating ICU selected a data collector (a nurse or a ward clerk not providing patient care), who received data collection training from the central trial coordinating office. The measure definition was the presence of one process-of-care indicator and no contraindications to receiving the practice. Data were collected once daily.

The findings were that, overall, patients in ICUs receiving active intervention were more likely to receive the targeted care practice than those in contemporaneous control ICUs receiving an active intervention for a different practice. There were however significant differences between care practices in their increase compared to the "control", suggesting some care practices are more amenable to improvement through this intervention than others.



Certainty about and generalisablity of the outcomes

The randomization and parallel switching of the interventions in the two groups does reduce the number of other explanations for the outcomes, but requires complex data collection and analysis. An observational study would not have been able to exclude temporal change as an explanation for the findings in the same way. The study also used a "decay-monitoring period" to assess if the improvements persisted and interviews with participants to possible ways in which the intervention had its effect and what helped and hindered.

As regards repeating the intervention, the time resources and effort to implement the trial were largely provided by the research and these resources may not be available elsewhere.

Strengths and weaknesses of the design in this example

As regards the outcome measures, these are indicators which are thought to correlate with patient outcomes, but the evidence of this correlation is contested. It is a long and fragile causal chain from carrying out a pressure ulcer risk assessment to reduce pressure ulcer incidence. Following the care practice cannot be said with certainty to then result in better patient outcomes.

Even though there were over 9000 ICU admissions, this was not a large enough sample size to find differences in patient outcomes. A trial to do this would need to be too large and expensive to be feasible and requiring patient-level outcomes would have slowed down the study.

Another strength of the design is the useful evidence it provided about how the intervention appears to have significantly improved three care practices (for VAP positioning, CRBS and pressure ulcer risk assessment) but had little impact on three others (DVT prophylaxis, daily SBT and early enteral nutrition).

This study deployed its data collection resources to collect a range of measures from each ICU. The disadvantage of this is there is not much research time and resources available to assess the validity and comparability of the data, so there may be questions about data accuracy – the report describes a site inspection and audit of data collection at each ICU during the trial but no further details are given.

Cluster randomized trials are complex to carry out successfully. The primary outcome (summary ratio of odds ratios) is complicated to compute and understand but is appropriate for the research question.

Example 3. Interrupted time series evaluation of a hand hygiene intervention

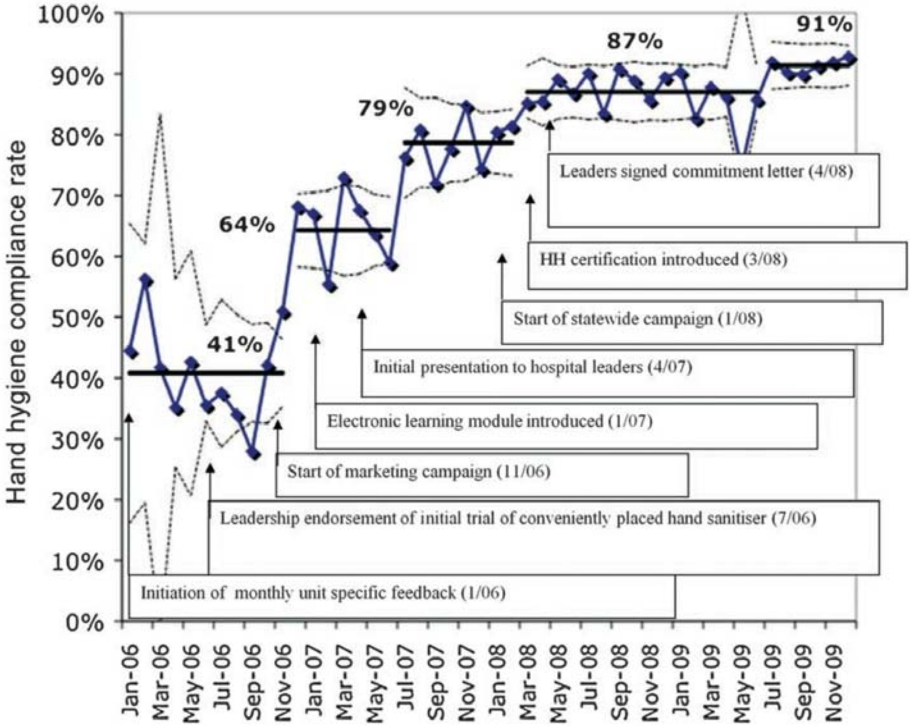
Kirkland KB, Homa KA, Lasky RA, et al. Impact of a hospital-wide hand hygiene initiative on healthcare-associated infections: results of an interrupted time series. BMJ Qual Saf 2012. Published Online First: 24 July 2012. doi: 10.1136/ bmjqs-2012-000800

Question addressed

Does a multifaceted hospital-wide intervention improve hand hygiene (HH) and reduce hospital associated infection in patients?

The intervention evaluated

public statements by leadership 2) measurement/feedback (monthly audits on all units published on an intranet site available to all staff, and reported to executive leadership, clinical leaders and board members, 3) increased hand sanitizer availability; (4) education/training using internet and a certification programme for competency; and (5) marketing/communication with posters and screen savers, stories in medical centre publications and local news outlets, and direct communications with staff about expectations and progress towards goals.



The design

Three-year interrupted time series, with multiple sequential interventions, and 1-year post-intervention follow-up.

Outcomes measured and discovered

Implementation fidelity: measured the number of HH audits, the inventory of hand sanitizers consumed, and the number of HH-related posters, screensavers and articles in internal publications. To assess staff exposure to the interventions, measurement of the number of monthly visits to the report card website, the completion rate for the electronic learning module and the number of staff who were certified 'competent' in HH.

Observed HH compliance (%): direct covert observation by infection prevention staff (not blinded to the interventions), with training to ensure 90% agreement in measurement during simultaneous observation periods. Monthly observations in all units counted HH 'opportunities' (before and after contact with patients or their immediate environments) and documented whether HH was performed (how was this done without staff being aware is not described in the report). Compliance% = number of times HH performed by the total number of opportunities.

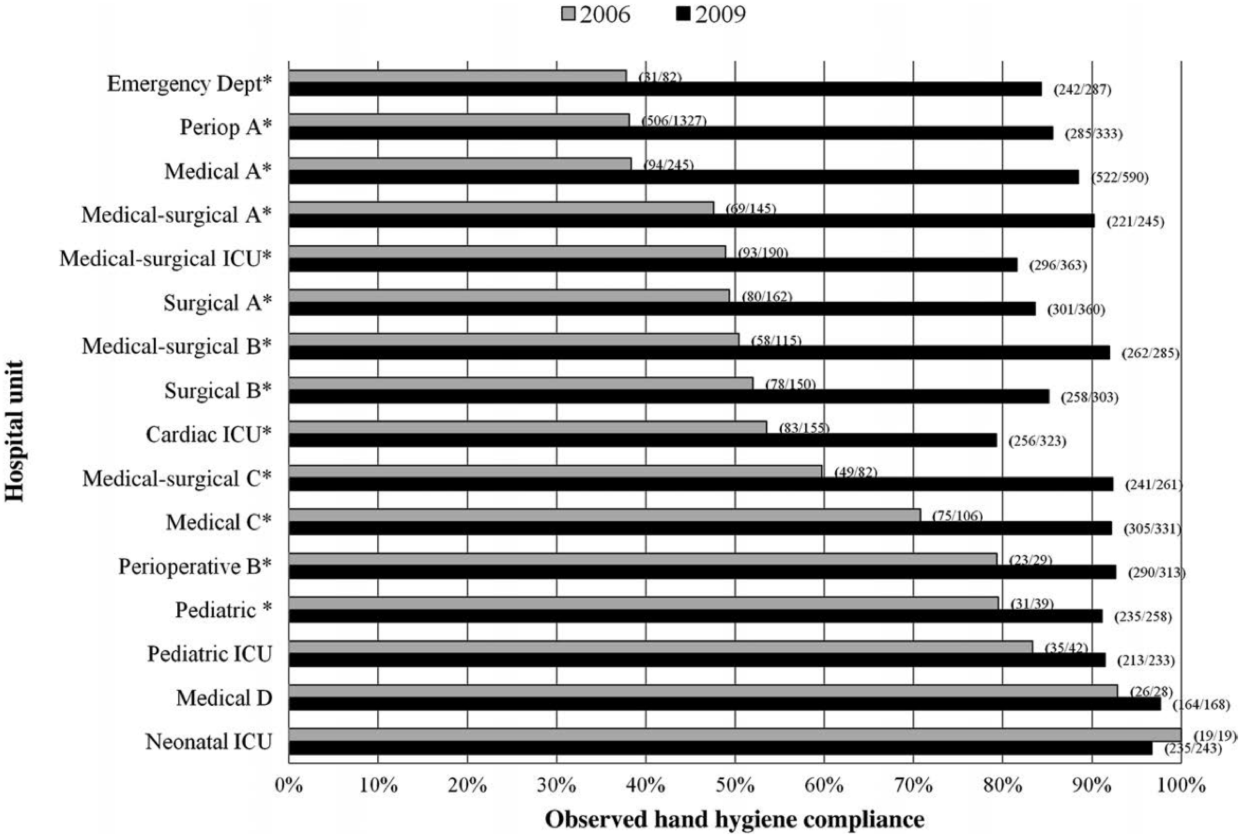
Rates of healthcare-associated infection per 1000 inpatient days: measured by daily review of microbiology data, with medical record review, and infection prevention staff applied standard definitions to identify all cases of bloodstream infection due to any organism, clinical infection at any site due to Staphylococcus aureus and Clostridium difficile infection.

Results

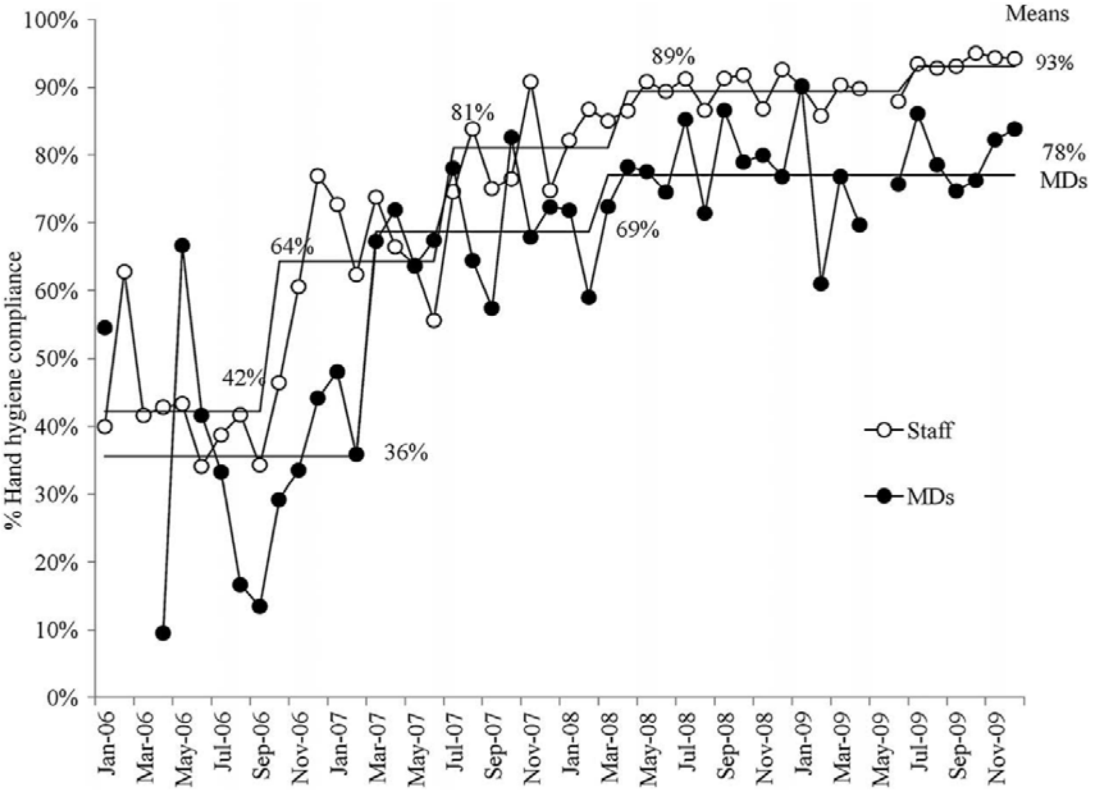
1) Quite high intervention fidelity (details p 1021)

2) Improved HH compliance:

Compliance rates per units:

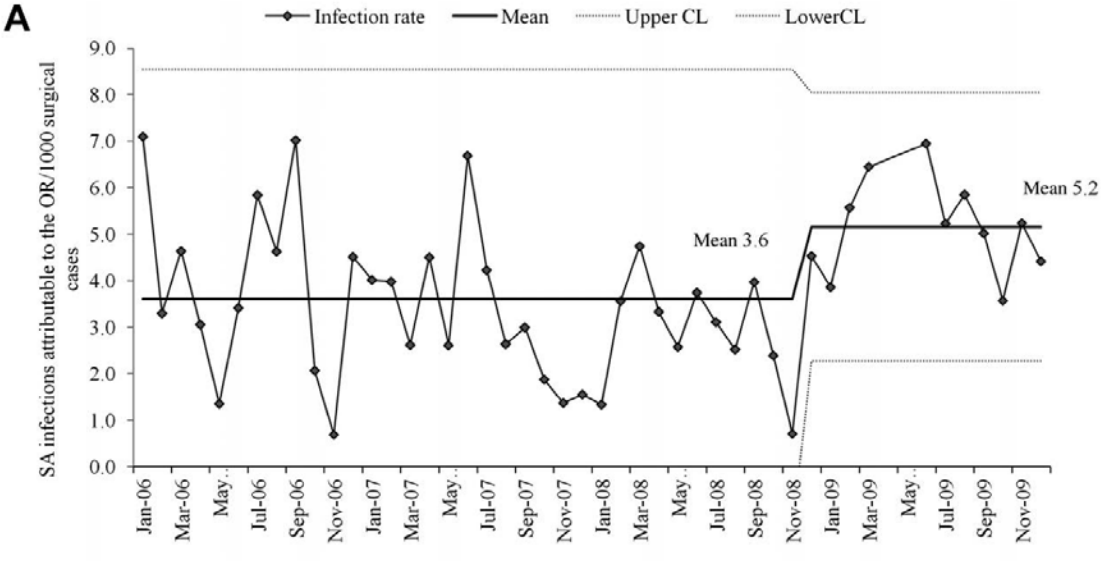


Physician HH rates using statistical process control analysis:

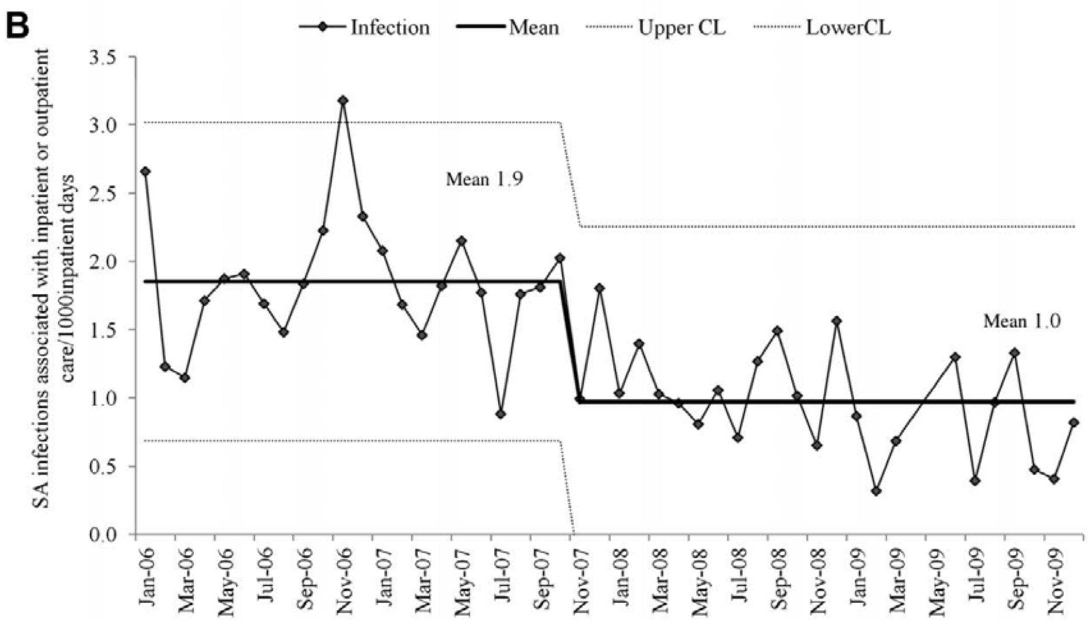


3) Infection rates

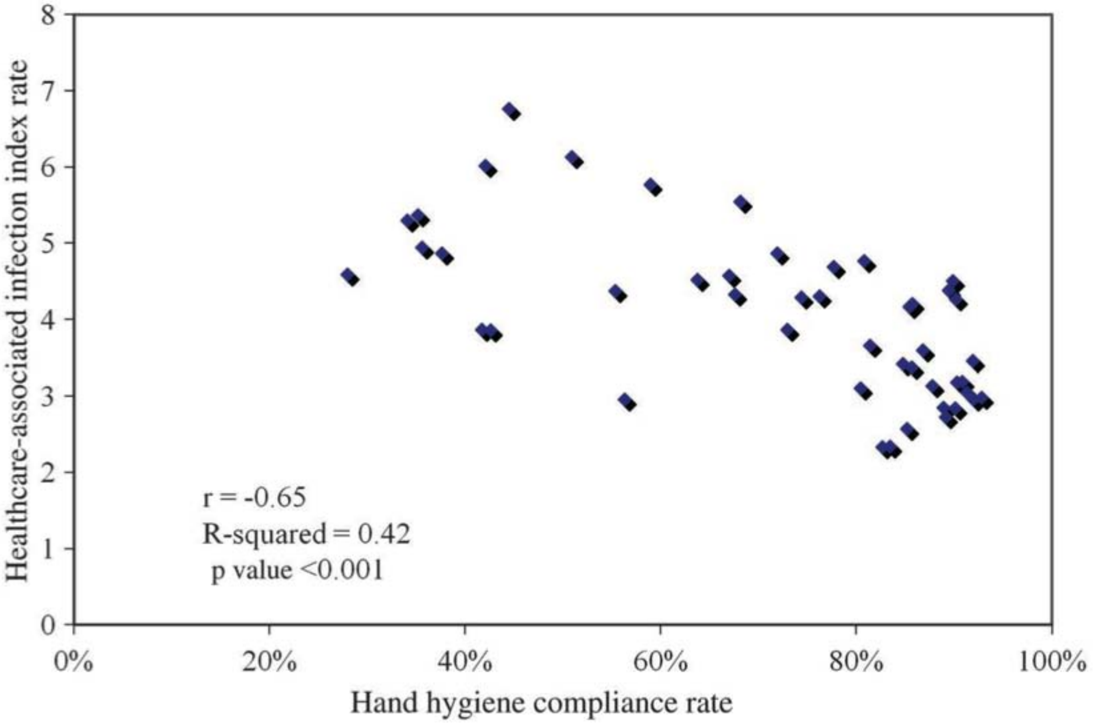
A below: Infections attributable to OR/1000 surgical cases



B below: Infections associated with inpatient and outpatient care /1000 inpatient days



C below: Scatter plot of monthly healthcare-associated infection index rates and hand hygiene compliance rates.



Certainty about and generalisablity of the outcomes

One strength was the assessment of stages in the causal chain which increased certainty that the outcomes were due to the intervention: 1) extent of implementation, 2) HH compliance, 3) infection rates. Another was the actions taken to select and collect a few key measures, the validity of these measures for the question, and the accuracy of the data on 1), 2) and 3). Both these are not necessarily intrinsic to the type of design (other observational designs include these elements), but show best practice in using this type of design.

Certainty of attribution may be reduced by the observation method which may be an intervention in its own right: staff must have known about monthly "direct covert observation" so the intervention is not just feedback of data but also either awareness of being observed or knowing an observer is present. Also a simultaneous state- wide campaign - the 'High Five for a Healthy NH' may have contributed to the results

Reproducibility of the intervention may be questionable because of the high cost. Timely feedback as an essential part of many CSIs and both process (compliance) and outcome data (infections) are collected for credibility. However the cost of this, especially monthly observations of all units, may be too high for many hospitals, so costs may make generalisablity difficult without savings data and a return on investment assessment.

Strengths and weaknesses of the design in this example

Sustained hand hygiene compliance is one of the most difficult changes to achieve and requires a CSI which includes an infrastructure and systems to sustain it (the latter was not described in the study). One strength was the assessment of stages in the causal chain noted above to increase the certainty of attribution. Also, the study used a 'tracer condition' less sensitive to hand hygiene - infections attributed to the operating room - as a comparison: that this rose at the same time as the other data feel increases certainty that these falls were due to the intervention.

The design was also about to accommodate the evolution of the programme, but it should be noted it did not give exact details the interventions introduced or intensified over the three years. This does not allow us to assess exactly which interventions were more impactful than others, or how much the results depended on a whole programme and a synergy between the components.

Another strength of the study, but not intrinsic to this design, was the use of process control charts which helped to look ahead and decide if the change was random or due to a special cause, and also to monitor the impact of the phased introduction of the interventions.

As with other studies of this type, it is the differences between units which are of interest: although the design was able to show this, it could not explain it. An additional retrospective study could have been made using another design and could have used compliance-observers insights to build a theory.

Example 4. Mixed methods prospective observational evaluation of an intervention to change primary care physicians behavior

Nazareth, I Freemantle N Duggan C Mason J Haines A 2002 Evaluation of a complex intervention for changing professional behavior: the Evidence Based Out Reach (EBOR) Trial, Journal of Health Services Research & Policy Vol 7 No 4, 2002: 230–238

(Related RCT: Freemantle N, Nazareth I, Eccles M, Wood J, Haines A and the EBOR Trialists. A randomized trial of the effect of educational outreach by community pharmacists on prescribing in primary care. British Journal of General Practice 2002; 52: 290–295)

<u>Question addressed</u>

Do educational outreach visits by pharmacists change primary care practitioners (GPs) prescribing and what are the barriers and enablers of change?

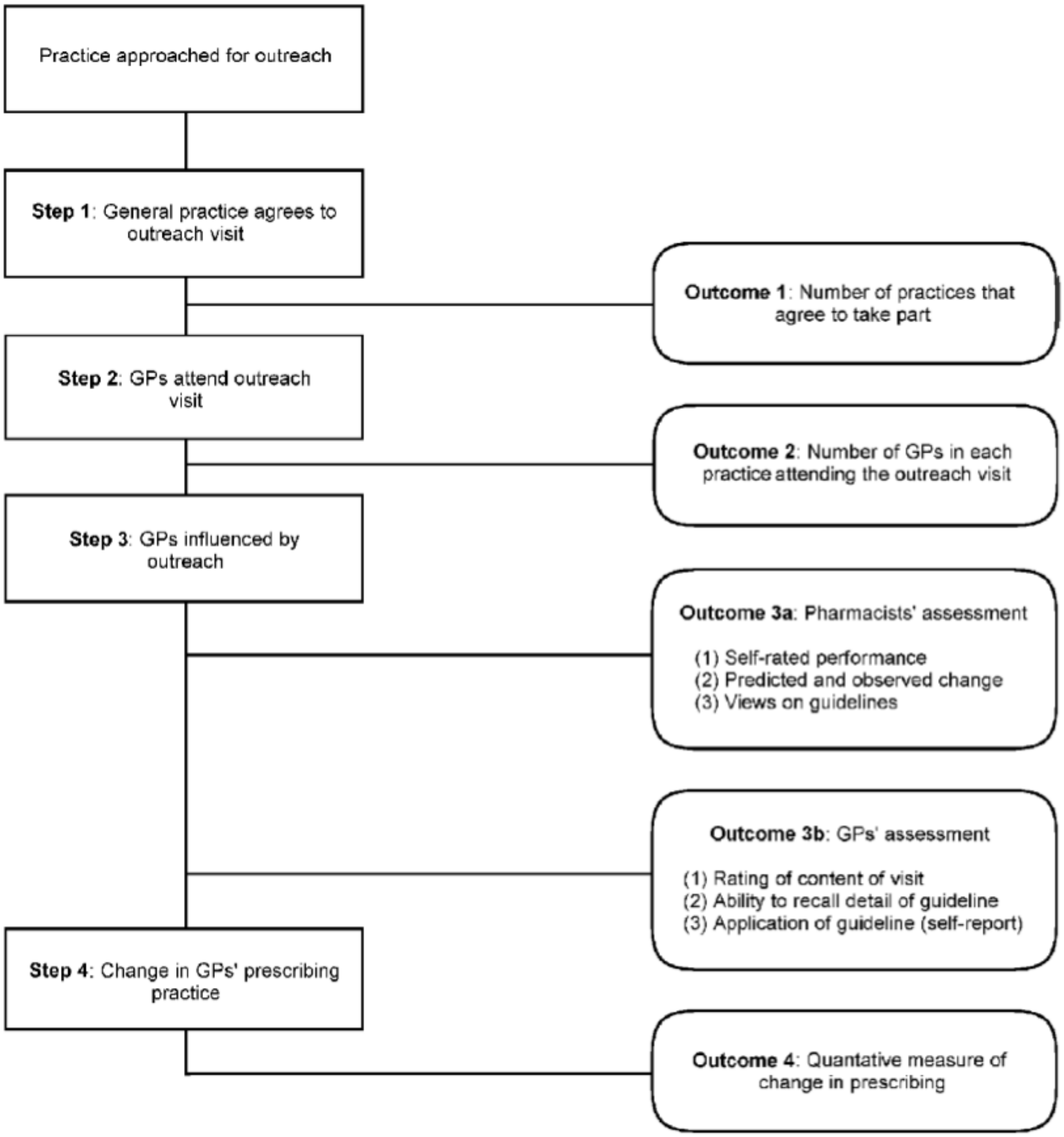
(a separate RCT assessed the effect of the intervention, reporting on changes in prescribing)

<u>The intervention evaluated</u>

Community pharmacists visited PCPs to give education on treatment recommendations for: angiotensin-converting enzyme inhibitors for the management of heart failure; aspirin for the secondary prevention of cardiovascular diseases; antidepressant therapy for treating depression; and non-steroidal anti-inflammatory drugs for the management of osteoarthritis.

<u>The design</u>

Mixed methods prospective observational evaluation. Each practice received an outreach visit for two topics and served as a control for the other two topics.


Practice approached for outreach
Step 1: General practice agrees to outreach visit
Outcome 1: Number of practices that agree to take part
Step 2: GPs attend outreach visit
Outcome 2: Number of GPs in each practice attending the outreach visit
Step 3: GPs influenced by outreach
Outcome 3a: Pharmacists' assessment
(1) Self-rated performance
(2) Predicted and observed change
(3) Views on guidelines
Outcome 3b: GPs' assessment
(1) Rating of content of visit
(2) Ability to recall detail of guideline
(3) Application of guideline (self-report)
Step 4: Change in GPs' prescribing practice
Outcome 4: Quantative measure of change in prescribing

Pathway of change in general practitioners' (GPs') prescribing practice.

Outcomes measured and discovered

Staged outcomes Participation, attendance at educational meeting, pharmacists assessment of the visits; GPs assessments of the visit (questionnaire), change in prescribing of medications compared to guideline (reported in the RCT study (Freemantle et al 2002)).

Data on pharmacists assessment of the visits was by their assessing acceptability of the message to the GPs, their own overall performance, the rapport established with the doctors & prediction of whether the doctors' prescribing practices would change. Nominal group interviews were used to assess the pharmacists' views of the interventions. Two group interviews were held the 1st, soon after the pharmacists had completed their initial visits and the second just before they had completed their last visit.

The details of the findings for each of the staged outcomes are given in the paper as well as the barriers discovered to changes in prescribing. As regards these changes, a 4% change in prescribing for the antidepressant guideline was observed and 2% change for the angiotensin-converting enzyme inhibitors guideline and 7% for the aspirin guideline.

Certainty about and generalisablity of the outcomes

The RCT study was to answer the, "does it work?" question. In contrast this observational study was able to being to answer the, "why does it work?" and the "what helps and hinders?" questions as well as the reasons for different impact on different types of prescribing.

As regards reproducing the intervention, the details provided would allow generalization, but whether the outcomes would be the same may depend more on the context in another setting, such as financing for drugs budget and incentives to reduce prescribing, which were not investigated in the study.

Strengths and weaknesses of the design in this example

Interestingly this study alone could have given a less certain but useful and answer at a lower cost than an RCT to the "does it work?" question, because the causal chain was mapped and the links between the staged outcomes could be established. It provided data which made it possible to discover the barriers to changes in prescribing and explain the different changes which were achieved. One weakness was that the sample may have been biased towards change-receptive GP practices - a number declined to participate.

Example 5. Mixed methods prospective observational evaluation of large scale safety programme

Benning A, Ghaleb M, Suokas A, Dixon-Woods M, Dawson J, Barber N, et al. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. BMJ 2011;doi:10.1136/bmj.d195.

(see also: a) Benning A, Nwulu U, Ghaleb M, Dixon-Woods M, Dawson J, Barber N, et al. A controlled evaluation of the second phase of a complex patient safety intervention implemented in English hospitals. 2010. www.haps.bham.ac.uk/publichealth/psrp/EvalSPI.shtml.

b) Benn J Burnett, S Parand, A Pinto, A Iskander, S Vincent, C Studying large-scale programmes to improve patient safety in whole care systems: Challenges for research Social Science & Medicine 69 (2009) 1767–1776).

Question addressed

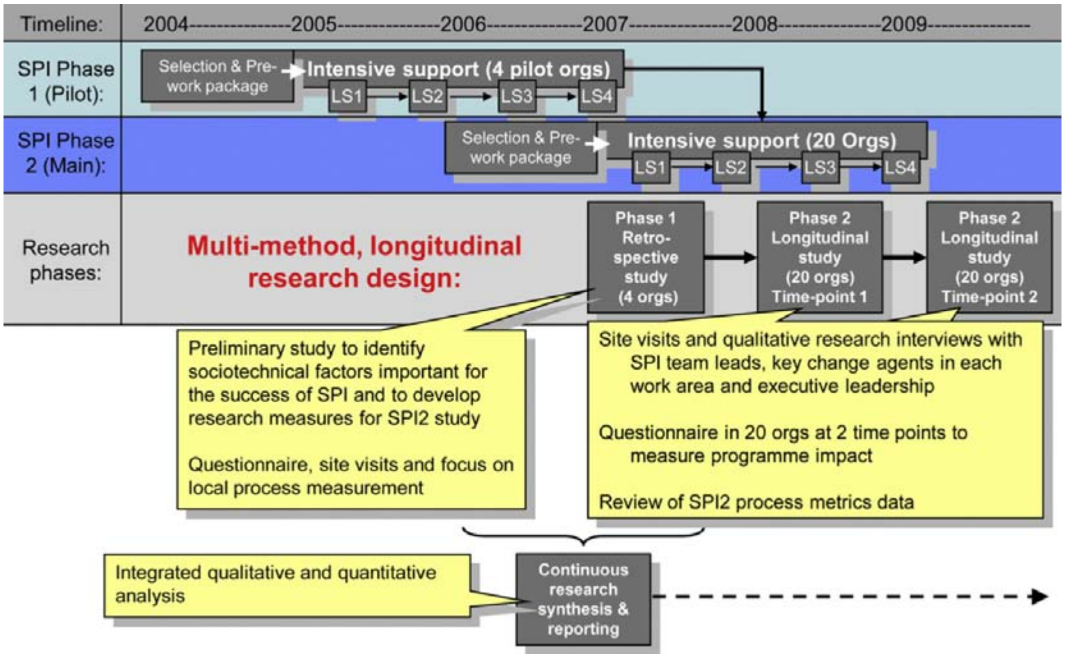
What are the effects of the first phase of a multi-component safer patients initiative (SPI) on four participating hospitals, each in different countries in the UK.

The intervention evaluated

Multi- component organisational intervention similar to the USA IHI 100k lives campaign using quality improvement breakthrough methods, with interventions to improve specific care processes in certain clinical specialties and to promote organisational and cultural change.

The design

Mixed method before and after evaluation with five sub studies with 4 intervention hospitals compared to 18 control hospitals.


Timeline:
2004--------------2005--------------2006-------------2007--------------2008--------------2009--------------
SPI Phase 1 (Pilot):
Selection & Pre-work package
Intensive support (4 pilot orgs)
LS1
LS2
LS3
LS4
SPI Phase 2 (Main):
Selection & Pre-work package
Intensive support (20 Orgs)
LS1
LS2
LS3
LS4
Research phases:
Multi-method, longitudinal research design:
Phase 1 Retro-spective study (4 orgs)
Phase 2 Longitudinal study (20 orgs) Time-point 1
Phase 2 Longitudinal study (20 orgs) Time-point 2
Continuous research synthesis & reporting

Preliminary study to identify sociotechnical factors important for the success of SPI and to develop research measures for SPI2 study

Questionnaire, site visits and focus on local process measurement

Site visits and qualitative research interviews with SPI team leads, key change agents in each work area and executive leadership

Questionnaire in 20 orgs at 2 time points to measure programme impact

Review of SPI2 process metrics data

Integrated qualitative and quantitative analysis

Outcomes measured and discovered

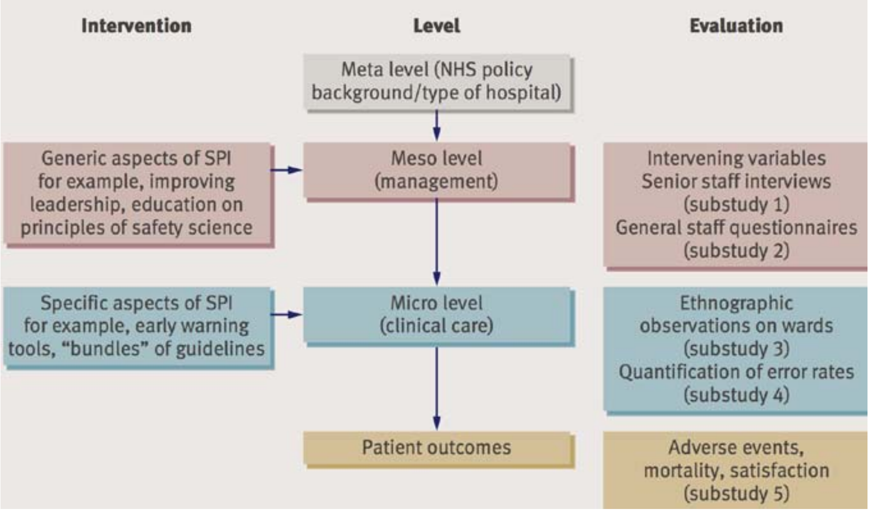
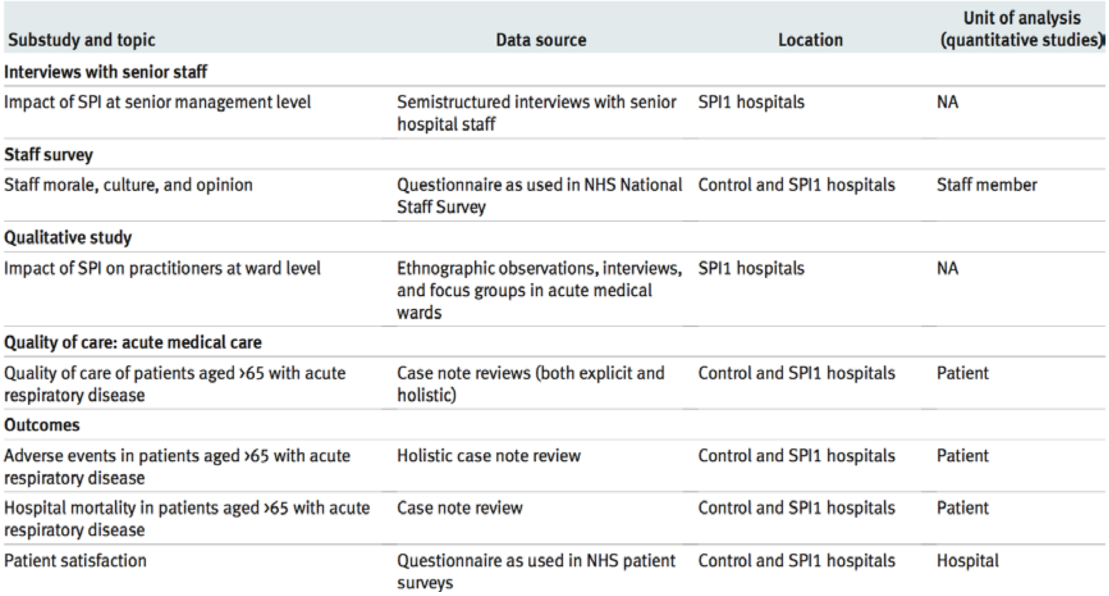

Intervention
Level
Evaluation
Meta level (NHS policy background/type of hospital)
Generic aspects of SPI for example, improving leadership, education on principles of safety science
Meso level (management)
Intervening variables Senior staff interviews (substudy 1) General staff questionnaires (substudy 2)
Specific aspects of SPI for example, early warning tools, "bundles" of guidelines
Micro level (clinical care)
Ethnographic observations on wards (substudy 3) Quantification of error rates (substudy 4)
Patient outcomes
Adverse events, mortality, satisfaction (substudy 5)

Table 2 | Summary of substudies comprising evaluation of phase one of Safer Patients Initiative (SPI1)



Substudy and topic	Data source	Location	Unit of analysis (quantitative studies)
Interviews with senior staff			
Impact of SPI at senior management level	Semistructured interviews with senior hospital staff	SPI1 hospitals	NA
Staff survey			
Staff morale, culture, and opinion	Questionnaire as used in NHS National Staff Survey	Control and SPI1 hospitals	Staff member
Qualitative study			
Impact of SPI on practitioners at ward level	Ethnographic observations, interviews, and focus groups in acute medical wards	SPI1 hospitals	NA
Quality of care: acute medical care			
Quality of care of patients aged >65 with acute respiratory disease	Case note reviews (both explicit and holistic)	Control and SPI1 hospitals	Patient
Outcomes			
Adverse events in patients aged >65 with acute respiratory disease	Holistic case note review	Control and SPI1 hospitals	Patient
Hospital mortality in patients aged >65 with acute respiratory disease	Case note review	Control and SPI1 hospitals	Patient
Patient satisfaction	Questionnaire as used in NHS patient surveys	Control and SPI1 hospitals	Hospital

Eleven of the National NHS Staff Survey questions were used at two time points to measure variables such as staff morale, attitudes, and aspects of "culture" that might be affected by the generic strengthening of organisational systems that SPI1.

For the qualitative study, one researcher undertook three rounds of data collection. Between April and September 2006, she visited one medical ward in each of the four hospitals for one week for 150 hours of ethnographic observations and 47 interviews with different types of ward staff, focusing on general issues relating to patient safety and SPI1. Medical case notes were reviewed for much of the other data.
Results Overall, the introduction of SPI1 was associated with improvements in one of the types of clinical process studied (monitoring of vital signs) and one measure of staff perceptions of organisational

climate. There was no additional effect of SPI1 on other targeted issues or on other measures of generic organisational strengthening.

Senior staff were knowledgeable and enthusiastic about SPI1. There was a small but significant effect in favor of the SPI1 hospitals in one of 11 dimensions of the staff questionnaire (organisational climate). Qualitative evidence showed only modest penetration of SPI1 at medical ward level. Of the five components to identify patients at risk of deterioration there was little net difference between control and SPI1 hospitals. Recording of respiratory rate increased; use of a formal scoring system for patients with pneumonia increased. There were no improvements in the proportion of prescription errors and no effects that could be attributed to SPI1 in non-targeted generic areas (such as enhanced safety culture). On some measures, the lack of effect could be because compliance was already high at baseline (such as use of steroids in over 85% of cases where indicated), but even when there was more room for improvement (such as in quality of medical history taking), there was no significant additional net effect of SPI1.

There were no changes over time or between control and SPI1 hospitals in errors or rates of adverse events in patients in medical wards. Mortality increased from 11% (27) to 16% (39) among controls and decreased from 17% (63) to 13% (49) among SPI1 hospitals, but the risk adjusted difference was not significant. Poor care was a contributing factor in four of the 178 deaths identified by review of case notes. The survey of patients showed no significant differences apart from an increase in perception of cleanliness in SPI1 hospitals.

<u>Certainty about and generalisablity of the outcomes</u>

The research used a pre-defined protocol, quantified safety practices and used independent case note reviewers who made observations across multiple hospitals (This reduced risk of bias in comparisons between institutions). Observations across the different levels within the hospitals made possible “triangulation”48 of data qualitative and quantitative collection and to increase the certainty about the findings.

The comparisons with control hospitals increased certainty that the outcomes were due to the intervention and not to other changes. Reproducing the principles and methods of the intervention would be possible as many details are presented in other documents, but whether the results would be similar is unlikely as the context would be different with different quality payment and other changes.
Strengths and weaknesses of the design in this example

This design involved comparisons with “control” hospitals and used a before and after “difference in difference” approach to quantitative measurements. Most other large scale evaluations rely on the hospitals own quality project’s data collection, but this study made more rigorous research assessments. One weakness, apart from the cost of the study, was that randomization was not used: SPI1 hospitals might have had less room for improvement, and controls might have had higher than average performance, particularly as half were also selected as future SPI2 intervention sites. The SPI1 hospitals were selected, not chosen at random and agreement to participate in the evaluation could have had a motivating effect in SPI1 hospitals compared to control hospitals.

Example 6. Process evaluation of an intervention to promote smoking cessation

Zapka, J Valentine-Goins, K Pbert L Ockene K 2004 Translating Efficacy Research to Effectiveness Studies in Practice: Lessons From Research to Promote Smoking Cessation in Community Health Centers Health Promot Pract 2004; 5; 245-255.

<u>Question addressed</u>

How was the "Quit Together" an intervention implemented to improve smoking cessation and relapse prevention among low-income pregnant and postpartum women who receive care at community health centers (CHCs).

<u>The intervention evaluated</u>

A theory- and evidence-based intervention with three components: 1) a provider-delivered smoking intervention with providers trained to deliver guideline- based, tailored counseling at all visits; 2) prompts to providers using a office practice management system tailored to each clinic (to screen for smoking status, deliver the intervention, document the encounter, distribute materials, and arrange follow-up for patients); 3) processes to facilitate systematic communication and documentation linkages between clinics to promote consistency and continuity of the smoking cessation strategies for women.

<u>The design</u>

Process evaluation of implementation participation and interactions. Smoking cessation services at the 6 comparison "usual care" sites were clinical intervention without protocols. A separate controlled trial was carried out to assess outcome.

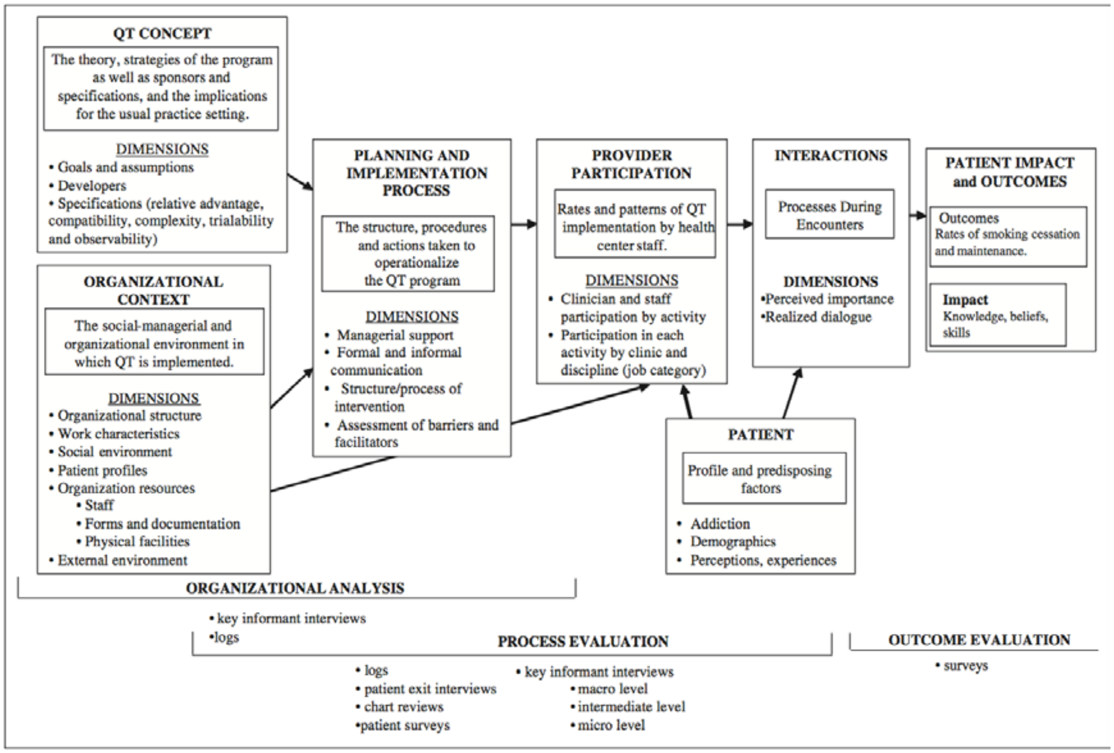
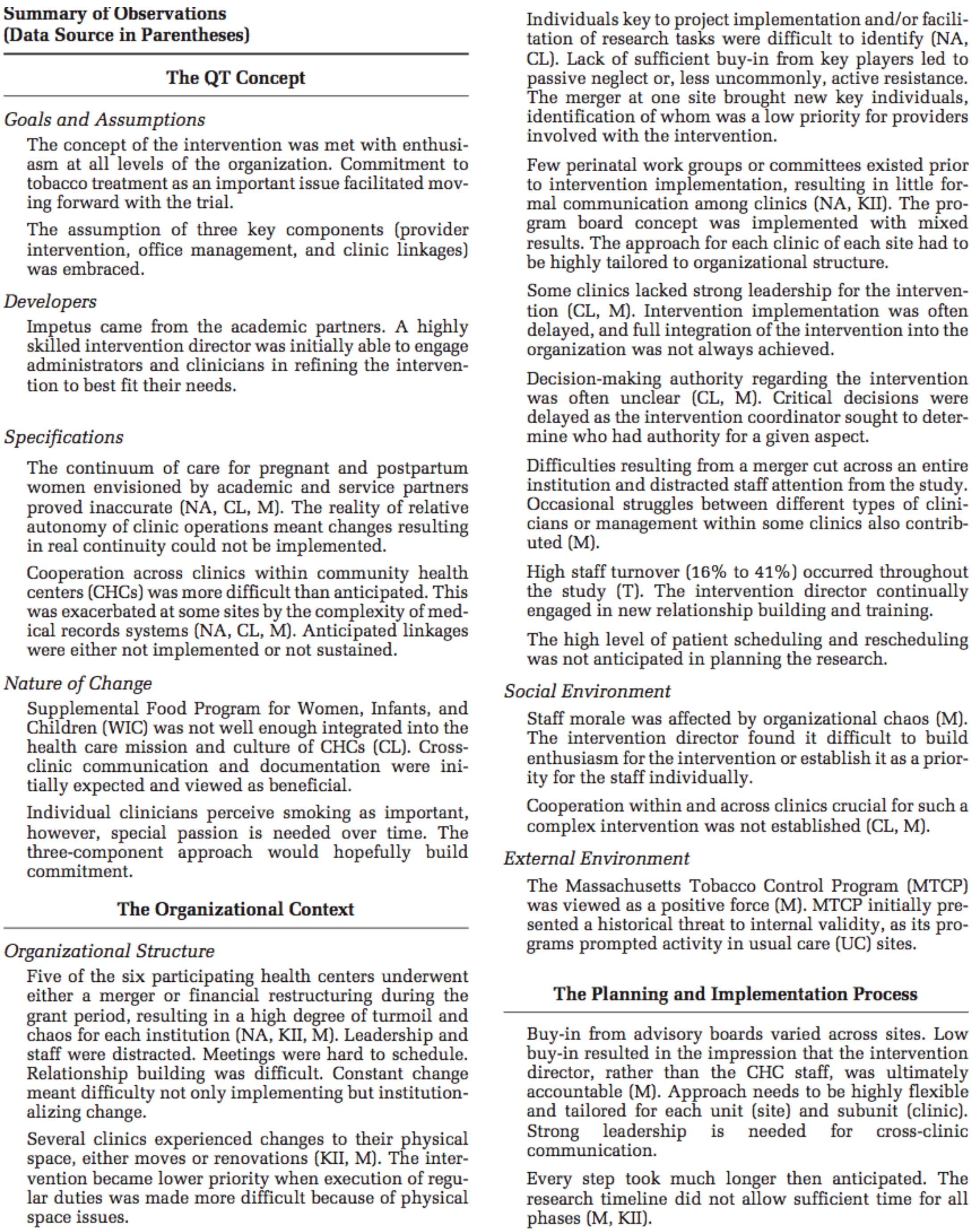

QT CONCEPT
The theory, strategies of the program as well as sponsors and specifications, and the implications for the usual practice setting.
DIMENSIONS
• Goals and assumptions
• Developers
• Specifications (relative advantage, compatibility, complexity, trialability and observability)
ORGANIZATIONAL CONTEXT
The social-managerial and organizational environment in which QT is implemented.
DIMENSIONS
• Organizational structure
• Work characteristics
• Social environment
• Patient profiles
• Organization resources
• Staff
• Forms and documentation
• Physical facilities
• External environment
PLANNING AND IMPLEMENTATION PROCESS
The structure, procedures and actions taken to operationalize the QT program
DIMENSIONS
• Managerial support
• Formal and informal communication
• Structure/process of intervention
• Assessment of barriers and facilitators
PROVIDER PARTICIPATION
Rates and patterns of QT implementation by health center staff.
DIMENSIONS
• Clinician and staff participation by activity
• Participation in each activity by clinic and discipline (job category)
INTERACTIONS
Processes During Encounters
DIMENSIONS
•Perceived importance
•Realized dialogue
PATIENT IMPACT and OUTCOMES
Outcomes
Rates of smoking cessation and maintenance.
Impact
Knowledge, beliefs, skills
PATIENT
Profile and predisposing factors
• Addiction
• Demographics
• Perceptions, experiences
ORGANIZATIONAL ANALYSIS
• key informant interviews
•logs
PROCESS EVALUATION
• logs
• patient exit interviews
• chart reviews
•patient surveys
• key informant interviews
• macro level
• intermediate level
• micro level
OUTCOME EVALUATION
• surveys

FIGURE 1 **Conceptual Model of Relationships Among Constructs Assessed and Integrated Into the Quit Together (QT) Program**

Outcomes measured and discovered

Process data were collected through eight methods:

- Organizational assessment (OA) reports. Collected from key informant interviews about program/clinic flow, structure, and process linkages among the clinics and key players. The interviews were analyzed to identify potential barriers and facilitators to integration of the intervention into each clinic of each SI health center's operations.
- Contact logs. A semi structured form to record the intervention coordinator's meetings with key personnel at each SI clinic, to track the expectations from each meeting and whether follow-up occurred. (It was reported that, from this, the coordinator "identified the underlying problem of a power struggle between two key individuals. She then tailored her actions to move the project forward").
- Meeting notes. For meetings of the program boards, clinic work groups, and individual personnel key to the project, recorded type of meeting, participants, agenda, issues discussed, and decisions made.
- Training records. A Microsoft Access database was developed to track all SI providers trained.
- Key informant interviews (KIIs). Follow-up interviews halfway and at the end of the trial.
- Patient exit interviews (PEI). After every WIC nutrition appointment.
- Patient report of exposure in periodic surveys. At the final prenatal interview, questions prompted the patient to think about her most recent visit to the OB clinic. At the postpartum interviews, questions asked the patient to confirm implementation of the intervention protocol steps at her most recent pediatric visit.
- Chart reviews. Clinic staff at SI sites (a) placed a cue on each woman's chart to remind providers to intervene for smoking and (b) placed an algorithm inside the chart for providers to record what they had done. A random sample of charts of women eligible for the study was examined to confirm presence of the cue and algorithm in the record and completeness of documentation.



Summary of Observations
(Data Source in Parentheses)

The QT Concept

Goals and Assumptions

The concept of the intervention was met with enthusiasm at all levels of the organization. Commitment to tobacco treatment as an important issue facilitated moving forward with the trial.

The assumption of three key components (provider intervention, office management, and clinic linkages) was embraced.

Developers

Impetus came from the academic partners. A highly skilled intervention director was initially able to engage administrators and clinicians in refining the intervention to best fit their needs.

Specifications

The continuum of care for pregnant and postpartum women envisioned by academic and service partners proved inaccurate (NA, CL, M). The reality of relative autonomy of clinic operations meant changes resulting in real continuity could not be implemented.

Cooperation across clinics within community health centers (CHCs) was more difficult than anticipated. This was exacerbated at some sites by the complexity of medical records systems (NA, CL, M). Anticipated linkages were either not implemented or not sustained.

Nature of Change

Supplemental Food Program for Women, Infants, and Children (WIC) was not well enough integrated into the health care mission and culture of CHCs (CL). Cross-clinic communication and documentation were initially expected and viewed as beneficial.

Individual clinicians perceive smoking as important, however, special passion is needed over time. The three-component approach would hopefully build commitment.

The Organizational Context

Organizational Structure

Five of the six participating health centers underwent either a merger or financial restructuring during the grant period, resulting in a high degree of turmoil and chaos for each institution (NA, KII, M). Leadership and staff were distracted. Meetings were hard to schedule. Relationship building was difficult. Constant change meant difficulty not only implementing but institutionalizing change.

Several clinics experienced changes to their physical space, either moves or renovations (KII, M). The intervention became lower priority when execution of regular duties was made more difficult because of physical space issues.

Individuals key to project implementation and/or facilitation of research tasks were difficult to identify (NA, CL). Lack of sufficient buy-in from key players led to passive neglect or, less uncommonly, active resistance. The merger at one site brought new key individuals, identification of whom was a low priority for providers involved with the intervention.

Few perinatal work groups or committees existed prior to intervention implementation, resulting in little formal communication among clinics (NA, KII). The program board concept was implemented with mixed results. The approach for each clinic of each site had to be highly tailored to organizational structure.

Some clinics lacked strong leadership for the intervention (CL, M). Intervention implementation was often delayed, and full integration of the intervention into the organization was not always achieved.

Decision-making authority regarding the intervention was often unclear (CL, M). Critical decisions were delayed as the intervention coordinator sought to determine who had authority for a given aspect.

Difficulties resulting from a merger cut across an entire institution and distracted staff attention from the study. Occasional struggles between different types of clinicians or management within some clinics also contributed (M).

High staff turnover (16% to 41%) occurred throughout the study (T). The intervention director continually engaged in new relationship building and training.

The high level of patient scheduling and rescheduling was not anticipated in planning the research.

Social Environment

Staff morale was affected by organizational chaos (M). The intervention director found it difficult to build enthusiasm for the intervention or establish it as a priority for the staff individually.

Cooperation within and across clinics crucial for such a complex intervention was not established (CL, M).

External Environment

The Massachusetts Tobacco Control Program (MTCP) was viewed as a positive force (M). MTCP initially presented a historical threat to internal validity, as its programs prompted activity in usual care (UC) sites.

The Planning and Implementation Process

Buy-in from advisory boards varied across sites. Low buy-in resulted in the impression that the intervention director, rather than the CHC staff, was ultimately accountable (M). Approach needs to be highly flexible and tailored for each unit (site) and subunit (clinic). Strong leadership is needed for cross-clinic communication.

Every step took much longer then anticipated. The research timeline did not allow sufficient time for all phases (M, KII).

Results: The main findings were about challenges to implementation of the QT strategy which explained program implementation failure, specifically concerning organizational context, planning and implementation.

<u>Certainty about and generalisablity of the outcomes</u>

The process evaluation triangulated and compared data from the KII, logs, and chart reviews were examined across clinic types (OB, PED, and WIC) for each site, across sites for each clinic type, and across intervention condition (SI vs. UC). ("Constant comparison method" (Lincoln & Guba, 1985). These methods increased the certainty about the findings.

Strengths and weaknesses of the design in this example

Relevant data from the PEIs and periodic patient surveys also were reviewed. Observations are reported according to the key domains identified in the pre-study theory of the intervention bearing in mind three types of potential failure, theory, program implementation, and measurement failure. Translating the contextual factors into lessons for practitioners was conducted considering these three areas.

This design was able to examine context factors affecting complex programmes of this type and show the "lessons learned" about barriers and approaches needed for implementation in similar programmes.

Example 7. Case evaluation of a programme for electronic summaries of patients' medical records (mixed-methods)

Greenhalgh, T Stramer, K Bratan, T Byrne, E Russell, J Potts, H 2010 Adoption and non-adoption of a shared electronic summary record in England: a mixed-method case study BMJ 2010;340:c3111

Question addressed

- What is the usability, use, functionality, and impact of the summary care record (SCR), and what explains variation in its adoption and use?
- How was the SCR programme shaped by influences at the macro, meso, and micro level?
- What are the transferable lessons for practice and policy?

The intervention evaluated

Implementation of a SCR (a structured summary held on a national database and accessible to authorized staff over a secure internet connection) as part of a wider programme to implement an electronic medical record in the English public National Health Service (NHS)

The design

Case evaluation, using mixed methods informed by "utilization-focused evaluation," and a method for interpretive field studies in large scale information systems (described in detail in Greenhalgh 2010). The former approach views complex programmes as having multiple stakeholders, each with different expectations of the programme and the evaluation (Patton 1997), and the latter approach is to make continuous, iterative comparison of findings in one part of the project with an emerging description of the whole programme (Klein and Myers' 1990).

The design aimed to describe the intervention at a macro level (e.g. national policy, wider social norms and expectations), a meso level (e.g. organisational processes and routines), and a micro level (e.g. particular experiences of patients and professionals) using qualitatively and quantitative data.

Outcomes measured and discovered

The findings reported included narrative descriptions of "What stakeholders expected of the summary care record" and "Implementing the programme" - the different actions and experiences of participants. Data on use and non-use of summary care records was quantitative data: those collected and reported included, proportion of patients with a summary care record; access rates to this record by providers (low); trends in accesses; and impact on consultation times. Qualitative data included interviews with clinicians (which were used to explain the above); ethnographic data on 214 clinical consultations showed that non-access of the SCR had many different reasons, and benefits associated with summary care record use.

Certainty about and generalisablity of the outcomes

The mixed methods and theory-informed design made it possible to explain the low use and identify improvements needed. The understanding of the programme was enhanced by apply socio-technical network theory to the data to view it a complex, dynamic, and unstable socio-technical network with multiple interacting sub-networks with individuals and organizations representing four different institutional “worlds”—political, clinical, technical, and commercial.

Strengths and weaknesses of the design in this example

The literature review allowed pre-study identification of issues to explore and building a theoretical model to inform data collection. The researchers propose that “This enabled us to combine qualitative and quantitative techniques to highlight the competing conceptualizations and complex interdependencies of the SCR programme and to bring into frame numerous social, technical, ethical, and political explanations for why particular goals and milestones set by policy makers and implementation teams were or were not reached”.

A possible weakness is the credibility to the users of the evaluation of this design. The researchers suggest that some healthcare information systems researchers define rigor in terms of experimental or quasi-experimental studies of the “deployment” of a technology and its “impact” on predefined outcomes, ”The profound epistemological differences and lack of dialogue between healthcare information systems research (led largely by doctors with an interest in information technology) and mainstream information systems research (led by interdisciplinary teams of organisational sociologists, computer scientists, and political scientists) have been highlighted previously”.

Example 8. Formative realist case evaluation of large scale "transformation" of health services in London

Greenhalgh, T Humphrey, C Hughes, J Macfarlane, F Butler, C Pawson, R 2009 How Do You Modernize a Health Service? A Realist Evaluation of Whole-Scale Transformation in London, The Milbank Quarterly, Vol. 87, No. 2, 2009 (pp. 391–416).

See also: Greenhalgh, T., C. Humphrey, J. Hughes, F. Macfarlane, C. Butler, P. Connell, and R. Pawson. 2008. The Modernisation Initiative Independent Evaluation: Final Report. London: University College London. Available at http://www.ucl.ac.uk/openlearning/research.htm. Ref Type: Report (accessed March 30, 2009).

Question addressed

"What works, for whom, under what circumstances?"

"What are transferable lessons about effective change?"

The intervention evaluated

A change programme to a public healthcare system to make stroke, kidney, and sexual health services more efficient, effective, and patient centered.

The design

Formative realist case evaluation, informed by realist context-intervention-mechanism theory, with mixed data sources and collection methods of ethnographic observation, semi-structured interviews, and document analysis used in a "pragmatic and reflexive manner to build a picture of the case and follow its fortunes over the three-year study period".

The design uses inductive interpretation of the data to build a theory, involving methods to increase validity like noting participants reactions to interpretations, triangulating data sources, and seeking disconfirmations for initial interpretations. The aim is to identify which mechanisms are triggered in certain contexts by the intervention to then result in service changes which were thought to improve patient outcomes. The steps through which this is done are:

- Organizing primary data and producing preliminary thematic summaries of these,
- Repeating this at a later time to document changes,
- Presenting, defending, and negotiating particular interpretations of actions and events both within the research team and also to the stakeholders themselves,
- Testing these interpretations by explicitly seeking disconfirming or contradictory data,
- Considering other interpretations that might account for the same findings,
- Using cross-case comparisons to determine how the same mechanism (such as "integrating services across providers") or sub mechanism (such as "introducing boundary-spanning roles") plays out in different contexts and produces different outcomes, thereby allowing inferences about the "generative causality" of different contexts.

Outcomes measured and discovered

Six "Mechanisms of change" and associated "lessons" for others seeing similar changes.

Mechanism 1: Integrating Services across Providers: through establishing boundary-spanning roles, shared guidelines, protocols, pathways, IT systems and common data sets, developing networks and supporting networking.

Lessons: efforts to achieve integration across providers are more likely to succeed when

- Relationships between organizations have trust, a history of collaboration, and compatibility of values rather than mismatch or competition,
- Approaches to integration are imaginative, locally responsive, negotiable, and supported by technology rather than rigid and driven by technology.
- External incentives (e.g., policies) are designed to reward collaborative performance and do not pit organizations against one another.
- The strategy for integration includes both "soft" and "hard" approaches rather than focusing exclusively on one or the other of these.
- Solutions are participatory rather than developed by one party and imposed on others

The following represents the context – mechanism - outcome understanding presented by the study for the first of the six mechanisms (the context is "enabling and constraining factors that appeared to make each mechanism more or less likely to produce a desired outcome in any particular set of circumstances").

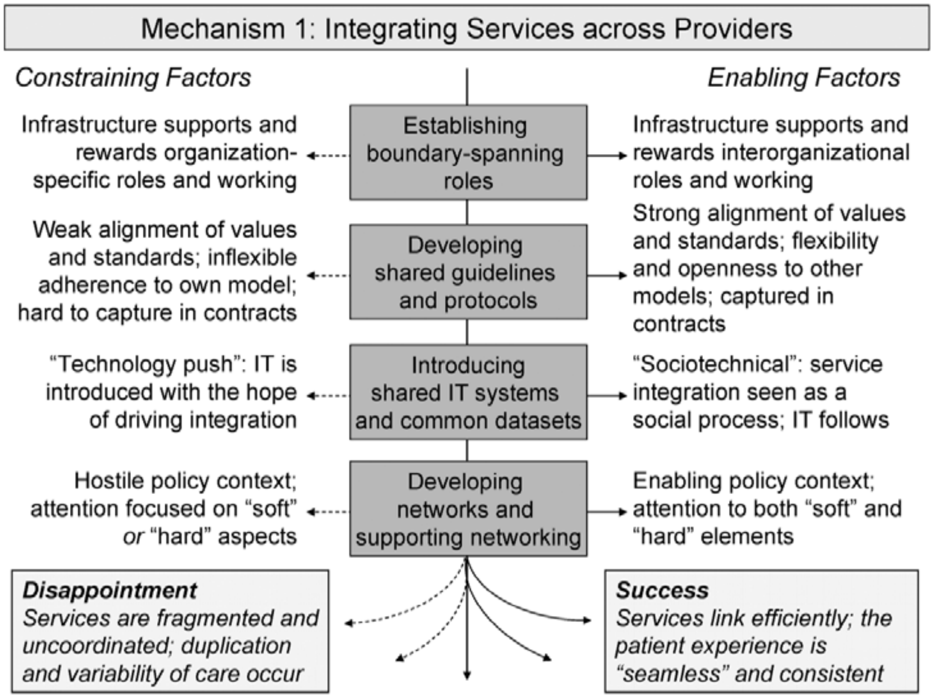

Mechanism 1: Integrating Services across Providers
Constraining Factors
Enabling Factors
Infrastructure supports and rewards organization-specific roles and working
Establishing boundary-spanning roles
Infrastructure supports and rewards interorganizational roles and working
Weak alignment of values and standards; inflexible adherence to own model; hard to capture in contracts
Developing shared guidelines and protocols
Strong alignment of values and standards; flexibility and openness to other models; captured in contracts
"Technology push": IT is introduced with the hope of driving integration
Introducing shared IT systems and common datasets
"Sociotechnical": service integration seen as a social process; IT follows
Hostile policy context; attention focused on "soft" or "hard" aspects
Developing networks and supporting networking
Enabling policy context; attention to both "soft" and "hard" elements
Disappointment
Services are fragmented and uncoordinated; duplication and variability of care occur
Success
Services link efficiently; the patient experience is "seamless" and consistent

FIGURE 1. Realist Analysis of Attempts to Modernize by Integrating Services across Providers

The other mechanisms of change also had associated “lessons” but only three were presented in the study “for length of article reasons”. (Mechanism 2: Finding and Using Evidence, Mechanism 3: Involving Service Users in the Modernization Effort) Mechanism 4: supporting self-care, Mechanism 5: developing the workforce, and Mechanism 6: extending the range of services.

Outcomes are described in more detail in another report: “hard” (evidence-based protocols; extended opening hours; shorter waiting lists; governance structures for inter-organisational working) and “soft” (improved staff attitudes and motivation, greater user satisfaction, and what one senior manager described as a “precious, extraordinary” cultural shift) (Greenhalgh et al 2008).

Certainty about and generalisablity of the outcomes

The certainty about the outcomes is maximized if we trust that the researchers appropriately applied the accepted methods of interpretive case study, with a team culture which challenged each other to apply these methods: carefully defining and justifying the “case,” “immersion in the case” (i.e., spending enough time at the field site to understand what is going on), systematic data collection and analysis, reflexivity in both researchers and research participants, developing theory iteratively as emerging data are analyzed, seeking disconfirming cases and alternative explanations, and defending interpretations to both the research participants and academic peers.

The generalisablity is increased by providing details of the change actions and of the context to allow others to assess whether if they applied these actions in their setting, they are likely to get similar results. The description of the “mechanisms” or change principles and how they were triggered by the intervention in the setting enables others to try to reproduce the mechanism in their different setting using possibly different interventions.

The discussion section of the study elaborates on the design, and proposes “a greater understanding of these underlying mechanisms will help inform similar change programs in the future” and that the researchers “deliberately not passed judgment on the program’s overall “success.”…”If a team sets out to achieve X but along the way learns things or encounters challenges that convince it that Y is a more appropriate (or practicable) goal, then it will have “succeeded” if it achieves something approaching Y.” and go on to comment that, “Building the evaluation criteria for expensive, large-scale change programs on such shifting sands is relatively controversial when judged by conventional clinical research criteria, which define “rigor” as the systematic pursuit of well-defined goals, objective measurement of progress, and robust accountability procedures”.

Strengths and weaknesses of the design in this example

The design gives a narrative understanding of the process of change in a context, rather than showing a few measured outcomes attributed to a change. It can answer some questions about implementation and necessary conditions, but does not give clear-cut answers to the “did patients benefit” question, because other changes are not controlled for and research resources and data focus was on process not later end-outcomes. The design is also dependent on the researchers experience and skills with the interpretive methods, and the rigor with which the research team apply the methods, which is true for all designs but perhaps more so with this type of design.

The discussion comments on the methods, as it is one of the first studies to use a design of this type. The researchers report that identifying the mechanisms of change was far more difficult than the text book on realist evaluation suggests (Pawson & Tilley 1997). They concluded that researchers “must anticipate the mismatch between the realist evaluation’s assumption that a set of more or less well-defined “mechanisms of change” can be articulated and tested and the empirical reality in which these mechanisms may prove stubbornly hard to nail”. Alternative “mechanisms of change” explanations could be proposed and it was not possible in this study to decide which was operating.

Others using this design found that “surfacing theories of change,” did not allow understanding of the power dynamics among different groups, and recommended the design also draws on other theories to document and understand the politics of change (e.g. neo-institutional theory (Barnes, Matka, and Sullivan 2003)). A further comment is about the participatory or action role of the researcher-evaluators (Øvretveit 2002, Patton 1997, Guba & Lincoln 1989). In this study the researchers not only fed back to implementers interim findings, they also discussed and tested the ideas about mechanisms and context with actors which also was likely to change the programme. The researchers argue that the design is thus incompatible with a detached objectivist approach to research.

Moore (2012) gives a summary of limitations of and issues with the approach:

- Distinguishing between context and mechanism

- Is there an endpoint? When to “close”
- Time-consuming
- Can realist evaluation studies be replicated?
- How does Realist Evaluation differ from traditional mediator/moderator analysis?

Appendix 3: Discussion of observational-qualitative or mixed method designs

There are a range of observational-qualitative or mixed method designs (termed “naturalistic” approaches), used to evaluate CSI effectiveness and answer other decision makers questions. Program and case evaluation include, at one extreme, inductive approach which build theory of how the intervention has its effects from the data gathered, usually qualitative. Deductive approaches start with theory and test or explore theoretical propositions in the study. There are also or “snowball” approaches which draw-in and test different theories about how the intervention is effective during the research, building and using theory in interaction with and during the data gathering. In addition program and case evaluation approaches can be classified as positivist and objectivist, assuming causal mechanisms and using quantitative measures; or as subjectivist and qualitative, assuming “interventions” work through people who make choices about whether to respond depending on their interpretations, motives and values, and influenced by culture and other social factors.

Programme- and process- evaluation

Program evaluation is a term used to describe evaluations of social and policy interventions aimed at populations, such as educational and public health service programs, as well as programs whose focus is on organizations, such as accreditation programs. Most program evaluations use quasi-experimental designs, with before/after data gathering and control or comparison groups to answer both effectiveness questions and questions about whether the program had achieved its goals.

Since 1980 there have been developments in this field which are relevant to answering questions about CSIs. For large scale social programs, quasi-experimental designs proved inconclusive in establishing associations between the program and client outcomes. This led to working back from client outcomes to establish whether the program had achieved intermediate changes and effects which were thought necessary for intended client outcomes. More attention was given to describing the extent to which the original program plan was implemented (e.g., through “process evaluation”, sometimes used to develop the program as it is implemented, as in “action evaluation” (Øvretveit 2002), and “formative evaluation” (Shadish et alShadish et alShadish et alShadish et alShadish et alShadish et alShadish et al and identification of intermediate and unanticipated effects.

The second and related development increased emphasis on program theory (or logic model):

> *“…Program evaluation identified that its major purpose was to examine the theory or conceptual basis of the program. Comprehensive evaluations address the theory by carefully defining the components of the program and their relationships, and then examining the implementation of these components and how they mediate outcomes.”*
>
> Bickman (2008)

Experimental designs use “theory” in the sense that the evaluation is designed as a prospective test of a hypothesis, In contrast, in theory-informed program evaluation the program theory is either a prospective model of how the components lead to the intended results, or a retrospective explanation of how or why the program progressed as it did (Bickman 1987, Weiss 1997a,b). This is sometimes termed a "logic model" (Wholey 1983) which proposes a chain of events over time in cause-effect patterns in which the dependent variable (event) at an earlier stage becomes the independent variable (causal event) for the next stage. An important difference from experimental designs is that influences other than the program (exogenous influences) are assessed for their influence on the program outcomes, i.e., the program is only one of a number of independent variables that are examined for their influence on the dependent variables.

A number of researchers emphasize the value of process or theory-informed program evaluations in parallel to controlled trials (Bradey et al. 1999, Byng et al. 2005), because the former research allows an explanation of the results found in the controlled trial and adds answers to questions which controlled trials cannot answer.

Examples of studies using these methods to study safety, quality or other interventions to organizations include those by Bradey et al. (1999), Rousseau et al. (2003), and Foy et al. (2005).

Case study evaluation (CSE)

Case study methods were not originally developed for evaluation. They are used to, "investigate a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" (Yin 2003). Yin also proposes case studies use multiple sources of data and triangulation to increase validity, and "prior development of theoretical propositions to guide data collection and analysis." Case studies do not standardize the content or implementation of a CSI, but document how it is carried out, and various outcomes in ordinary settings. Usually there is no control group or site, but there may be comparison or multiple cases to understand the CSI in different settings. Yin represents the deductive theory testing school, whereas Stake (1995) represents the inductive school within this approach which argues for the value of theory building about the intervention and effects as part of data gathering.

Keen & Packwood (1995) describe case study evaluation (CSE) as,

> *"valuable where broad, complex questions have to be addressed in complex circumstances…when the question being posed requires an investigation of a real life intervention in detail, where the focus is on how and why the intervention succeeds or fails, where the general context will influence the outcome and where researchers asking the questions will have no control over events. As a result, the number of relevant variables will be far greater than can be controlled for, so that experimental approaches are simply not appropriate."*

Some case study approaches are based on the assumption that experimental methods successful for assessing effects of discrete drug or surgical interventions on human bio-physiology are not suitable for assessing effects of social interventions on clinicians' behavior or social organizations. These approaches start from the premise that experimental methods are based on assumptions about causality in physical objects which are not suited to understanding social change, where effectiveness depends on how actors choose to interpret and respond to the intervention within a social culture and system.

Examples of studies using these methods to study safety, quality or other interventions to organizations include those by Walshe & Shortell (2004), Øvretveit & Aslaksen (1999), Ovretveit et al. (2007).

Realist evaluation

Realist evaluations identify context-mechanism-outcome (CMO) configurations in complex interventions in different settings, and aim to establish “what works for whom in which settings” (Pawson & Tilley 1997). The assumption, based on some evidence from education and criminal programs, are that, in social interventions, outcomes are a function of “the mechanism” through which the intervention works, and the context (which includes multiple levels) in which it is applied: there are different effects in different settings even if the same intervention is used.

The aim is not only to describe the intervention but to clarify the “generative mechanism”: the essential idea or “active ingredient” which is the basis for the intervention (e.g., performance feedback). Using this approach, superficially different interventions can be grouped and compared through their underlying logic. Another aim is to examine how much and how the mechanism depends on or interacts with the context to produce different effects.

The aims are to use a program model or theory (sometimes called a logic model) to select programs and test hypotheses about CMO configuration for one intervention in one setting, and then to study “similar” programs in other settings to examine how the interactions between C, M, and O vary. Interventions are viewed as “theories in practice”. Discovering poor or no outcomes of a similar intervention in a different setting is an opportunity to refine the logic model of the CMO configuration.
This approach thus emphasizes studying in a variety of situations the mechanism which is thought to generate certain results rather than one intervention at one site. The approach is similar to case study in describing and understanding outcomes as the product of an intervention implementation in context, but differs from some case study evaluations in emphasizing the logic model testing and the comparison between different implementations, as well as elucidating the essential feature of the mechanism to allow comparison of a variety of superficially different changes.

A possible limitation of the realist approach for studying PSPs are that the concepts of context, mechanism and outcome are not well defined and only illustrated in a few studies. It is also unclear exactly how “mechanism” is elucidated: “mechanism” does not just describe the intervention components or implementing actions, but how this higher level conceptualization of “mechanism” is created is unclear - of how the actions work (“generative mechanism”), which is different from their interaction with context.

Examples of studies using these methods to study safety, quality or other interventions to organizations include those by Redfern et al (2002), Blaise & Kegels (2004), Byng et al (2005 and 2008), Kennedy et al

(2005) and Greenhalgh et al (2009). Some limitations are described by Davis (2005).

Resource requirements

The financial costs of these approaches varies: limited case study evaluations relying on interviews for descriptions of intervention content, implementation and assessments of outcome can be completed within 6 months and for less than $100,000 for one case. Normally, however, these approaches require a minimum of 12 months to gather and analyze data from a variety of sources and to relate the data to previous research, involving costs of $100,000 - $300,000 per year or over depending on how many cases.

As regards researcher skills resources, qualitative researchers and those who are able to integrate mixed methods data and relate to theory, and who are familiar with health services, are in short supply. Often teams are required with different skills, where the skill and experience of the team leader to integrate the inputs and relate to theory is the critical resource.

Expertise and experience is necessary for high quality research using these methods because detailed procedures have not been formulated and because of the importance of relating theory to the data collected.

As regards realist evaluation, a study of one CMO configuration might be considered of limited use: the value of the approach is comparing CMO combinations in different settings in a program of research. This would require substantial resources as well as specific skills which are not common in health researchers. Researchers in this field have explored “realist reviews” of studies which allow some comparison between CMO configurations which researchers, usually different teams, have studied (Pawson et al. 2005, Greenhalgh et al. 2007). However little research has been carried out and published which allow such comparisons or reviews, especially in the field of patient safety.

Theory in evaluating CSIs: logic models program theory and other types of theory

Theories at two levels of abstraction are relevant for evaluating CSIs.

The first higher level theory is what is termed the “logic model” or “program theory” of the intervention. These are different ways of representing the assumptions about the influences which lead to certain outcomes: the sequence of actions and intermediate and later effects expected.

A logic model describes how an intervention is understood or intended to produce particular results (Rogers 2005). The logic model proposes a chain of events over time in cause-effect patterns in which an event at an earlier stage causes the next stage (Wholey 1983). The logic model is the simplest of models as it usually only describes inputs, activities and immediate and later effects. Although logic models often do not delineate theories they usually draw on assumptions or an implicit theory of behavior or organizational change, which are lower level of abstraction (more concrete) theories discussed later in this section.

"Treatment theory" describes the process through which an intervention is expected to have its effects on a specified target population," In the case of CSIs, the target population is usually providers or organizations (Lipsey 1993). This theory is not a protocol that requires very specific prescribed actions. Instead it is a set of principles that together are hypothesized to bring about change in the particular situation. These principles might be enacted in several different ways, but they would all achieve the same "functions" (Hawe et al 2004) and intermediate objectives in a chain of events which leads ultimately to improved patient outcomes.

In the field of program evaluation, program theory is defined as the "conceptual basis" of the program: "Comprehensive evaluations address the theory by carefully defining the components of the program and their relationships, and then examining the implementation of these components and how they mediate outcomes" (Bickman 2008). Experimental designs use "theory" in the sense that the evaluation is designed as a prospective test of a hypothesis. In contrast, in theory-informed program evaluation, the program theory is either a prospective model of how the components lead to the intended results, or a retrospective explanation of how or why the program progressed as it did (Bickman 1987, Weiss 1997a,b).

A "theory of change" is usually used to describe how those responsible for implementation understand an intervention to work (Sullivan & Stewart 2006, Mason & Barnes 2007, Connell & Klem 2000). It may be explicit, or may exist as a theory in a sense of being unspoken assumptions or beliefs. Dixon-Woods et al. 2010 describe a theory of change as identifying "plans for change and how and why those plans are likely to work, and indicates the assumptions and principles that allow outcomes to be attributed to particular activities." This is different from an explanation derived from empirical research on possible influences on outcomes.

At a lower and more concrete level there are other theories relevant to evaluating CSIs. These theories can be used either in forming a pre-data-gathering model to decide which data to collect to test hypotheses about how the change works or does not work, or post-data-gathering to explain effects discovered. Most such theories come from a variety of disciplines: psychology, sociology, anthropology, behavioral economics, and management sciences. A longer overview of these theories in quality improvement is given by Grol et al. (2007).

An illustration of one way to use theory in CSI evaluation is given in Dixon-Woods et al. (2010).
Many theories focus on the intervention and conceptualize it as a chain of events, often in a linear sequence, which leads through intermediate changes (including changes in provider and organizational behavior) to final results (clinical or cost outcomes). Other theories pay more attention to context factors and seek to understand or explain their influence on or interaction with the intervention, and may view the implementation as a number of interacting components with a synergistic and system effect.

Best practice in implementation as well as in evaluation proposes using theory to design the intervention and its implementation, in order to make it possible to revise theory in the light of results and thus build up understanding of what is critical to intervention success.

Improving evaluations of CSIs using theory is likely to call for greater collaboration with researchers from psychology, sociology and management sciences to choose and apply the theories before or after evaluation data gathering. CSI evaluation is likely to need inter-disciplinary collaborations that can bring in one or more theoretical perspectives.

Earlier research is useful for deciding which theories are most relevant to the evaluation of the CSI. General conceptual frameworks may be useful as a starting point: different reviews and consensus reports have integrated factors from individual theories into broader conceptual frameworks (Michie et al. 2005 (psychology), Ward et al. 2009, and Cabana et al. 1999). Tools for applying theories are discussed in volume 2 such as the theory of planned behavior, or normalization process theory (Francis et al. 2004, May et al. 2010).

Difference between a theory and a conceptual framework or model: Miles and Huberman (1994) define a conceptual framework as a representation of a given phenomenon that ''explains, either graphically or in narrative form, the main things to be studied—the key factors, concepts, or variables'' (p. 18) that comprise the phenomenon.

(Miles, M., & Huberman, M. (1994). Qualitative data analysis: An expanded sourcebook (2nd ed.). London: Sage)

7. Appendix: Conceptualising complex social interventions

There are three aspects to an intervention to a patient, population or organization which lead to defining the intervention as complex:

- the outcome of the intervention (the before/after difference it makes for the target it is applied to)
- the content of the intervention (the components or active ingredients)
- the implementation: the actions or strategies taken to change the target to produce the outcome

An intervention may be complex,

- in content (multiple component, such as a bundle for VAP) or
- in implementation (many different actions taken to put the components into practice, some possibly at different phases, for example, training, checklists, feedback, supervision, and reward programs).

Complex-content interventions are more challenging to evaluate for effectiveness using controlled trials, especially if the content needs to be adapted to local circumstances. Controls and randomization can be used (the intervention content does not need to be standardized) and this increases internal validity. However a detailed description of the content which was implemented is needed and this requires non-experimental research methods. In addition, it is useful to describe why these specific adaptations were made, which usually requires qualitative methods.

As regards evaluating the effectiveness of a simple implementation strategy such as training or providing written guidance, experimental designs can be used. However complex implementation strategies are more difficult to evaluate using such designs.

The complexity of the intervention which makes evaluation difficult is because the components may interact synergistically or in a negative way, or may need to be provided in different ways or different intensities in different situations. The actions to create change are often also multiple. There will need to be intermediate and later effect measures and at different levels of change. All this may be carried out in a changing health service or system and these changes may affect the components and the implementation.

These issues are examined in a literature which considers both experimental and other methods for evaluating complex social interventions (Craig et al. 2008, Wolff 2001, Oakley et al. 2006, Hawe et al. 2004, Judge 2000).

Conceptualization of a CSI and its effects – the naturalistic approach

The naturalistic approach to evaluating complex social interventions conceives of the intervention and its effects using these ideas:

<u>Intervention:</u> this is both the idea or physical artifact (“content of the intervention”) and the actions taken by “implementers” to put this into practice or to enable others to change (“intervention actions”).

> *A hospital medical emergency team (MET) is a set of ideas about providing specialist critical care competence to nurses and doctors who are concerned about a patient’s deterioration, when local expertise is not quickly available. To establish a MET a number of implementation actions need to be taken such as defining local criteria for calling a MET, agreeing how it will operate with local physicians, training personnel about when to call, and to establishing, and training the MET members.*

<u>Complex</u>: the intervention has a number of components and actions, changing over time and sometimes interacting with each other (synergistic).

> *The MET intervention involves training the team and personnel and a number of other actions. Over time the calling criteria or team membership may be changed to improve the intervention, or to respond to fewer or more resources in the hospital.*

<u>Social:</u> the intervention aims to change how people relate to each other, think and behave in groups and larger collectives. The implementers interact with each other and others outside the implementation group: the implementation is a social activity. They interact with those who are being enabled to change in the “target” audience or organization.

> *Those implementing the MET interact with each other as a team as they plan and progress the implementation, and interact with many others to establish the MET.*
> *The aim of the intervention is to change social relations and behaviors. The training and other actions aim to allow nurses and doctors to relate to a MET directly when they decide, and to bypass the usual chain of command. Members of the MET team often train personnel during the MET call to stop the patient’s deterioration rather than providing this care directly.*

<u>An environment interacting with the intervention:</u> for the implementation actions to be carried out there need to be resources and other conditions. These conditions change during implementation and also include a history which made the implementation possible.

> *The chief medical officer who did not support the idea left and was replaced by one who did. More financing was made available for training. An authoritative organization recommended all hospitals should have a MET team.*

<u>Results</u>: any change resulting from the intervention.

> *Immediate changes may be that personnel start calling the MET when local seniors are not available, or that the workload of critical care personnel staffing the MET increases. There are also changes to policies, procedures and systems.*

<u>A causal chain sequence</u>: There are intermediate results in a causal chain sequence which is intended to end with better patient experience or clinical results or less waste. Intermediate results can be changes to procedures, structures or systems which are related to changes in provider behavior. The latter in turn may lead to improved patient results.

A causal chain sequence can be the assumptions of the implementers about how the intervention may lead to improved patient care. It can also be the program theory or model constructed by the researchers before data collection to test hypotheses, or after data collection to explain what was implemented and the results.

<u>Outcomes:</u> before/after differences in a variable which can be attributed to the intervention. Usually patient results cannot be unambiguously attributed to the intervention (e.g., changes in mortality), but process indicators of intermediate results (e.g., procedures established, number of calls to a MET team) may be more easy to relate to the intervention, and then these may be related through the causal chain to other results.

Criteria for assessing the quality of studies for CSI evaluations which use naturalistic observational approaches

There are no published criteria or guidelines assessing research using these approaches in the same way there are criteria for experimental or quality improvement research. In addition, assessment criteria to a large extent depend on the specific questions which the research aims to address. There are, however, a number of discussions about how to increase the validity (Mays & Pope 1995, 2000, Popay et al 2008).

From these discussions, the following general criteria were derived for assessing evaluations of CSIs are aimed at understanding and evaluating CSIs in routine settings.

Protocol?

Is there a plan which describes the questions to be answered, the theoretical propositions to be explored and the data collection and analysis methods?

Data collection?

- Did the study use multiple sources of evidence and triangulation which describe each of the "facts" or events by more than a single source of evidence?
- Did the research create a case study data-base which could allow other investigators to review the evidence directly, which includes interview notes?
- Does the study cite specific documents, interviews and/or observations in support of each conclusion and provide a "chain of evidence"?
- Does the study follow accepted good practice in procedures for gathering data in the methods used, such as interviews, document abstraction, surveys, and observational methods?

Data analysis?

- Does the study take account of the variety of evidence gathered?
- Does the study present the evidence separate from interpretation or explanation?
- Does the study adequately consider alternative interpretations or explanations?

Use of theory?

- Does the study define the theory or model which was used to select or conceptualize the case or intervention?
- Does the study test different theories or hypotheses for the events and results?
- If the study seeks to find repeated patterns in the analysis (pattern matching), did the study predict a sequence of events and to what extend are these observed in the empirical data?
- To what extent does the study build an explanation and how well does it do this? (for example, does it make an initial theoretical statement or proposition about the process or behavior of the case; comparing the findings of the case against the statement; revising the statement; compare other details of the case against the revision; compare the revision to other cases)

- If the study uses time series analysis how well does the analysis describe the events over time, their relationship, and explain them? (does it compare the chronology with that predicted by a theory)
- If the study uses a logic model, how well does the study describe, before data gathering, the activities and the immediate outcomes, and relate these to intermediate outcomes and to the final outcomes, and then test this against the data and revise the model in the light of the data and/or other explanations?

8. References

- Adams R, Bessant J, Phelps R. 2006 Innovation management measure- ment: a review. I J Manag Rev 2006;8:21–47
- Alexander JA, Weiner BJ, Shortell SM, Baker LC, Becker MP. 2006 The role of organizational infrastructure in implementation of hospitals' quality improvement. Hosp Top. 2006 Winter;84(1):11-20.
- Altman, D 1991 Practical Statistics for Medical Research. London: Chapman & Hall, 1991.
- Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. 1996 Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-9.
- Benning A, Ghaleb M, Suokas A, Dixon-Woods M, Dawson J, Barber N, et al. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. BMJ 2011;doi:10.1136/bmj.d195.
- Berwick D 2008 The Science of Improvement. JAMA 2008;299:1182 – 4
- Bickman L. 2000 Summing up program theory. In: Rogers PJ, Hacsi TA, Petrosino A, Huebner TA, eds. Program Theory in Evaluation Challenges and Opportunities: New Directions for Evaluation, No. 87. San Francisco, CA: Jossey-Bass; 2000;103-112.
- Bickman, K 2008 The science of quality improvement (letter to the editor) JAMA. 2008;300(4):391
- Bickman, L ed 1987 Using program theory in program evaluation, v 33 of New Directions in Program Evaluation, San Francisco: Jossey-Bass, 1987.
- Blaise P Kegels G 2004 A realistic approach to the evaluation of the quality management movement in health care systems: a comparison between European and African contexts based on Mintzberg’s organizational models Int J Health Plann Mgmt 2004; 19: 337–364.
- Booth A, Falzon F. 2001 Evaluating information service innovations in the health service: ‘If I was planning on going there I wouldn’t start from here’. Health Inform Jl 2001;7:13–9
- Bradley F, Wiles R, Kinmonth A, Mant D, Gantley M. 1999 Development and evaluation of complex inter ventions in health services research: case study of the Southampton heart integrated care project (SHIP). BMJ 1999;318:711-5.
- Brown C, T Hofer, A Johal, R Thomson, J Nicholl, B D Franklin, and R J Lilford 2008a An epistemology of patient safety research: a framework for study design and interpretation. Part 1. Conceptualising and developing interventions Qual Saf Health Care 2008;17 158-162 (see also other 3 papers in the series).
- Brown, C. A. and R. J. Lilford 2006 The stepped wedge trial design: a systematic review BMC Med Res Methodol 6.
- Byng R Norman I Redfern S 2005 Using Realistic Evaluation to Evaluate a Practice-level Intervention to Improve Primary Healthcare for Patients with Long-term Mental Illness Evaluation Vol 11(1): 69–93
- Byng R Norman I Redfern S Jones R 2008 Exposing the key functions of a complex intervention for shared care in mental health: case study of a process evaluation BMC Health Services Research 2008, 8:274
- Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PAC, Rubin HR: Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999, 282:1458-1465. AND
- Campbell DT, Stanley JC, Gage NL. 1966 Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally College Pub Co; 1966.
- Campbell M, Fitzpatrick R, Haines A, Kinmouth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000;321:694-6.
- Carey, R & Lloyd, R 1995 Measuring Quality Improvement in Healthcare, Quality Resources, New York.
- Coleman, E Parry, C Chalmers, S Min, S Chalmers 2006 The Care Transitions Intervention Results of a Randomized Controlled Trial Arch Intern Med. 2006;166:1822-1828
- Concato J, Horwitz RI. 2004 Beyond randomised versus observational studies. Lancet. 2004;363(9422):1660-1661.
- Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. 2010 Observational methods in

comparative effectiveness research. Am J Med. 2010;123 (12)(suppl 1):e16-e23.

- Concato, J 2012 Is It Time for Medicine-Based Evidence? JAMA, April 18, 2012—Vol 307, No. 15 1641-3.
- Connell JP, Klem AC. 2000 You can get there from here: using a theory of change approach to plan urban education reform. J Educ Psychol Consult 2000;11:93e120.
- Coyte PC, Holmes D. 2007 Health care technology adoption and diffusion in a social context. Policy Polit Nurs Pract 2007;8:47–54
- Craig, P., P. Dieppe, et al. (2008). "Developing and evaluating complex interventions: the new Medical Research Council guidance." BMJ 337: a1655. ID Number 1047.http://www.ncbi.nlm.nih.gov/entrez (see also Craig, P et al (2008) Developing and evaluating complex interventions: new guidance, MRC, London: Medical Research Council www.mrc.ac.uk/)
- Damschroder L Aron D, Keith R Kirsh S, Alexander J Lowery J 2009 Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science, Implementation Science 2009, 4:50, 1-15.
- Davis, P 2005 The Limits of Realist Evaluation Surfacing and Exploring Assumptions in Assessingthe Best Value Performance Regime Evaluation Vol 11(3): 275–295.
- Devon E. Hinton D Chhean D Pich V et al 2005 A randomized controlled trial of cognitive-behavior therapy for Cambodian refugees with treatment-resistant PTSD and panic attacks: A cross-over design Journal of Traumatic Stress Volume 18, Issue 6, pages 617–629, December 2005.
- Dixon-Woods M, C Tarrant, J Willars, A Suokas How will it work? A qualitative study of strategic stakeholders' accounts of a patient safety initiative, Qual Saf Health Care 2010;19:74e78.
- Dixon-Woods, M Bosk, C Aveling, E Goeschel, C Pronovost, P 2011 Explaining Michigan: Developing an Ex Post Theory of a Quality Improvement Program The Milbank Quarterly, Vol. 89, No. 2, 2011 (pp. 167–205).
- Dreyer NA, Tunis SR, Berger M, et al. 2010 Why observational studies should be among the tools used in comparative effectiveness research. Health Aff (Millwood) 2010;29:1818e25.
- EPOC 2009 Cochrane Effective Practice and Organisation of Care Group http://www.epoc.cochrane.org/en/authors.html
- Fan E Laupacis A Pronovost P et al. 2010 How to Use an Article About Quality Improvement, JAMA. 2010;304(20):2279-2287.
- Feinstein AR. Multivariable Analysis: An Introduction. New Haven, CT: Yale University Press; 1996.

  9. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis. 1967;20(7):511-524.
- Ferlie, E Fitzgerald, L Wood M Hawkins, C The nonspread of innovations: the mediating role of professionals Academy of Management Journal 2005, Vol. 48, No. 1, 117–134.
- Foy R, Walker A, Ramsay C, Penney G, Grimshaw J, Francis JJ: Theory-based identification of barriers to quality improvement: induced abortion care. Int J Qual Health Care 2005, 17:147-155.
- Foy, R Øvretveit, J Shekelle, P Pronovost,J Taylor, S Dy, S Hempel, S McDonald, K Rubenstein, L Wachter, R 2011 The role of theory in research to develop and evaluate the implementation of patient safety practices, Quality and Safety in Health Care 2011 BMJ Qual Saf. 2011; 20(5): p. 453-459.
- Francis JJ, Eccles MP, Johnston M, Walker AE, Grimshaw JM, Foy R, Kaner EFS, Smith L, Bonetti D: Constructing questionnaires based on the theory of planned behavior. A manual for health services researchers. Centre for Health Services Research, University of Newcastle upon Tyne; 2004.
- Freeman T, Dickinson H, McIver S, McLeod H. 2006 Innovation in Service Delivery: A Literature Review on the Characteristics of Innovation and Improvement. Health Services Management Centre, University of Birmingham, 2006.
- Friedman CP, Wyatt JC. 2006 Evaluation Methods in Bio-Medical Informatics. 2nd edn. New York: Springer-Verlag, 2006.
- Friedman LM, Furberg CD, DeMets DL 1998 Fundamentals of clinical trials. 3rd edn. New York: Springer 1998.
- Glasgow RE, Lichtenstein E, Marcus AC 2003 Why don't we see more translation of health promotion research to practice? Rethinking the efficacy to effectiveness transition. Am J Public Health. 2003;93:1261–1267.

- Glasziou, P Meats E Heneghan C Shepperd S 2008 What is missing from descriptions of treatment in trials and reviews? BMJ 2008 Vol 336 1472-4.
- Gold, M Helms, D Guterman, S 2011, Identifying, Monitoring, and Assessing Promising Innovations: Using Evaluation to Support Rapid-Cycle Change, Issue Brief, The Commonwealth Fund pub. 1512 Vol. 12, Washington.
- Greene, J "Qualitative program evaluation", pp 530-544, in Denzin, N & Lincoln, Y (eds) Handbook of Qualitative Research, Sage, London, 1993.
- Greenhalgh T & Russell J 2010, Why Do Evaluations of eHealth Programs Fail? An Alternative Set of Guiding Principles, PLOS Medicine 1 November 2010, Volume 7, Issue 11, e1000360
- Greenhalgh T, Humphrey C, Hughes J, Macfarlane F, Butler C, Pawson R. 2009 How do you modernize a health service? A realist evaluation of wholescale transformation in London, UK, Milbank Quarterly 2009; 87: 391-416.
- Greenhalgh T, Robert G, MacFarlane F, Bate P (2004). Diffusion of innovation in service organisations: systematic review and recommendations. Milbank Quarterly 82(4):581-629.
- Greenhalgh, T Humphrey, C Hughes, J Macfarlane, F Butler, C Pawson, R 2009 How Do You Modernize a Health Service? A Realist Evaluation of Whole-Scale Transformation in London, The Milbank Quarterly, Vol. 87, No. 2, 2009 (pp. 391–416).
- Greenhalgh, T Stramer, K Bratan, T Byrne, E Russell, J Potts, H 2010 Adoption and non-adoption of a shared electronic summary record in England: a mixed-method case study BMJ 2010;340:c3111
- Grol, RPTM, Bosch, MC, Hulscher, MEJL, Eccles, MP, Wensing, M. Planning and studying improvement in patient care: The use of theoretical perspectives. The Milbank Quarterly 2007, 85(1):93-138.
- Hart, E & Bond, M (1996) Action Research for Health and Social Care, Open University Press, Milton Keynes.
- Hawe P, Shiell A, Riley T. 2004 Complex interventions: how "out of control" can a randomised trial be? BMJ 2004;328:1561-3.
- HealthyPeople.gov. 2011 Healthy People 2020, Determinants of Health. Washington DC: US Department of Health and Human Services, 2011. Available at: http://www.healthypeople.gov/2020/about/DOHAbout.aspx Accessed June 29, 2011.
- Helfrich CD, Li YF, Sharp ND, Sales AE: Organizational readiness to change assessment (ORCA): Development of an instrument based on the Promoting Action on Research in Health Services (PARiHS) framework. Implementation Science 2009, 4(1):38.
- IHI 100k summarised in Berwick, D. M., Calkins, D. R., McCannon, C. J., & Hackbarth, A. D. (2006). The 100,000 lives campaign: setting a goal and a deadline for improving health care quality. JAMA, 295(3), 324-327.
- JCSEE (1994) The Programme Evaluation Standards: How to assess evaluations of educational programs, (2nd edtn) Sage, Thousand Oaks. (The US Joint Committe on Standards for Educational Evaluation).
- Judge, K 2000 Testing the limits of evaluation: Health Action Zones in England, Journal of Health Services Research and Policy, 5:1.
- Keen, J & Packwood, T Qualitative Research: Case study evaluation BMJ 1995;311:444-446 (12 August)
- Kessler, R Glasgow, R 2011 A Proposal to Speed Translation of Healthcare Research Into Practice Dramatic Change Is Needed Am J Prev Med 2011;40(6):637– 644.
- Kilo C. A framework for collaborative improvement: Lessons from the Institute for Healthcare Improvement's Breakthrough Series. Qual Manag in Health Care. 1998;6(4):1-13.
- Kirkland KB, Homa KA, Lasky RA, et al. Impact of a hospital-wide hand hygiene initiative on healthcare-associated infections: results of an interrupted time series. BMJ Qual Saf 2012. Published Online First: 24 July 2012. doi: 10.1136/ bmjqs-2012-000800
- Klein HK, Myers MD (1999) A set of principles for conducting and evaluating interpretive field studies in information systems. Mis Quarterly 23: 67–93.
- Landefeld CS, Shojania KG, Auerbach AD. 2008 Should we use large scale healthcare interventions without clear evidence that benefits outweigh costs and harms? no. BMJ. 2008;336(7656):1277.

- Langley GJ, Nolan KM, Nolan TW, Norman CN, Provost LP. 1996 The Improvement Guide: A Practical Approach to Enhanc- ing Organizational Performance. San Francisco: Jossey-Bass; 1996.
- Lehmann, H & Ohno-Machado, L 2011 Evaluation of informatics systems: beyond RCTs and beyond the hospital J Am Med Inform Assoc March 2011 Vol 18 No 2, 110-111
- Lipsey, 1993. M.W. Lipsey, Theory as method: Small theories of treatments. In: L.B. Sechrest and A.G. Scott, Editors, Understanding causes and generalizing about them. New Directions for Program Evaluation 57, Jossey-Bass, San Francisco, CA (1993), pp. 5–38.
- Liu JL, Wyatt JC. The case for randomized controlled trials to assess the impact of clinical information systems. J Am Med Inform Assoc 2011;18:173e80.
- Lukas CV, Holmes SK, Cohen AB, Restuccia J, Cramer IE, Shwartz M, Charns MP. 2007 Transformational change in health care systems: an organizational model. Health Care Manage Rev. 2007 Oct-Dec;32(4):309-20
- Mann, C 2003 Observational research methods. Research design II: cohort, cross sectional, and case-control studies, Emerg Med J 2003;20:54–60
- Mason P, Barnes M. Constructing theories of change: methods and sources. Evaluation 2007;13:151e70.
- May, C., Murray, E., Finch, T., Mair, F., Treweek, S., Ballini, L., Macfarlane, A. and Rapley, T. (2010) Normalization Process Theory On-line Users' Manual and Toolkit. Available from http://www.normalizationprocess.org accessed 27dec2012.
- Mays, N & Pope, C (1995) Rigour and qualitative research", British Medical Journal, Vol. 311, pp 109-113.
- Mays, N and Pope 2000 Qualitative research in health care: Assessingquality in qualitative research BMJ 2000;320;50-52
- McCannon C.J., et al. 2006: Saving 100,000 lives in U.S. hospitals. BMJ 332:1328–1330, Jun. 3, 2006. See also McCannon C.J., Hackbarth A.D., Griffin F.A. 2007 Miles to go: An introduction to the 5 Million Lives Campaign. Jt Comm J Qual Patient Saf 33:477–484, Aug. 2007.
- McCormack B, McCarthy G, Wright J, Coffey A, Slater P: Development of the Context Assessment Scale. Belfast: University of Ulster 2008.
- Mdege, N Man, MTorgerson, D 2011 Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation Journal of Clinical Epidemiology Volume 64, Issue 9 , Pages 936-948, September 2011
- Mercer, S. L., B. J. DeVinney, et al. (2007). "Study designs for effectiveness and translation research :identifying trade-offs." Am J Prev Med 33(2): 139-154.
- Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A: Making psychological theory useful for implementing evidence based practice: a consensus approach. Qual Saf Health Care 2005, 14:26-33.
- Michie, S Fixsen, D Grimshaw, J Eccles, M 2009 Specifying and reporting complex behaviour: the Evidence Based Out Reach (EBOR) Trial, Journal of Health Services Research & Policy Vol 7 No 4, 2002: 230–238
- Nazareth, I Freemantle N Duggan C Mason J Haines A 2002 Evaluation of a complex intervention for changing professional behaviour: the Evidence BasedOut Reach (EBOR) Trial Journal of Health Services Research & Policy Vol 7 No 4, 2002: 230–238.
- Oakley,A, V Strange, C Bonell, E Allen, J Stephenson, RIPPLE Study Team, 2006, Process evaluation in randomised controlled trials of complex interventions, British Medical Journal, v.332, p. 413-416.
- Øvretveit J and Gustafson D. (2003) "Evaluation of Quality Improvement Programmes" British Medical Journal,vol 326, pp 759-761.
- Øvretveit, J (1998) Evaluating Health Interventions, Open University Press, Milton Keynes.
- Øvretveit, J (2002) Action Evaluation of Health Programmes and Change A handbook for a user focused approach , Radcliffe Medical Press, Oxford.
- Øvretveit, J 2011a Understanding the conditions for improvement: research to discover which context influences affect improvement success BMJ Qual Saf. 2011; 20(Suppl_1): p. i18-i23. http://qualitysafety.bmj.com/cgi/content/abstract/20/Suppl_1/i18?ct
- Øvretveit, J 2011b Widespread focused improvement: lessons from developing

countries for scaling up specific improvements to health services International Journal for Quality in Health Care 2011; Volume 23, Number 3: pp. 239–24610.1093/intqhc/mzr018

- Øvretveit, J 2012 Evidence Review: Do changes to patient-provider relationships improve quality and save money? Volume 2 Full review of research. The Health Foundation, London. http://www.health.org.uk/publications/do-changes-to-patient-provider-relationships-improve-quality-and-save-money/
- Øvretveit, J Andreen-Sachs M, Carlsson, J Gustafsson, H, Lofgren, S Mazzocato, P Keller, C Hansson, J Brommels, M 2012 Implementing Organization and Management Innovations in Swedish Healthcare: lessons from a comparison of 12 cases, Journal of Health Organisation and Management, 2012, Vol. 26 Iss: 2 pp. 237 – 257.
- Øvretveit, J Aslaksen, A (1999) The Quality Journeys of Six Norwegian Hospitals, Norwegian Medical Association, Oslo, Norway.
- Owen, J. & Rodgers, P. (1999) Program Evaluation: Forms and Approaches (London, Sage)
- Pawson, R., and N. Tilley. 1997. Realistic Evaluation. London: Sage.
- Popay, J Rogers, A Williams, G (1998) “Rationale and standards for the systematic review of qualitative literature in health services research”, Qualitative Health Research, Vol 8 No. 3 pp 341-51.
- PPACA 2010Patient Protection and Affordable Care Act of 2010, PL 111-148, sec. 6301.
- Pronovost PJ, Berenholtz SM, Goeschel C, et al. Improving patient safety in intensive care units in Michigan. J Crit Care 2008;23:207e21 (see also AHRQ 2011 clinical unit safety program (On the Cusp)
- Redfern S Christian S Norman I 2002 Evaluating change in health care practice: lessons from three studies Journal of Evaluation in Clinical Practice,9, 2, 239–249
- Robson, C (1993) Real World Research, Blackwell, Oxford.
- Rousseau N, McColl E, Newton J, Grimshaw J, Eccles M: Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. BMJ 2003, 326:314-318.
- Rubenstein, L. Chaney, E Ober, S Felker, B Sherman, S Lanto, A Vivell, S 2010 Using Evidence-Based Quality Improvement Methods for Translating Depression Collaborative Care Research Into Practice, Families, Systems, & Health.2010, Vol. 28, No. 2, 91–113 DOI: 10.1037/a0020302
- Rubenstein, L. V., L. S. Meredith, et al. (2006). “Impacts of evidence-based quality improvement on depression in primary care: a randomized experiment.” J Gen Intern Med. 21(10): 1027-35. Epub 2006 Jul 7.
- Scales DC, Dainty K, Hales B, et al. A multifaceted intervention for quality improvement in a network of intensive care units: a cluster randomized trial *JAMA*. doi:10.1001/jama.2010 .2000 [published online January 19, 2011]
- Schouten L, Hulscher, M van Everdingen, J Huijsman R Grol, R 2008 Evidence for the impact of quality improvement collaboratives: systematic review BMJ 2008;336;1491-1494;
- Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trial BMJ 2010;340: 698-702 doi: 10.1136/bmj.c332
- Shadish, W Cook, T Leviton, L (1991) Foundations of Programme Evaluation: Theories of practice, Sage, London.
- Shekelle PG, Pronovost PJ, Wachter RM, Taylor SL, Dy S, Foy R, Hempel S, Ovretveit J, Rubenstein L, and the PSP Technical Expert Panel. Assessing the Evidence for Context-Sensitive Effectiveness and Safety of Patient Safety Practices: Developing Criteria (Prepared under Contract No. HHSA-290-2009-10001C.) Rockville, MD: Agency for Healthcare Research and Quality. Jan 2010.
- Speroff T, O’Connor GT. Study designs for PDSA quality improvement research. Qual Manag Health Care. 2004;13(1):17-32.
- Stetler C Damschroder L Helfrich C Hagedorn H (2011) A Guide for applying a revised version of the PARIHS framework for implementation, Implementation Science 2011, 6:99; http://www.implementationscience.com/content/6/1/99
- Stetler CB, Legro MW, Wallace CM, Bowman C, Guihan M, Hagedorn H, Kimmel B, Sharp ND, Smith JL: 2006 The role of formative evaluation in implementation research and the QUERI experience. J Gen Intern Med 2006, 21(Suppl 2):S1-8.

- Sullivan H, Stewart M. Who owns the theory of change? Evaluation 2006;12:179e99.
- Tunis SR, Stryer DB, Clancy CM (2003) Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. JAMA 290: 1624– 1632.
- UKIHI 2001 UK Institute of Health Informatics. Evaluation of Electronic Patient and Health Record Projects. Winchester: ERDIP Programme, NHS Information Authority, 2001.
- Wagenaar, A Komro, K 2011 Natural Experiments: Design Elements for Optimal Causal Inference, PHLR Methods Monograph Series, from Public Health Law Research Program or Department of Health Outcomes and Policy, College of Medicine, University of Florida.
- Wagner EH, Austin BT, Davis C, Hindmarsh M, Schaefer J, Bonomi A. 2001 Improving chronic illness care: translating evidence into action. Health Affairs 2001; 20(6): 64-78.
- Walshe, K Shortell, S What happens when things go wrong: how healthcare organisations deal with major failures in care Health Affairs May 2004, 23, 3: 103-11.
- Ward, V., House, A. Hamer, S. Developing a framework for transferring knowledge into action: a thematic analysis of the literature. Journal of Health Services Research & Policy 2009; 14(3): 156-164.)
- Waterman, H Tillen, D Dickson, R de Koning, K Action research: a systematic review and guidance for assessment, Health Technology Assessment 5, 23, 2001.
- Weiss, C 1997a How can theory based evaluation make greater headway" Evaluation review, 21: 501-8.
- Weiss, C 1997b Theory-based evaluation: past present and future, New directions for evaluation, 76 41-55.
- Wheeler, D (1993) Understanding Variation, SPC Press, Tennesse.
- WHO (1981) "Health Programme Evaluation", World Health Organisation, Geneva.
- Wholey, J (1983) Evaluation and Effective Public Management, Little, Brown, Boston.
- Williams I 2011 Organizational readiness for innovation in health care: some lessons from the recent literature, Health Services Management Research 2011; 00: 1–6.
- Wolff N: Randomised trials of socially complex interventions: promise or peril? Journal of Health Services Research and Policy 2001, 6(2):123-126.
- Yin, R 2003 Case Study Research: Design and Methods, Third Edition. Thousand Oaks, CA: Sage Publications, 2003.
- Zapka, J Valentine-Goins, K Pbert L Ockene K 2004 Translating Efficacy Research to Effectiveness Studies in Practice: Lessons From Research to Promote Smoking Cessation in Community Health Centers Health Promot Pract 2004; 5; 245-255