Contract Final Report

Assessing the Evidence for Context-Sensitive Effectiveness and Safety of Patient Safety Practices: Developing Criteria

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. HHSA-290-2009-10001C

Prepared by:
RAND Health
Santa Monica, CA

‘

Project Leaders
Paul G. Shekelle, M.D., Ph.D.
Peter J. Pronovost, M.D., Ph.D.
Robert M. Wachter, M.D.

AHRQ Publication No. 11-0006-EF
December 2010

The project leaders gratefully acknowledge the contributions of the following individuals:

Content Experts:
Lauren Carr, M.P.H.
Robbie Foy, MBChB, Ph.D., M.Sc
Sydney Dy, M.D
Susanne Hempel, Ph.D
Kathryn M. McDonald, M.M.
John Øvretveit Ph.D., C.Psychol.
Lisa Rubenstein, M.D., MSPH, FACP

Project Director
Stephanie Taylor, Ph.D.

Librarian:
Roberta Shanman, M.L.S.

Research Assistant:
Breanne Johnsen, B.S.

Administrative Experts:
Jerome Hawkins, B.A.
Aneesa Motala, B.A.
Patricia Smith

This report was prepared for the Agency for Healthcare Research and Quality by RAND Health, Santa Monica, CA, under contract HHSA-290-2009-10001C. The opinions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Department of Health and Human Services or the Agency for Healthcare Research and Quality.

Suggested Citation:

Shekelle PG, Pronovost PJ, Wachter RM, Taylor SL, Dy S, Foy R, Hempel S, McDonald, K, 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). AHRQ Publication No. 11-0006-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2010.

Contents

Appendixes

Note: Appendixes are available on request from Patty Smith, RAND Health, 1776 Main Street, Mailstop M4E, Santa Monica, CA 9040; phone 310-393-0411; e-mail psmith@rand.org.

Executive Summary

Introduction

The emergence of new kinds of interventions to improve health care quality and safety has led to a rethinking of traditional health services and clinical research. Interventions intended to improve quality and safety are often complex sociotechnical interventions whose targets may be entire health care organizations or groups of providers, and they may be targeted at extremely rare events. As such, evaluation of patient safety practices (PSPs) must be evaluated along two dimensions: the evidence regarding the outcomes of the safe practices and the contextual factors influencing the practices' use and effectiveness.

The methodological criteria for assessing the quality of clinical intervention research and evaluation studies may be insufficient for studies of the effectiveness of organizational and behavioral change required to implement a safety practice. Indeed, researchers of PSPs often have to assess, as clinical researchers do, whether an intervention works. They also, as organizational and behavioral researchers do, need to determine whether such practices will work in their own settings, (i.e., will they benefit patients in their own organization with its unique attributes). In addition to questions of effectiveness (whether, how, and why interventions work), it is also important to consider unintended adverse consequences of implementing the safety practice. In other words, like medications, quality improvement (QI) and safety interventions can have side effects, which must be anticipated and measured.

Origin of this Report

Over the past decade, major concerns about the quality and safety of medical care have surfaced. Influential factors in our health care system such as government payers, accreditors, and employers have responded by creating a variety of incentives to promote quality and safety. The lack of consensus about the standards creates a risk that the substantial investment in new knowledge will be undermined by poor study design, flawed execution, or inappropriate interpretation of study results. In addition, policymakers are encouraging or requiring provider organizations to implement safe practices in the absence of explicit criteria for evaluating the strength of the evidence supporting the practice under consideration or evidence about the likelihood that patients will benefit.

Recognizing this major gap in knowledge and understanding, AHRQ supported the development of a report to identify criteria for assessing the context-sensitive effectiveness and safety of PSPs. Context is a particularly crucial issue because it is believed to be a key factor differentiating the interpretation of PSPs from clinical interventions. Researchers, policymakers, and providers evaluating PSPs care not only whether robust evidence supports the PSP, but also whether and how they can implement the PSP in their organizations to improve patient outcomes.

To address these gaps, the Agency for Healthcare Research and Quality issued a Request for Proposals (RFP) focused on developing criteria to assess the effectiveness and safety of PSPs. In the RFP guiding this project, PSPs are described as "interventions; systems, organizational, and behavioral interventions; and various combinations of these." To provide a real-world basis for

committee deliberations regarding the research questions, the study investigators, working with a panel of experts, chose to focus on five PSPs representing various aspects of the patient safety research field:

1. Checklists for catheter-related bloodstream infection prevention.
2. The Universal Protocol for preventing wrong procedure, wrong site, wrong person surgery.
3. Computerized order entry/decision support systems.
4. Medication reconciliation.
5. Interventions to prevent in-facility falls.

Methods

In this 1-year project, we assembled a 22-member Technical Expert Panel (TEP) comprising international patient safety leaders, clinicians, policymakers, social scientists, and methodologists. We met with the TEP three times, performed many literature reviews, conducted five Internet surveys, and achieved consensus on the points below.

Key Findings

1. Important evaluation questions for these PSPs are:

    a. What is the effectiveness of the PSP?
    b. What is the implementation experience of the PSP at individual institutions?
    c. What is the success of widespread adoption, spread, and sustainability of the PSP?

Interpretation and significance: Evaluations of PSPs should explicitly consider these three questions. Journals should consider asking researchers to report on them separately. Also, implementers will want to assess their experience across all three questions.

2. High-priority contexts for assessing context-sensitive effectiveness at individual institutions are:

    a. Structural organizational characteristics (such as size, location, financial status, existing quality and safety infrastructure).
    b. External factors (such as regulatory requirements, the presence in the external environment of payments or penalties such as pay-for-performance or public reporting, national patient safety campaigns or collaboratives, or local sentinel patient safety events).
    c. Patient safety culture (not to be confused with the larger organizational culture), teamwork, and leadership at the level of the unit.
    d. Availability of implementation and management tools (such as staff education and training, presence of dedicated time for training, use of internal audit-and-feedback, presence of internal or external individuals responsible for the implementation, or degree of local tailoring of any intervention).

Interpretation and significance: Context is considered important in determining the outcomes of PSPs. The study investigators and the TEP judged these four domains as the most salient areas of context. This recommendation has broad implications for a variety of audiences. Researchers should be encouraged to measure and report on these contexts when describing a study of a PSP. Consumers of research will want to look for such reports, which will influence their interpretation of the study results and affect the applicability of the PSP to their setting. Accreditors and regulators should be reluctant to mandate adoption of a given PSP if it appears to be very dependent on context. In that case, they should also provide guidance on how that PSP might need to be modified depending on local contexts.

3. There is insufficient evidence and expert opinion to recommend particular measures for patient safety culture, teamwork, or leadership. Given the plethora of existing measurement tools we identified and reviewed, our recommendation is to use whichever method seems most appropriate for the particular PSP being evaluated.

    a. For patient safety culture, the measurements methods with the most support were the AHRQ Patient Safety Culture Surveys, the Safety Climate Scale, and the related Safety Climate Survey.
    b. For teamwork, the most support was given to the ICU [Intensive Care Unit] Nurse-Physician Questionnaire; no other measure received more than half the votes of respondents.
    c. For leadership, the measures receiving the most support were the ICU Nurse-Physician Questionnaire, the Leadership Practice Inventory, and the Practice Environment Scale.

Interpretation and significance: Because the four areas of context described in Point 2, above, are judged highest priority, it will be crucial to develop and use valid measures of them in PSP studies. Researchers' use of common validated instruments would better enable readers to evaluate whether published results are applicable to their own settings. The state of the science here is immature, and funders and researchers are encouraged to continue to develop standard measures of the key domains of context.

4. The PSP field would advance by moving past considering studies of effectiveness as being "controlled trials" versus "observational studies." Although controlled trials offer greater control of sources of systematic error, they often are not feasible, either in terms of time or resources. Also, controlled trials often are not possible for PSPs requiring large-scale organizational change or PSPs targeted at very rare events. Hence, strong evidence about the effectiveness and comparative effectiveness of PSPs can be developed using designs other than randomized controlled trials. However, PSP evaluators are to be discouraged from drawing cause-and-effect conclusions from studies with a single pre- and post-intervention measure of outcome. More sophisticated designs (such as a time series or stepped wedge design), are available and should be used when possible.

Interpretation and significance: Given the major push to improve patient safety and the focus on evidence-based practices (which are rapidly embedded in national standards such as those issued by the National Quality Forum, the Joint Commission, the Institute for Healthcare Improvement, and others), it will be crucial to develop standards for appropriate evaluations to answer key

safety-oriented questions. The results above will help journal editors, funders, researchers, and implementers adopt robust study methods for PSPs, methods that most efficiently answer the key questions without undue bias.

5. Regardless of the study design chosen, criteria for reporting on the following items in a PSP evaluation are necessary, both for an understanding of how the PSP worked in the study site, and whether it might work in other sites:

    a. An explicit description of the theory for the chosen intervention components, and/or an explicit logic model for "why this PSP should work."
    b. A description of the PSP in sufficient detail that it can be replicated, including the expected change in staff roles.
    c. Measurement of contexts in the four domains described in Point 2, above.
    d. Details of the implementation process, what the actual effects were on staff roles, and how the implementation or the intervention changed over time.
    e. Assessment of the impact of the PSP on outcomes and possible unexpected effects. Including data on costs, when available, is desirable.
    f. For studies with multiple intervention sites, an assessment of the influence of context on intervention and implementation effectiveness (processes and clinical outcomes).

Interpretation and significance: These criteria (items a-f) are deemed necessary for an understanding of PSP implementation and effectiveness and the degree to which these elements are sensitive to context. Future AHRQ-supported evaluations of PSP implementation should adhere to the criteria developed by this project. Only through repeated assessments and measurements will it be possible to determine the context-sensitivity of PSPs, build the evidence base for which contexts are most important, and determine how they should be measured and reported.

Recommendations for Future Research

Based on the group discussions and a formal vote by the TEP, the most important needs for future research are:

1. **Developing and validating measures of patient safety culture.** Discussion at the panel meetings indicated that several technical experts considered patient safety culture to be the overarching important construct. This view may explain why patient safety culture received majority support as a high priority for future research, whereas research on leadership and teamwork measures did not. Specific suggestions for future research included:

    a. Developing validated measures of cultural adaptability to change.
    b. Assessing the potential distinction between a culture of safety, a culture of excellence, and organizational culture.
    c. Establishing connections between aspects of patient safety culture and patient outcomes or processes of care.
    d. Assessing correlations between measures.

Additional comments that we received can be summarized as “we think teamwork and leadership are important," "several measures are currently available," and "the most important thing at this point is for people to use them so we can start building some evidence about this construct.”

2. **Developing criteria and recommendations, for what constitutes "reporting the intervention in sufficient detail that it can be replicated.”** More precise criteria for how PSP interventions should be described warrant additional research. In particular, the guidance described here, along with that provided by Standards for Quality Improvement Reporting Excellence (SQUIRE) and the National Quality Forum (NQF), need to be evaluated. Doing so will help determine which PSP elements need to be described in order to evaluate whether the PSP is truly effective. This also will help maximize the possibility of successful PSP replication with similar outcomes. Further research could also evaluate the effect of applying these draft criteria regarding PSP descriptions on the quality of PSP projects and published articles. Clearly, thoroughly describing PSPs also can help readers determine the relevance of an evaluation study to other PSPs or other contexts. For example, if a PSP requires an individual behavior change such as hand-washing, then knowing intervention details may help readers of the study assess whether the given results are relevant only to hand-washing interventions or if they could be applied to other types of PSPs requiring individual behavior change. Knowing the details of the intervention also could help readers of the study determine how much the success of the PSP implementation depended on contextual issues (e.g., organization or teamwork).

3. **Understanding the important items to measure and report on for implementation.** Experts consider having comprehensive information about implementation key to being able to replicate a PSP. However, little empirical evidence exists about what makes a description of the PSP adequate for reporting. Assessing what implementers need to know, if they are to be able to implement or adapt an intervention in their own settings, is critical. Most experts considered "understanding the important items to measure and report on for implementation" to be related to or even the same as "reporting the intervention in sufficient detail that it can be replicated." This view suggests that the distinction between “the intervention” and “the implementation” may be an arbitrary line, and that ideal evaluations of PSP interventions need to consider the implementation as part of the intervention.

4. **Developing a theory-based taxonomy or framework with which to describe and evaluate key elements of interventions, contexts, and targeted behaviors.** Although the current project made a promising start on meeting this need, progress in this area will require additional development to produce a taxonomy that would be both sufficiently broad based and flexible enough to be widely useful. Issues to be considered include whether a taxonomy is the preferable way to proceed, or whether a more useful strategy might be to create an explicit methodology that researchers could apply to specific problems and contexts. Yet another approach might be to devise an “assessment framework.” Some experts sounded cautionary notes on this topic. They reported that outpatient PSP research may be too new to apply a taxonomy at this stage. They also reported that a single “unified” taxonomy may not be sufficiently flexible for diverse PSPs, and multiple taxonomies may be needed in any case. The countervailing view to these cautionary notes was that the field would not be well-served

by having a proliferation of taxonomies. Instead, they reported, what is needed is a coherent, sufficiently comprehensive taxonomy that can accommodate the challenges of the subject.

5. **Refining a framework for assessing the strength of a body of evidence.** We did developmental work on an adaptation of the GRADE and Evidence-based Practice Center (EPC) systems for assessing the strength of evidence across studies of a PSP. This work warrants further development.

6. **Generating empirical evidence that the contextual factors identified in this project influence the success of the PSP.** We acknowledge that most of the recommendations in the report have a thin empirical evidence base, which simply reflects the relatively immature state of research in this still relatively young field. Building a stronger evidence base will help future efforts at refining the recommendations presented here.

Chapter 1. Introduction

Patient safety research is a fairly young field that received substantial investment in the United States after the Institute of Medicine's 1999 landmark report *To Err is Human* sounded the alarm and resonated with the public. They heard the salient sound bite that one "jumbo jet" of patients dies each day from a medical error. In rapid response, the lead Federal agency for health care quality research, the Agency for Healthcare Quality and Research (AHRQ), commissioned an evidence-based practice center (EPC) team of researchers to develop an evidence report of "patient safety practices" (PSP). The resulting report, *Making Health Care Safer: A Critical Analysis of Patient Safety Practices*, identified 79 practices, ranging from targeted clinical interventions (e.g., the use of antibiotic-impregnated catheters to prevent urinary tract infections), to clinical procedural enhancements (e.g., visualizing central line placement to avoid inadvertent lung puncture), to broad system changes (e.g., promotion of a culture of safety and teamwork to reduce a range of possible failures in patient safety).[1] Because the evidence for the effectiveness of these PSPs was scant, according to the established evidentiary review lens available to the EPC at that time (which was in use by the global consortium of systematic review experts, the Cochrane Collaboration and their Effectiveness of Practice and Organization of Care (EPOC) group),[2] the complex, systems-oriented PSPs did not rise to the top of the list of PSPs recommended for further implementation.

These EPC recommendations, while explicitly based on only one potential approach to distill the evidence on practices, stimulated an important debate about whether the evidentiary lens needed adjustment for application to patient safety practices. The issues on either side of the debate are well presented in two "Controversies" articles published in JAMA in 2002,[3, 4] and taken up again more recently in other publications.[5-10] The interest in determining the approach to evidence evaluation for patient safety was also highlighted at the Second National Summit on Patient Safety Research sponsored by the Quality Inter-Agency Coordination Task Force (QuIC) in November 2003.[11] The panel reinforced that often it is not possible or practical to evaluate implementation performance using randomized controlled trials (RCTs), and as a result, concluded that other types of research designs should be considered. In addition, the panel recommended that AHRQ develop standards for patient safety research, as well as for synthesis of a body of evidence on a given PSP, based on a range of suitable research designs and analytic methods. Over the ensuing years, further efforts by national and international organizations (e.g., the National Quality Forum (NQF), the Joint Commission's International Center for Patient Safety) have focused on approaches to identifying, prioritizing, and recommending further development and dissemination of PSPs or "Patient Safety Solutions."[12]

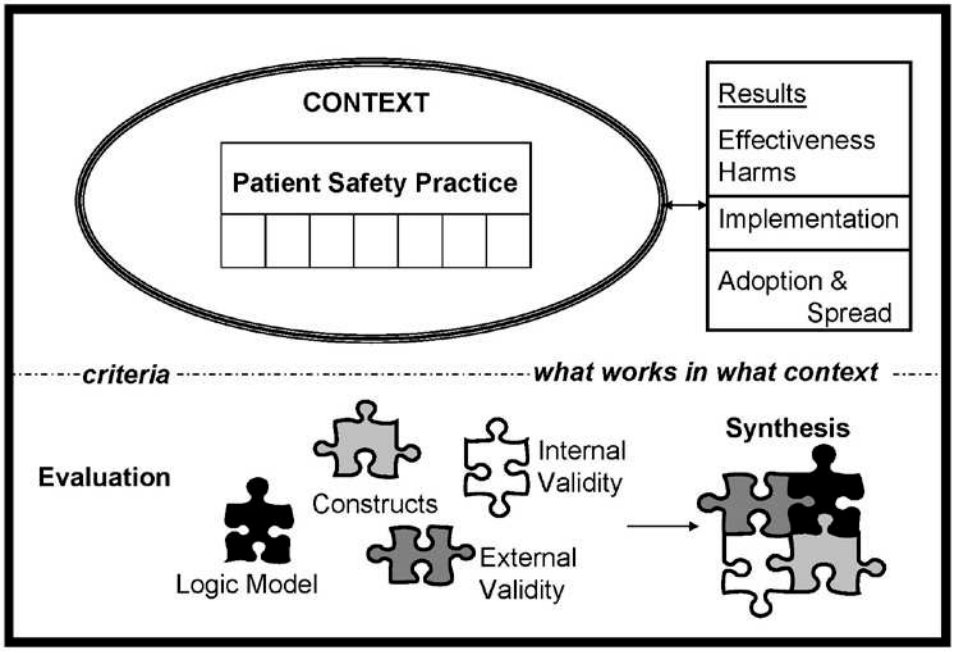
AHRQ initiated the current project to respond to the debate on what constitutes evidence in patient safety by engaging in a structured process with experts from a broad range of pertinent fields, including human factors, organizational behavior, management sciences, public health, evaluation sciences, implementation sciences, biostatistics, clinical medicine, evidence-based medicine, and patient safety. The overarching charge to the research team and expert panel was to assist AHRQ in developing "criteria for assessing the evidence base for the context-sensitive effectiveness and safety of patient safety practices," or how the contexts (within which a patient safety practice is implemented) can affect the effectiveness of that implementation. This charge emanates from a well-articulated rationale described by AHRQ per the Request for Proposal (RFP) that guides this project, and it is summarized by the project team in Figure 1.

The diagram's upper part displays a PSP, its context, and potential results from localized testing or full-scale roll out. The lower half of the diagram stylizes the key components of evaluation and how they need to be fit together in evidence synthesis. The middle line represents the needs of patient safety stakeholders for criteria to assess which patient safety practices work and in what context. Essentially, this project aims to strengthen the line between the top and bottom half of this diagram. Each part of the diagram is described further in the following sections to provide a brief rationale for the project.

Diagram Component: Patient Safety Practices

In the RFP that led to this project, AHRQ defined patient safety practices as "interventions, strategies, or approaches intended to prevent or mitigate unintended consequences of the delivery of health care and to improve the safety of health care for patients. PSPs may include clinical interventions, systems' organizational and behavioral interventions, and various combinations of these."

As implied by this definition, PSPs often include components that often are constructed differently at different points in time or in different settings. Figure 1 shows a generic PSP, with small empty boxes as placeholders to describe the PSP's components. The definition also highlights the diversity of PSPs and the potential for combining them to develop new PSPs.



.

Figure 1. Rationale for examining patient safety practices to assess their effectiveness

Diagram Component: Context

The oval around the PSP in Figure 1 represents the organizational, behavioral, and broader environmental context in which the PSP is embedded. Numerous leaders in patient safety research have articulated the importance of context. In a forthcoming review for the World Health Organization, John Øvretveit and colleagues state that an intervention's effectiveness and safety may vary according to context because of implementation differences, the need for adaptation of implementation, and interactions between contextual factors and the intervention, which result in modification to the intervention over time (Personal communication). Some PSPs address specific evidence-based therapies, while other PSPs are more abstract or diffuse, such as "training clinicians in teamwork." Local factors (such as staffing considerations) may require changes in order to make the PSP implementable given the local or wider context. Thus, interventions that appear to be the same or carry the same label may in fact be quite different when implemented in various places and timeframes; and such differences may account for different outcomes. These considerations support a requirement that studies provide precise descriptions of the evaluated intervention, along with relevant features of the intervention context, including implementation processes.

For many complex interventions, there is a paucity of information about context and its interplay with the PSP. For example, the 2006 AHRQ EPC report on Health Information Technologies (health IT) by Shekelle and colleagues found that the interventions studied included not just the technical aspects of the computer and software, but also the human factors, the project management, and the organizational process change; and that these contextual factors were not adequately described, making it is difficult for others to apply the study results to actual health care settings.[14] In another AHRQ EPC project on care coordination interventions to improve health care quality and patient safety, McDonald and colleagues noted the need for context-flexible evaluations tied to theory, as well as actual needs of quality improvement implementers.[15] The authors called for more detailed descriptions of both the interventions and the contexts in which they were tested to make any conclusions about outcomes more readily interpretable to those choosing potential intervention strategies for their particular circumstances. Thus, EPC investigators have also recognized the importance of context.

There is no standard definition of "context." It may include detailed information about processes of implementation, as well as barriers and facilitators related to the organizational and policy environment in which a PSP is implemented. These factors have been shown to be critically important to understanding the success or failure of a PSP. For example, Pronovost and colleagues discussed the importance of considering local context while maintaining standardized measures and evidence in their effort to reduce blood stream infections in Michigan.[16] They found that it was both efficient and effective to standardize the technical aspects of quality improvement programs while encouraging local modification of how the evidence is put into practice. Similarly, a recent evaluation of the World Health Organization (WHO) surgical checklist found an overall reduction in adverse events; yet this was not consistent at all sites (see http://www.who.int/patientsafety/safesurgery/ss_checklist/en/index.html). Evaluation of the implementation effectiveness or barriers and facilitators will be important in attempts to disseminate the WHO surgical checklist across the world.[17]

Diagram Component: Results

Use of a PSP in a particular context may result in positive and negative outcomes (including unintended harms), shown as effectiveness and harms respectively in the box on the right hand side of Figure 1. In addition to these critical outcomes, other potentially important effects include those related either to implementation (e.g., uptake, cost, and ease of implementation initially) or widespread adoption and spread of a PSP. Figure 1 is a simplification, but nevertheless, it posits that the effectiveness, safety, and other outcomes of a PSP may be affected by its specific components; where, when, and how the PSP is implemented; and with whom and for what purposes the PSP is used; as well as by features of the external environment or larger context.

Diagram Component: Criteria and Knowing What Works

The middle dotted line in Figure 1 sets up the overarching objective of the current project. To inform stakeholders interested in improving patient safety about what works in which contexts, AHRQ has called for context-sensitive criteria to assess PSPs. Therefore, the goal of developing criteria and guidance on evaluations of PSPs is to understand the relationships between PSPs and their intended and unintended results in particular contexts and configurations. Specifically, the agency suggests that:

> Establishing more appropriate criteria for evidence reviews of patient safety practices can be expected to have three closely related effects. First, the criteria should broaden the scope of patient safety practices that can be assessed for effectiveness and safety based on scientific evidence. Second, the availability of the criteria will strengthen research studies that are assessing those practices. Third, if developed in a way that is usable to implementers of patient safety practices beyond researchers (e.g., individual clinicians, health policymakers, and patient advocates), criteria can be applied in situations where PSPs should be evaluated for individual and institutional learning without regard to publication in peer-reviewed journals (per the RFP for this project; see www.ahrq.gov/fund/contarchive/rfp0910001.htm).

Diagram Component: Evaluation and Synthesis

For context-sensitive evaluation of PSPs, evidence synthesis promises to assemble information from individual studies, ultimately determining how to draw together information for each of the four puzzle pieces (Figure 1):

Constructs about the PSP, its components, context factors, outcomes, and ways to measure accurately these constructs.
Logic model or conceptual framework about the expected relationships among these constructs.
Internal validity to assess the PSP results in a particular setting.
External validity to assess the likelihood of being able to garner the same results in another setting.

A number of individual studies, with a broad range of research and evaluation designs, may be needed to answer satisfactorily the many questions of interest to the patient safety field for a given PSP. Initial key questions for evaluation and synthesis put forth by AHRQ include those focused on effectiveness, implementation, and adoption or spread.

In summary, this project aims to advance the patient safety field by using targeted literature reviews and structured expert panel consultation to present a conceptual framework and a set of rigorous evaluation criteria for assessing and guiding studies on PSPs and their contexts. The report presents an initial conceptual framework, initial criteria, and a path toward developing a comprehensive set of rigorous evaluation criteria for assessing and guiding studies on PSPs and their contexts. Based on the framework and criteria, we identify the types of research and evaluation models and methods that experts judge to be most useful for advancing the field of patient safety. We develop specific criteria for assessing the rigor of individual studies; we also lay out methods and criteria for synthesizing sets of studies to assess the overall body of evidence related to specific PSPs and their contexts. Finally, we identify issues and questions for future analysis of and research on PSP methodology.

The litmus test for the project will ultimately come from those on the frontlines of patient safety improvement efforts. What information will help those who are accountable for their health system or the Nation's performance in terms of health care quality and patient safety? What methodology guidance will enable those who are conducting systematic evidence reviews to address key questions about PSPs? What material in the report will ease the process of primary knowledge generation for researchers and evaluators of PSP interventions? What take-home messages will support research funders' ability to continue to move the field forward to its ultimate goal of making health care substantially safer for the public? These questions shape the reporting of our approach and recommendations in the subsequent chapters.

References for Chapter 1

1. Shojania K, Duncan B, McDonald K, Wachter R, eds. Making health care safer: A critical analysis of patient safety practices. Evidence Report/Technology Assessment No. 43 (AHRQ Publication No. 01-E058). Rockville, MD: Agency for Healthcare Research and Quality; 2001.
2. Bero LA, Grilli R, Grimshaw JM, et al. Closing the gap between research and practice: An overview of systematic reviews of interventions to promote the implementation of research findings: The Cochrane Effective Practice and Organization of Care Review Group. *Br Med J*. 1998; 317:465-8.
3. Leape L, Berwick D, Bates D. What practices will most improve safety? Evidence-based medicine meets patient safety. *JAMA.* 2002; 288(4):501-7.
4. Shojania K, Duncan B, McDonald K, Wachter R. Safe but sound: Patient safety meets evidence-based medicine. *JAMA.* 2002; 288(4):508-13.
5. Auerbach A, Landefeld C, Shojania K. The tension between needing to improve care and knowing how to do it. *N Engl J Med.* 2007; 357(6):608-13.
6. Berwick D. The science of improvement. *JAMA.* 2008; 299(10):1182-4.

7. Landefeld CS, Shojania KG, et al. (2008). Should we use large scale healthcare interventions without clear evidence that benefits outweigh costs and harms? No. *Br Med J*. 2008; 336(7656):1277
8. Crump B. Should we use large scale healthcare interventions without clear evidence that benefits outweigh costs and harms? Yes. *Br Med J*. 2008; 336(7656):1276.
9. Pronovost P, Wachter R. Proposed standards for quality improvement research and publication: One step forward and two steps back. *Qua Safety Health Care.* 2006; 15:152-3.
10. Davidoff F. Heterogeneity is not always noise: Lessons from improvement. *JAMA*. 2009; 302(23):2580-6.
11. Quality Interagency Coordination Task Force (QUIC). Paper presented at Second National Summit on Patient Safety; http://www.quic.gov/summit/index.htm. Accessed Aug. 4, 2008.
12. Safe practices for better healthcare. 2008 Update of *Safe practices for better healthcare*. Washington, DC: National Quality Forum; 2008. Available at: www.qualityforum.org/publications.
13. Patient Safety Solutions--Development Process. Washington, DC: The Joint Commission; 2008. Available at http://www.ccforpatientsafety.org/Patient-Safety-Solutions/.
14. Shekelle PG, Morton SC, Keeler EB. Costs and benefits of health information technology. Evidence Report/Technology Assessment No. 132. (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-02-0003.) AHRQ Publication No.06-E006. Rockville, MD: Agency for Healthcare Research and Quality; 2006.
15. McDonald KM, Sundaram V, Bravata DM, et al. Care coordination. In Shojania KG, McDonald KM, Wachter RM, Owens DK, eds. Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies, Volume 7. Technical Review 9 (Prepared by the Stanford University-UCSF Evidence-based Practice Center under contract 290-02-0017). AHRQ Publication No. 04(07)-0051-7. Rockville, MD: Agency for Healthcare Research and Quality; 2007.
16. Pronovost P, Needham D, Berenholtz S, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. *N Engl J Med*. 2006; 355(26):2725-32.
17. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med*. 2009; 360(5):491-9.

Chapter 2. Methods

This project represents a collaboration between the project team and an interdisciplinary group of patient safety and methods experts, called the Technical Expert Panel, or TEP. Each key step of the project required both preparatory work on the part of the project team and then consideration and discussion by the TEP, with synthesis of the TEP discussion and decisions then made by the project team. The figures on the following pages give an overview of the methods, which are more fully described in Appendix A, Part 1.

The five goals of the project were to:

1. Form an interdisciplinary panel of experts to assist with all phases of the project.
2. Identify a diverse and representative set of patient safety practices to be used as initial subjects. As noted in the RFP, "to help iteratively develop criteria for rigorous and systematic assessment of the context-sensitive effectiveness and safety of PSPs. They should be in actual use, promising in terms of underlying logic models for achieving effectiveness, safety, and generalizability, …address high impact and diverse patient safety problems, and represent the contexts in which patient safety is an important concern…"
3. Identify research and evaluation models, methods and designs to rigorously evaluate the patient safety practices identified and, "in considering research designs and methods," identify or develop approaches that measure contexts and implementation processes in PSP interventions and suggest how collection of contextual and process data needed for assessing the generalizability of the PSP can be combined with designs that are strong on internal and construct validity. Pay close attention to assessing both the positive and negative impacts of PSPs. Pay close attention to identifying appropriated measures of aspects of the PSP."
4. Develop a set of criteria, including criteria for strength of evidence, to be used for assessing future studies and reports. Criteria are necessary to guide both (a) future assessments of evidence and safety relative to the effectiveness, implementation, and adoption of the identified types of PSPs; and (b) systematic reviews of patient safety evidence.
5. Identify specific needs for future development of theories, constructs, and research/evaluation designs and methods to further strengthen evaluations of PSPs and criteria for systematic review.

In Figure 2, the selection of the "diverse and representative patient safety practices" (goal 2 above), the project team used the literature, expert input, and information from other sources (such as the project officer, the RFP, etc.) to develop a list of candidate PSPs. This list was then voted on by the TEP, and the results of the vote were used by the project team to select four PSPs, based on a number of criteria such as setting, regulated use, etc., presented in more detail later. There were remaining questions about the need for a possible fifth PSP, and this too was put to the TEP in an e-mail vote. The results of this process led to the final set of five diverse and representative PSPs, which was affirmed by the TEP.

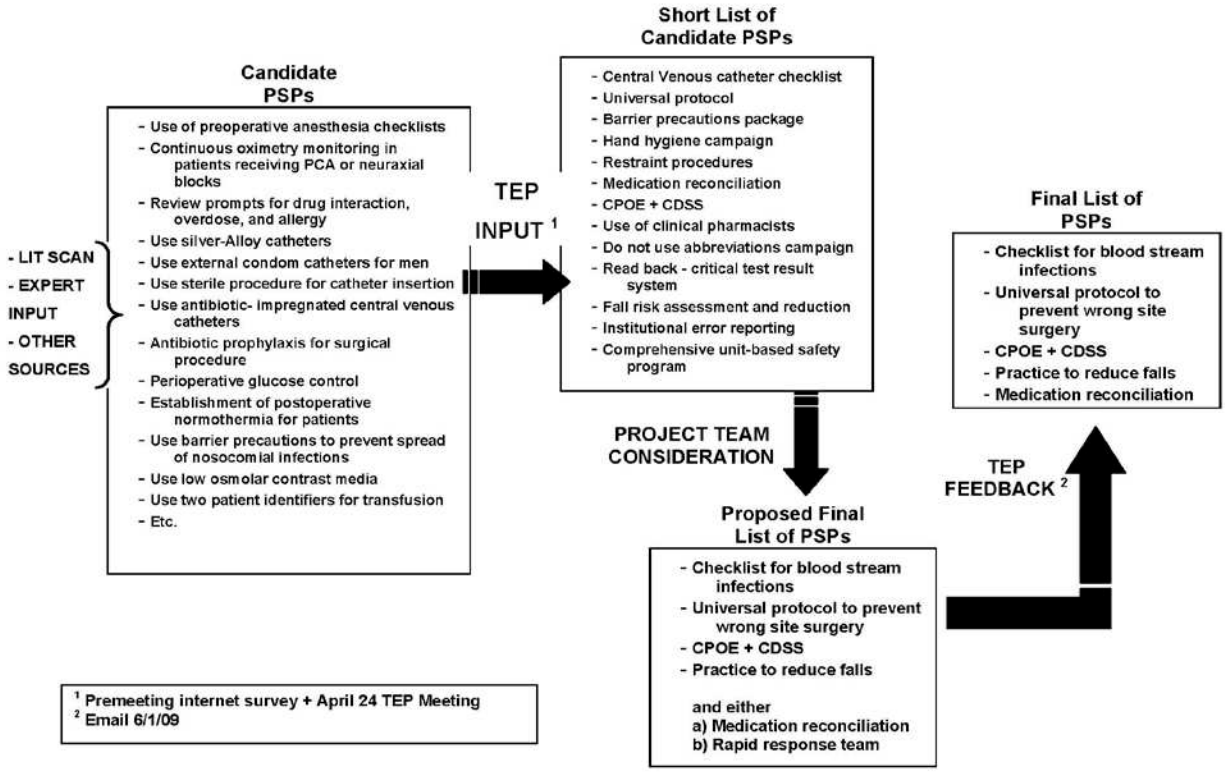


Figure 2. Selection of the diverse and representative patient safety practices (PSPs)

In Figure 3, regarding evaluation questions, the project team (again using the literature, input from experts, and their own experiences in quality improvement and patient safety research) developed a draft monograph proposing three basic types of evaluation questions. This monograph was reviewed by the TEP and then discussed at the July 17, 2009 TEP teleconference. A revised set of evaluation questions was then prepared reflecting the TEP's input.

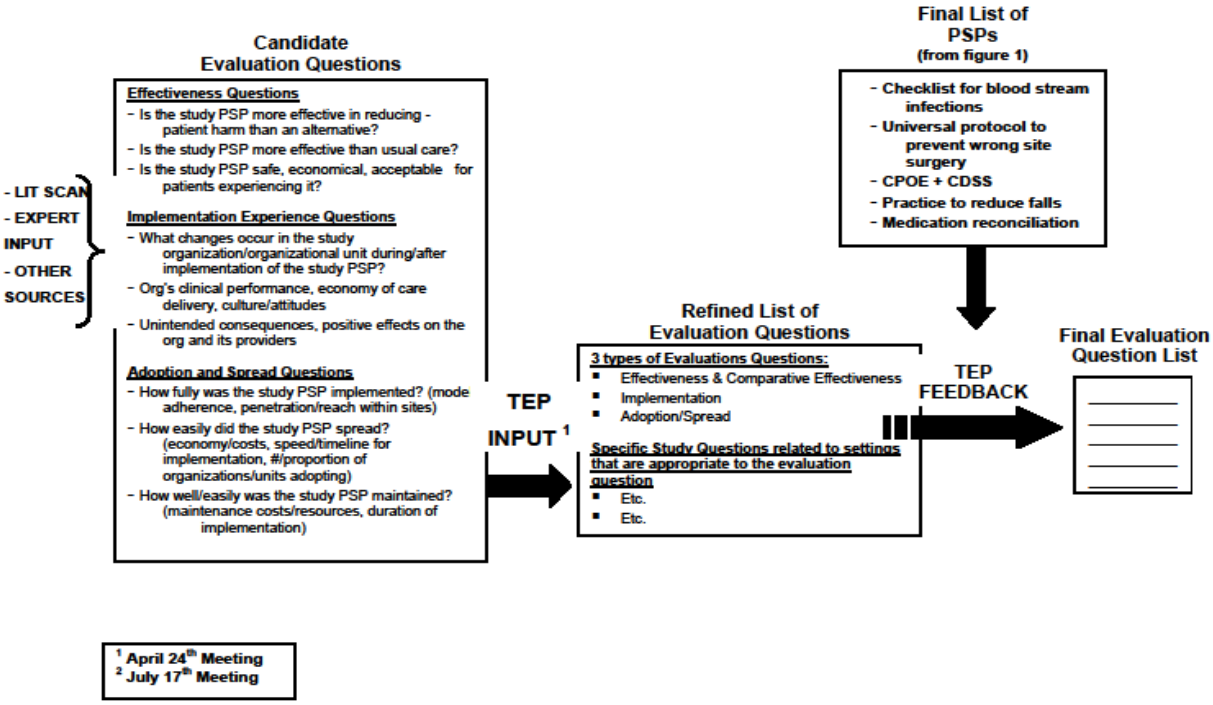


Figure 3. Evaluation questions

In Figure 4, regarding study designs, the project team used existing literature plus input from experts, including key methods experts on the TEP, to come up with a framework of study designs linked to evaluation questions and contexts. The issue about study design continued to be a topic of discussion at the July 17, 2009 TEP teleconference, as well as the November 4-5, 2009 face-to-face TEP meeting. The results informed the report chapter on study design, presented in Appendix I, as well as the criteria for evaluating the body of evidence. An important result of this process was the TEP's recognition that prior arguments conceptualizing the issue as "randomized controlled trials" versus "observational study designs" obscured important elements of assessment that should be included in any well-done evaluation. Another important result of this process was the TEP's agreement on which of those assessment elements were most critical.

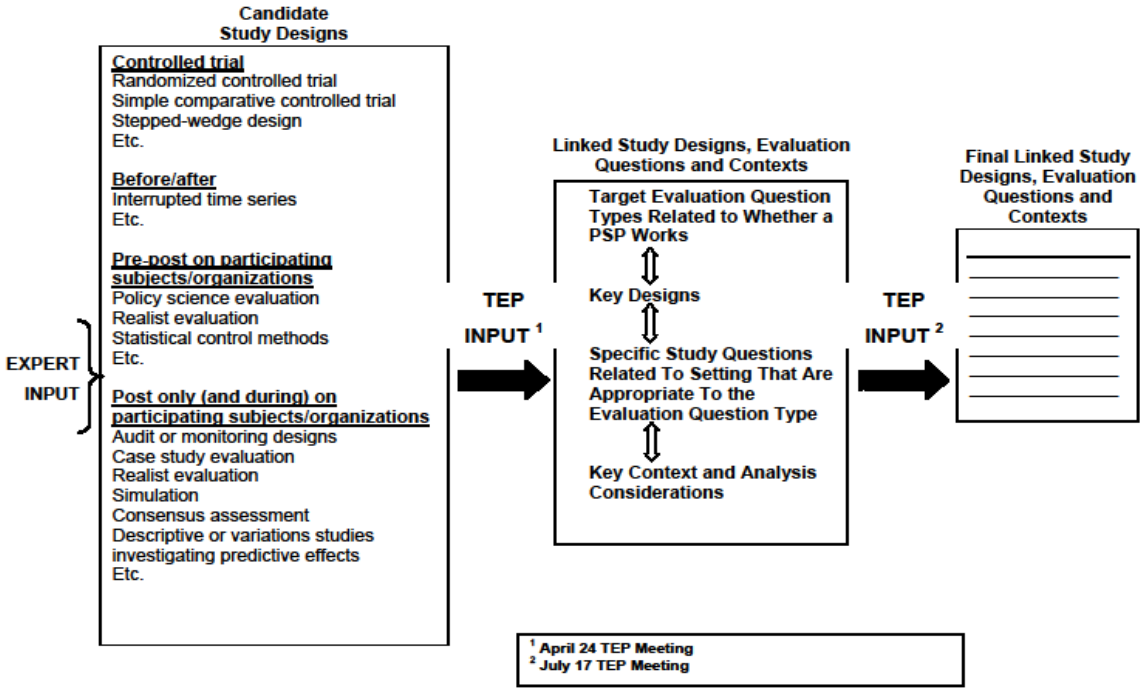


Figure 4. Selecting study designs

In Figure 5, regarding the selection of contexts, the project team again used existing literature, theory, and expert input to come up with a candidate list of potential contexts important for assessment in this project. This list was shortened as a result of TEP input in an Internet survey plus discussion at the July 17, 2009 TEP teleconference. This shortened list was then the subject of a literature review by the project team, assessing the evidence for the influence of these contexts on implementation effectiveness or outcomes. This information helped guide a discussion by the TEP at the November 4-5, 2009 meeting. Subsequently, a revised Internet survey was completed by the TEP, resulting in the final list of contexts.

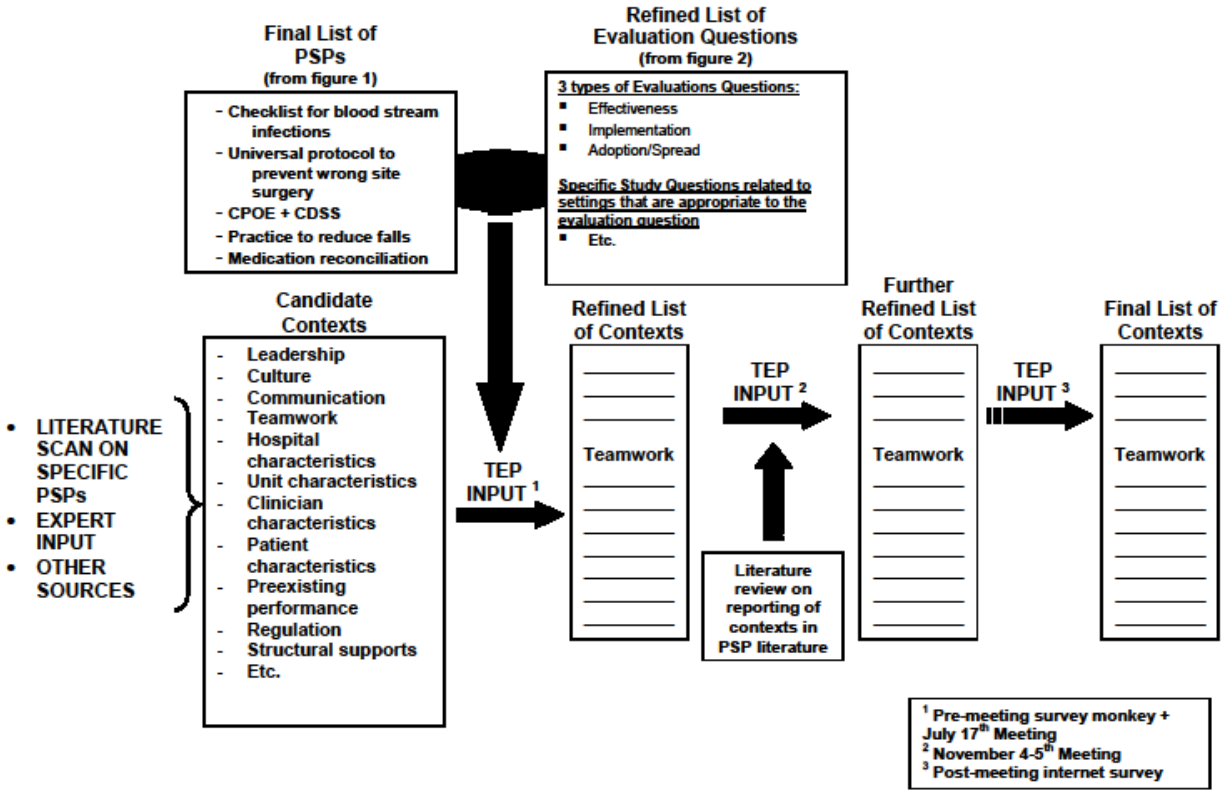


Figure 5. Selecting contexts

In Figure 6, regarding selection of criteria for assessing context-sensitivity, the project team took the shortened list of contexts and reviewed available methods of measuring the key contexts that present measurement challenges (teamwork, leadership, patient safety culture, and organizational complexity). This literature review, in addition to the review of evidence developed in Figure 5, was then used by the TEP to select criteria for measuring these contexts. This was done during a discussion at the November 4-5, 2009 TEP meeting and in a subsequent Internet survey.

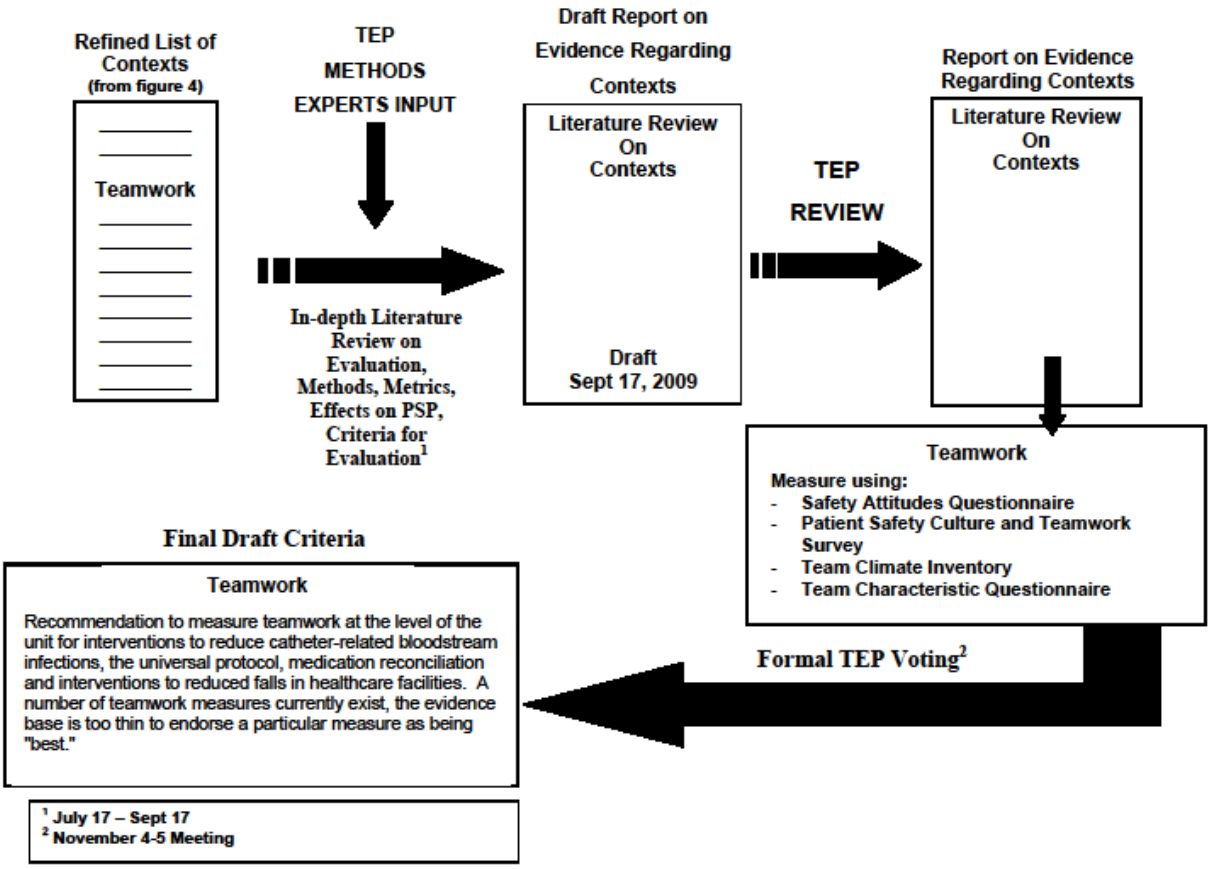


Figure 6. Selecting criteria

Finally, to identify specific needs for future development and research, we first surveyed the project team. We then received feedback from the project officer as well as the project team before surveying the entire TEP after the November 4-5, 2009 meeting.

Chapter 3. Forming an Interdisciplinary Panel of Experts

The technical expert panel (TEP) is a very important aspect of the project. Panel composition matters, both in terms of the panel output and the external credibility of the output. In short, one wants the most relevant disciplines to be represented by the experts with the greatest external credibility in their field, since others in their field will in part base their acceptance of the resulting criteria on their trust in the experts who contributed to the development.

With these principles in mind, we formed an expert panel with broad representation in terms of both the methods of evaluation of effectiveness, implementation, and safety and in terms of diverse groups of patient safety literature stakeholders. We included recognized experts from different patient safety topic areas (such as health IT, hospital-acquired infections, etc.), plus front line health care delivery experts and a journal editor. The members of the TEP are listed in Table 1, and their bioparagraphs can be found in Appendix A, Part 2.

Table 1. Expert panel members: Stakeholders and methodologists

Name	Qualifications
Dr. Alyce Adams	Brings expertise on the determinants of suboptimal use health care services among disparities populations, including racial differences in medication adherence and the impact of changes in health policy on access to high quality health services and health outcomes for vulnerable patients.
Dr. Peter Angood	Inaugural Senior Advisor for Patient Safety at the National Quality Forum (NQF), overseeing development and maintenance of the NQF Safe Practices program and the NQF Serious Reportable Events program and providing oversight for NQF-endorsement of Measures for Patient Safety.
Dr. David Bates	Brings an information technology and medication safety perspective and helps lead the Center of Information Technology Leadership at Partners HealthCare System, Inc; served as the Center Director on one of three national Centers of Excellence in Patient Safety and Research supported by AHRQ.
Dr. Leonard Bickman	Brings expertise in quantitative methods, health services research, and program evaluation; Betts Professor of Psychology and Director of the Center for Evaluation and Program Improvement, Peabody College, Vanderbilt University.
Dr. Pascale Carayon	Brings expertise in human factors engineering and is the Procter & Gamble Bascom Professor in Total Quality and the Director, Center for Quality and Productivity Improvement at the University of Wisconsin, Madison.
Professor Sir Liam Donaldson	Chief Medical Officer for England, United Kingdom; principal advisor to the United Kingdom Government on health matters and one of the most senior officials in the National Health Service (NHS). International leader in health care quality and safety and director of the WHO World Alliance for Patient Safety.

Name	Qualifications
Dr. Naihua Duan	An accomplished practicing biostatistician with research interests in health services research, prevention research, and sample design and experimental design, a Professor of Biostatistics (in Psychiatry) at Columbia University, and the Director of the Division of Biostatistics and Data Coordination at New York State Psychiatric Institute.
Dr. Donna Farley	An expert at conducting rigorous program evaluations of quality improvement initiatives and patient safety interventions who led RAND's Patient Safety Evaluation Center, funded by AHRQ, to evaluate the Federal Government's national patient safety initiative.
Dr. Trisha Greenhalgh	A general practitioner with research expertise in complex innovation in health care, especially electronic health records and the use of narrative methods in health services research.
Dr. John Haughom	Senior Vice President of Clinical Quality and Patient Safety for PeaceHealth, a non-profit, integrated health care system in the Pacific Northwest; responsible for clinical improvement, patient safety initiatives, health services research, outcomes measurement, and all information systems initiatives.
Dr. Eileen Lake	Brings expertise on the contributions of the nurse's work environment and clinical nursing expertise to patient outcomes, as well as expertise on methods for outcomes research.
Dr. Richard Lilford	A physician with expertise in Bayesian analysis and interests in patient safety (particularly as applied to obstetrics/gynecology) who co-authored the recent four-part series on "An epistemology of patient safety research." Dr. Lilford's associate, Dr. Celia Brown, attended the first TEP meeting in his place. Dr. Brown brings expertise in the epistemology of patient safety research and in economics, public health, epidemiology, and biostatistics.
Dr. Kathleen Lohr	Brings over 35 years of experience in health services and policy research; was founding director of the RTI International–University of North Carolina Evidence-based Practice Center and the RTI DEcIDE Center; now serves as senior advisor to both.
Dr. Gregg Meyer	Expert in quality improvement; Senior Vice President for the Center for Quality and Safety at Massachusetts General Hospital; Co-chairman, NQF Executive Institute Task Force on Safe Practices; and previously, Director of AHRQ's Center for Quality Improvement and Patient Safety.
Dr. Marlene Miller	Expert in pediatric quality and patient safety; Vice Chair for Quality and Safety, Johns Hopkins University Children's Center; Vice President of Quality Transformation for the National Association of Children's Hospitals and Related Institutions.
Dr. Duncan Neuhauser	A health services researcher, nationally recognized for research in hospital management, quality improvement, and clinical decision analysis.
Dr. Gery Ryan	An expert in a wide range of qualitative methodology and research and evaluation design.

Name	Qualifications
Dr. Sanjay Saint	Director of the VA/University of Michigan Patient Safety Enhancement Program; his research focuses on hospital-acquired infections.
Dr. Kaveh Shojania	A leader in identifying evidence-based patient safety interventions and effective strategies for translating evidence into practice; co-authored EPC patient safety reports with Wachter.
Dr. David Stevens	Leader of the SQUIRE project, the editor of the journal Quality and Safety in Health Care, is with AHRQ's Center for Quality Improvement and Patient Safety, and directs the Association of American Medical Colleges' Institute for Improving Care.
Dr. Steve Shortell	Dean of the UC Berkeley School of Public Health and an expert in organizational management and behavior, quality improvement, and health services research.
Dr. Kieran Walshe	A health services researcher who is an expert in theory-driven evaluation and in clinical and organizational governance.

Chapter 4. Determining the Target Patient Safety Practices

We selected five types of patient safety practices (PSPs) for the diverse and representative set of practices on which the rest of this project focused:

1. Checklists for catheter-related bloodstream infection prevention.
2. Universal Protocol for preventing wrong procedure, wrong site, wrong person surgery.
3. Computerized physician order entry and decision support system.
4. Medication reconciliation.
5. Interventions to prevent in-facility falls.

We selected these five PSPs after conducting a series of activities, one of which was a survey of the TEP. The full results of that survey are found in Appendix B. Our definitions for the PSPs follow.

Universal Protocol

The Universal Protocol for Preventing Wrong Site, Wrong Procedure, Wrong Person Surgery was created by the Joint Commission and became effective in 2004.[1] The protocol consists of three components: conducting a pre-procedure verification process, marking the procedure site, and a 'time out" session before starting the procedure. The protocol targets a very rare event but one that presumably is a preventable event.[2] It was designed to address surgery errors with tragic consequences but has since been adopted in other fields or has been expanded to non-surgical fields.[3] The Joint Commission recommends the use of a checklist but does not mandate it. Checklists seem to be a prominent way to implement the Universal Protocol and to ensure that its components actually take place.

Medication Reconciliation

Medication reconciliation is the practice of acquiring an accurate medication history at each transition in care.[4] It aims to reduce adverse drug events that result because of medication information that is lost as patients transfer from one setting to another. Many different medication reconciliation interventions have been developed for use by health care providers, but most rely on two main components:

1. Development of forms and procedures to capture information and compare for discrepancies from different sources (e.g., primary care, admission, discharge).
2. Work flow and role assignment among providers (and sometimes patients).

In addition, interventions often include education of providers (and sometimes patients) on the new processes and paperwork (or electronic tools) and audit and feedback regarding compliance with the process and the benefits of reconciliation.

Computerized Physician Order Entry (CPOE) and Decision Support Systems (DSS)

CPOE can be thought of as direct entry of medical orders into the computer. DSS has been described as "a wide range of computerized tools directed at improving patient care, including computerized reminders and advice regarding drug selection, dosage, interactions, allergies, and the need for subsequent orders."[5] However, DSS vary substantially in their features and capabilities.[6] In this context, DSS refers to decision support regarding prescribing to help reduce adverse drug events (check for dosing errors, drug-drug interactions, etc.).

Fall Prevention Programs

Many different interventions have been developed to prevent falls, including multifactorial falls risk assessment and management, exercise, environmental modifications, education, and review of drugs, and programs that target risk factor reduction (identifying and reducing fall risk factors that can be removed or reduced). Risk factor reduction is one component in most programs (e.g. a clinical medication review by a pharmacist and treatment of care home residents). Most falls prevention interventions in institutions are a combination of components (multi-factorial) that may be prescribed for the implementers by label in a "bundle" (e.g., "implement an education program for staff and residents, risk assessment, non-slip mats, and medications review – how you do this is up to you") or not prescribed for the implementers, instead it is a "menu" of labels and examples from which implementers choose.

Blood Stream Infection Prevention Efforts

A large variety of patient safety interventions have been evaluated for reducing central line-associated bloodstream infections (CLABSI).[7] Most are technical, such as avoiding the femoral insertion site and use of specific skin disinfection solutions. However, more recently, a few studies have been oriented towards quality improvement and human factors issues, including elements such as staff education, infection control programs, and feedback. We defined the patient safety practice for catheter-related infection or CLABSI prevention as practices, policies, or checklists to reduce the rate of infections acquired as a result of placement and maintenance of intravascular catheters in hospitalized patients.

References for Chapter 4

1. The Universal Protocol for Preventing Wrong Site, Wrong Procedure, and Wrong Person Surgery. Washington, DC: The Joint Commission; 2009. Available at http://www.jointcommission.org/PatientSafety/UniversalProtocol/. Accessed March 2010.
2. Kwaan MR, Studdert DM, Zinner MJ, Gawande AA. Incidence, patterns, and prevention of wrong-site surgery. *Arch Surg*. 2006; 141:353–7.
3. Angle JF, Nemcek AA, Cohen AM, et al. Quality improvement guidelines for preventing wrong site, wrong procedure, and wrong person errors: Application of The Joint Commission "Universal Protocol for Preventing Wrong Site, Wrong Procedure, Wrong

Person Surgery" to the practice of interventional radiology. *J Vasc Interv Radiol*. 2008; 19(8):1145-51.

4. AHRQ Patient Safety Network – Glossary. Available at http://www.psnet.ahrq.gov/glossary.aspx#M. Accessed September 30, 2010.
5. Wolfstadt JI, Gurwitz JH, Field TS, et al. The effect of computerized physician order entry with clinical decision support on the rates of adverse drug events: A systematic review. *J Gen Intern Med.* 2008; 23(4):451-8.
6. Wright A, Sittig DF, Ash JS, et al. Clinical decision support capabilities of commercially available clinical information systems. 2009; *JAMIA*. 16(5):637-44.
7. Ramritu P, Halton K, Cook D, et al. Catheter-related bloodstream infections in intensive care units: A systematic review with meta-analysis. *J Adv Nurs*. 2008; 62(1):3-21.

Chapter 5. Key Evaluation Questions

Our framework is based on three key types of questions that describe three different aspects of patient safety practice (PSP) evaluation (Figure 7).

1. Effectiveness Questions

- Is the study PSP more effective than usual care?
- Is the study PSP more effective in reducing patient harm than an alternate approach?

2. Implementation Experience Questions

- What changes occur in the clinical or economic performance of the study organization/organizational unit during/after implementation of the study PSP?
- What changes occur in staff/clinician culture/attitudes during/after implementation?

3. Adoption or Spread Questions

- How easily did the study PSP spread (economy/costs, speed/timeline for implementation, #/proportion of organizations/units adopting)?
- How well/easily was the study PSP maintained (maintenance costs/resources, duration of implementation)?
- Were there any unintended changes or incidents during or after implementation?

Figure 7. Three types of evaluations assessing the results of implementing PSPs

We hypothesized that fully addressing context-sensitive effectiveness and safety would require studies addressing all three types of PSP evaluation questions, and that each evaluation question type would imply different methodological approaches. While an important result of this project is a more expansive view of the characteristics of ideal studies of PSP implementations, effectiveness questions are often assessed using experimental or quasi-experimental designs; implementation experience questions frequently use a pre-post design; and adoption, spread, or sustainability questions would require, at a minimum, observational or descriptive designs. Questions on context sensitivity could be framed in terms of each type of evaluation. The information on effectiveness; implementation experience; and adoption, spread or sustainability in relationship to context would be part of the full picture of final judgments on context-sensitive effectiveness and safety.

The conceptual framework we defined is in line with the multi-modal approaches often used in the PSP field. Saying that an effectiveness study should have information on the process or completion of implementation relative to context, for example, will be meaningless if recognition is not also given to the multiple designs indicated by the different types of questions. In other words, we are not interested in someone's opinion of the implementation process or of how context affected it; we are interested in data. The randomized trial that reports an overall PSP effectiveness result without data on context often cannot be effectively applied. Similarly, knowing that small size practices did worse on PSP outcomes than larger ones (context) may be somewhat useful, but it is not likely to be as useful if data have not been collected to understand what went wrong in the smaller practices. For example, was the

PSP fully implemented in these practices? Did small size practices that succeeded better in implementing the PSP show greater impact on expected outcomes? Thus, combined approaches are favored that simultaneously seek rigor in precision regarding effectiveness while also reporting key contextual features and, if possible, assessing contextual inferences on the outcomes of effectiveness, implementation experience, and adoption or spread.

Within the timeline for this project, we focused most of our efforts on the first evaluation question – effectiveness and how it is influenced by context – and did not at all address the evaluation questions about adoption and spread.

Chapter 6. Description of Ideal Evaluation Methods: Overview

The overriding finding of our project is that in order to better understand the context-sensitivity of the effectiveness and safety of patient safety practices (PSPs) we need to move past the discussion of the merits of the traditional study designs aimed at assessing causality (e.g., "randomized trials" versus "observational studies"). We also need to pay far more attention to other important features currently missing from most published reports of PSP implementations. These features include:

- A presentation of why or how the PSP should work. What is the theory supporting why this particular intervention should influence the target patient safety outcome? What is the logic model for how the PSP should work?
- A description of the PSP in sufficient detail that readers could replicate it. PSPs are often complex interventions and cannot be described in only a few sentences.
- A description of key contextual domains.
- A description of the implementation process. For many PSPs, the line between the intervention and the implementation is not sharp, and the intervention and implementation may be considered to be a single construct.
- An assessment of what actually happened during implementation of the PSP. What went as planned, and what happened that was unexpected?
- An assessment of the results achieved, including benefits and harms.
- An analysis of how the effectiveness and safety of the PSP varied as a function of the key contextual domains.

The remaining seven chapters of the report address these features in more detail.

Chapter 7. The Importance of Theory

"There is nothing so practical as a good theory"
Kurt Lewin (1952)

Hand-washing by hospital staff is a patient safety practice widely advocated to reduce hospital infections. But how does hand-washing work to reduce infections? On one level, it is because bacteria cause disease, and hand-washing kills bacteria. On another level, a handwashing policy works because—and only to the extent that—staff regularly wash their hands between patients. Those in charge of implementing handwashing policies might come up with a range of ideas for achieving regular handwashing, such as installing motion-activated alcohol-based antibacterial dispensers at every room entrance in the hospital or instituting an educational campaign that emphasizes doorways as the reminder to wash your hands (e.g. "every time you pass through a door, wash your hands").

The above paragraph contains two types of "theories" regarding the effectiveness of handwashing. The first is a theory of how handwashing reduces hospital acquired infections. The second is a theory of how to establish handwashing as business-as-usual in a particular organization. More generally, theories about patient safety practices (PSPs) may fall into two types: theories about how a given PSP results in better patient outcomes (sometimes called the "PSP action theory") and theories about how to establish and implement PSPs (so-called "PSP implementation theory"). Both types of theory are important. It would be difficult to promote a particular PSP without a rationale for why it might reduce harm, and knowing the PSP implementation theory enables decisionmakers and those responsible for implementation to understand the mechanisms of action at the study site and thus to devise ways to carry out similar actions and changes in their situation. In practice, it may be difficult to reliably distinguish between the two types of theory (how a given PSP works and how to implement it), as PSPs are often multifaceted or embedded within more complex programs.

Changing provider and organizational behavior to apply effective PSPs in routine clinical practice is challenging. The implementation of PSPs has only recently become the subject of research. Implementation success is known to vary. This variation may be due to differences in implementation methods, in implementation fidelity, and also in differences in the context in which implementation is performed. However, such studies of change rarely describe the implementation or the context and do not allow "generalization through theory," often a more efficient and appropriate method of generalization than study replication in many possible settings. Neither do they provide theories that might explain variations in outcomes. Without these descriptions or explanations, decisionmakers lack information to make choices about what is required for successful implementation of PSPs in their service and how to implement them effectively.

One way forward is to carry out multiple studies of PSP implementation in many settings so that decisionmakers can learn from studies in settings similar to theirs. For example, audit and feedback are variably effective in changing provider behavior and clinical outcomes.[1] The effects of this intervention may vary according to elements such as content of feedback (e.g. comparative or not, anonymous or not), intensity (e.g. monthly, annually), method of delivery

(by peer, or non-peer), duration (6, 12 or 24 months), and context (intensive care or nursing home). Varying only five elements produces 288 combinations, without accounting for the need for replication or addition of co-interventions.[2] An alternative and more realistic and efficient approach is to use theories relevant to PSP implementation within evaluations and provide information that allows decisionmakers to better assess implementation feasibility or how best to implement the PSP. For example, an evaluation of implementation of an electronic medical record at a hospital in Sweden was based on theories of implementation. The authors' finding that implementation success was associated with factors in Rogers' Theory of the Diffusion of Innovation (plus additional factors postulated by previous research) strengthens our confidence in the usefulness of that theory and those factors to predict successful implementation in other settings.[3]

What is "the PSP implementation theory"?

The concept of the PSP implementation theory builds on related ideas such as the "logic model,"[4] "treatment theory,"[5] "program theory,"[6] or "theory of change."[7-10] A longer overview of theory in quality improvement has been published by Grol and colleagues.[11]

A logic model describes how an intervention is understood or intended to produce particular results.[4] The logic model 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.[12] "Treatment theory" describes the process through which an intervention is expected to have its effects on a specified target population," in this case, providers or organizations.[5] This "small 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 all would achieve the same "functions" [13] and intermediate objectives on a chain of events that ultimately lead 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."[14] 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.[6, 15, 16]

A "theory of change" is usually used to describe how those responsible for implementation understand an intervention to work.[7-9] It may be explicit, or it may exist as a theory in the sense of being unspoken assumptions or beliefs. Dixon-Woods et al[10] 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.

These types of theories focus on the intervention and conceptualize it as a chain of events, often in a linear sequence, that lead through intermediate changes (including changes in provider and organizational behavior) to final results (clinical or cost outcomes). More sophisticated variants, often relevant to some combined or "bundled" safety interventions, view the implementation as a number of interacting components with a synergistic and system effect.

A wider conceptualization of "PSP intervention theory" goes beyond the focus on the intervention and its causal chain to include an understanding of contextual influences and how they help and hinder implementation of the PSP. A contemporary example of this conceptualization is the realist evaluation idea of context-mechanism-outcome configurations—a theory of an intervention "triggers" action only in a particular context "primed" to be responsive to the intervention and where the intervention can "take hold."[17] As yet, the details of how to design and carry out studies to build these more complex "context sensitive" intervention theories are still being developed. An important difference from experimental designs is that influences other than the program 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.

In summary, the "PSP implementation theory" builds on related concepts such as logic models, treatment theory, and program theory. In practice, use of any of these concepts would improve our current understanding of PSP implementations.

Why do we need to know the "PSP implementation theory"?

Systematic reviews of interventions to improve the quality and safety of care consistently indicate that most interventions, across different categories, are effective some of the time, but not all of the time, and that intervention effects range from none to large.[18] However, very few such reviews are explicit about the underlying PSP action or implementation theories, let alone use them to explore causes of variation in effectiveness. Many studies of interventions to promote safety currently categorize features of interventions, targeted practices, and contexts on a superficial basis, e.g. computerized decision support systems (CDSS), prescribing, and urban hospitalsm respectively.[2] Such classification systems are really descriptive typologies rather than theoretically meaningful groupings. They may be as unhelpful or misleading as classifying drugs into groups according to whether they are taken orally or intravenously or by the color and size of the pill.[19, 20] It is not surprising that systematic reviews based on such categories or typologies raise more questions than they answer and struggle to extract generalizable lessons about how interventions achieve their effects.[21] For example, a CDSS can work in a number of ways, such as by increasing knowledge of safe practice, reinforcing motivation, or prompting recall, and its effects may vary according to what types of clinical behavior are targeted, whether it is used with co-interventions, and so forth. The mechanisms by which more complex interventions work, such as those to reduce falls or rapid response teams, may be both more variable and more sensitive to contextual features.

Therefore evaluations of PSP implementation need to address the processes by which interventions interact with contextual features and outcomes. For example, RCTs ideally should be accompanied by parallel process evaluations that assess the changes in processes, both intended and unintended, that may have contributed to changes in outcomes.[22]

Improving safety research with "PSP implementation theory"

Theory has not commonly been used in the field of quality and safety research.[23] Within a review of 235 implementation studies, only 53 used theory in any way, and only 14 were explicitly theory-based.[24] Similarly, most reports of PSP evaluations do not provide theory or the logic model underpinning the intervention. Even for the five representative PSPs chosen for this project, which are among the most commonly studied PSPs, our review of publications of evaluations of the PSPs found only two articles that even partially reported a theory for why the PSP should work.

Theory can guide or be applied to patient safety research in a range of ways, including the following.

Explaining clinical and organizational behavior. Just as with clinical practice, it is important to diagnose the causes of adherence or non-adherence to recommended practice before intervening. For example, theories of human error suggest that there are several causes of discrepancies between intended plan and actual action, such as slips and lapses leading to the wrong execution of an action sequence.[25] Recognition of such human limitations has led to better equipment design (e.g. alarms within anesthesia machines).[26]

Selection or tailoring of patient safety interventions for a given problem and context. Previous research or practitioner reports can be used to create hypotheses or a provisional model of which actions may lead to which intermediate changes and which context factors may be important for implementation. Researchers can draw on this provisional implementation theory to decide which data to gather, or operationalize variables, to be able to describe implementation actions and intermediate changes, as well as which aspects of context were or were not helpful to the implementation actions. For example, McAteer et al.[27] developed an intervention to increase levels of providers' hand hygiene behavior using psychological theory for evaluation in a cluster randomised trial. This involved a review of effective behavior change techniques to inform the theoretical approach taken, development of intervention components with clinicians, and focus groups with the targeted provider groups. It may be that the customization of interventions is more necessary than we appreciate.

Evaluating implementation and mechanisms of action. Theory can be used to help predict or evaluate the process of implementation, potentially distinguish between action theory failure and implementation failure, identify mechanisms of action, and shed light on whether the PSP worked (or not) as hypothesized or by an alternative means, and identify unanticipated outcomes or unintended consequences. For example, Byng et al.[28] conducted a qualitative interview study alongside an RCT of a multifaceted intervention to improve the care of people with long-term mental illness. They used a realist evaluation approach to delineate which aspects of the intervention had the greatest impact.

It should be borne in mind that theory is not enough by itself to justify the implementation of a PSP. For example, a program theory may strongly suggest that an intervention works as predicted, but 'triangulation' via experimental or quasi-experimental data may fail to support this.[29]

Conclusions

The “PSP implementation theory” is a representation of how actions lead to changes in provider and organizational behavior as a result of the PSP and, ultimately, affect patient outcomes. Yet, theoretical perspectives have, hitherto, seldom been incorporated into PSP evaluations. This lack of description and explanation of the assumptions or logic behind the PSP makes it more difficult for others to reproduce or adapt the PSP. Future evaluations should be theory-driven, in order to enhance generalizability and help build a cumulative understanding of the nature of change.

References for Chapter 7

1. Jamtvedt G, Young JM, Kristoffersen DT, et al. Audit and feedback: Effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews 2006; Issue 2, Art. No: CD000259. DOI: 10.1002/14651858.CD000259.pub2.
2. The Improved Clinical Effectiveness through Behavioural Research Group (ICEBERG): Designing theoretically informed implementation interventions. *Implement Sci*. 2006; 23(1):4.
3. Øvretveit J, Scott T, Rundall TG, et al. Improving quality through effective implementation of information technology in healthcare. *Int J Qual Health Care*. 2007; 19(5):259-66.
4. Rogers, PJ. 'Logic models' in Mathison S. (ed). Encyclopedia of evaluation. Beverly Hills, CA: Sage Publications; 2005, p. 232.
5. Lipsey MW. Theory as method: Small theories of treatments. In: Sechrest LB, Scott AG, eds. Understanding causes and generalizing about them. New Directions for program evaluation. San Francisco, CA: Jossey-Bass; 1993, pp. 5-38.
6. Bickman L, ed. Using program theory in program evaluation.New directions in program evaluation, Vol. 33. San Francisco, CA: Jossey-Bass; 1987.
7. Sullivan H, Stewart M. Who owns the theory of change? *Evaluation*. 2006; 12:179e99.
8. Mason P, Barnes M. Constructing theories of change: Methods and sources. *Evaluation*. 2007; 13:151e70.
9. Connell JP, Klem AC. You can get there from here: Using a theory of change approach to plan urban education reform. *J Educ Psychol Consult*. 2000; 11:93e120.
10. Dixon-Woods M, Tarrant C, Willars J, Suokas A. How will it work? A qualitative study of strategic stakeholders’ accounts of a patient safety initiative. *Qual Saf Health Care*. 2010; 19:74e78.
11. Grol RPTM, Bosch MC, Hulscher MEJL, et al. Planning and studying improvement in patient care: The use of theoretical perspectives. *Milbank Q*. 2007; 85(1):93-138.
12. Wholey, J. Evaluation and effective public management. Boston, MA: Little, Brown; 1983.

13. Hawe P, Shiell A, Riley T. Complex interventions: How out of control can a randomised controlled trial be? *Br Med J*. 2004; 328:1561-3.

14. Bickman K. The science of quality improvement (letter to the editor). *JAMA*. 2008; 300(4):391

15. Weiss, C. How can theory-based evaluation make greater headway. *Eval Rev*. 1997; 21:501-8.

16. Weiss, C. Theory-based evaluation: past present and future. In: Progress and future directions in evaluation: Perspectives on theory, practice, and methods: New directions for evaluation, No. 76, pp. 41-55. San Francisco, CA: Jossey-Bass; 1997.

17. Pawson R. Evidence-based policy: A realist perspective. London: Sage; 2006.

18. Grimshaw JM, Thomas RE, MacLennan G, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technol Assess*. 2004; 8.

19. Shojania KG, Grimshaw JM. Evidence-based quality improvement: The state of the science. *Health Aff*. 2005; 24:138-50.

20. Foy R, Francis J, Johnston M, et al. The development of a theory-based intervention to promote appropriate disclosure of a diagnosis of dementia. *BMC Health Serv Res*. 2007; 7:207.

21. Foy R, Eccles M, Jamtvedt G, et al. What do we know about how to do audit and feedback? Pitfalls in applying evidence from a systematic review. *BMC Health Serv Res*. 2005; 5:50.

22. Oakley A, Strange V, Bonell C, et al. Process evaluation in randomised controlled trials of complex interventions. *Br Med J*. 2006; 332:413-6.

23. Walshe K. Understanding what works – and why – in quality improvement: The need for theory-driven evaluation. *Int J Qual Health Care*. 2007; 19(2):57-9.

24. Davies P, Walker AE, Grimshaw JM. A systematic review of the use of theory in the design of guideline dissemination and implementation strategies and interpretation of the results of rigorous evaluations. *Implement Sci*. 2010; 5:14

25. Parker D, Lawton R. Psychological contribution to the understanding of adverse events in health care. *Qual Saf Health Care*. 2003; 12:453-7

26. Thompson PW. Safer design of anaesthetic machines. *Br J Anaesth*. 1987; 59:913.

27. McAteer J, Stone S, Roberts J, et al. Use of performance feedback to increase healthcare worker hand-hygiene behaviour. *J Hosp Infect*. 2007; 66(3):291-2.

28. Byng R, Norman I, Redfern S, Jones R. Exposing the key functions of a complex intervention for shared care in mental health: Case study of a process evaluation. *BMC Health Serv Res*. 2008; 8:274.

29. Brown C, Hofer T, Johal A, et al. An epistemology of patient safety research: A framework for study design and interpretation. One size does not fit all. *Qual Saf Health Care*. 2008; 17:178-81.

Chapter 8. Description of Ideal Evaluation Methods: Describing Patient Safety Practices

One of the key issues in patient safety practice (PSP) research and literature is adequately defining and describing PSPs. The key goal in standardizing PSP descriptions is to provide sufficient detail on the PSP and its implementation to assess: (a) whether it improves safety, (b) the risks that random or systematic error influences the results, and (c) the applicability to other situations for others to replicate or adapt it. Other goals for standard descriptions include the ability to make biases transparent, evaluate inclusion criteria and determine heterogeneity in literature reviews, and improve the ability of journal editors and reviewers to assess the quality of the information. These goals require some basic information on the PSP itself and on key elements of what was done.

Describing the intervention in sufficient detail so that it can be replicated is a requirement included in reporting guidelines such as SQUIRE for quality improvement studies, and it also was endorsed by our technical expert panel (TEP). The key challenge in PSP studies—in contrast to pharmaceutical or surgical interventions that are concrete or highly standardized and can be precisely defined, described, and delivered across settings—PSPs generally lack sharp boundaries between the practice, implementation process, and context. Indeed, as the science regarding effective implementation strategies and the understanding of context increase, implementation and context often merge together to become part of the intervention. PSP interventions are inherently dependent on context and vice versa—some PSPs include influencing context as part of the intervention. These interventions are often fluid, since some PSPs incorporate into "the intervention" unforeseen necessary adaptations or lessons learned during the implementation process. While this could be a major problem for clinical outcome studies (for example, if physicians were allowed to vary the dose of the study drug or add additional co-interventions at their discretion), the TEP felt that such fluidity was welcome in PSP studies as long as it was measured and described well. The intervention is often iterative, evolving in response to outcome feedback and changing as context changes through the implementation process. Finally, fully developed PSPs are often complex and include multiple components. While the knowledge regarding how to effectively implement a PSP and the effect of context on implementing a PSP is growing rapidly, it is still a relatively immature science.

Existing descriptions of PSPs in the published literature vary widely, are often incomplete, and criteria to guide researchers on how to describe PSPs, the implementation strategies, or the contexts are lacking. Some organizations have produced criteria for describing a quality improvement intervention, which may have relevance to describing a PSP. For example, the SQUIRE guidelines are an important step in describing quality improvement studies generally aimed at directly changing provider behaviors.[1] Table 2 contains an excerpt of the SQUIRE guidelines (the full set of guidelines is in Appendix M; see also http://squire-statement.org/).

Table 2. SQUIRE: Relevant elements (http://www.squire-statement.org/guidelines)

Setting	
Planning the intervention	▪ Describes the intervention and its component parts in sufficient detail that others could reproduce it. ▪ Indicates main factors that contributed to choice of the specific intervention (for example, analysis of causes of dysfunction; matching relevant improvement experience of others with the local situation). ▪ Outlines initial plans for how the intervention was to be implemented: e.g., what was to be done (initial steps; functions to be accomplished by those steps; how tests of change would be used to modify intervention), and by whom (intended roles, qualifications, and training of staff).
Results	▪ Outcomes. ▪ Explains the actual course of the intervention (for example, sequence of steps, events or phases; type and number of participants at key points), preferably using a time-line diagram or flow chart. ▪ Documents degree of success in implementing intervention components. ▪ Describes how and why the initial plan evolved and the most important lessons learned from that evolution, particularly the effects of internal feedback from tests of change (reflexiveness).

The NQF also lists requirements for describing quality improvement practices (See Table 3). These guidelines and requirements are generic for quality improvement interventions.

Table 3. National Quality Forum practice description key elements

The practice must be a clearly and precisely defined process or manner of providing health care services.	▪ Target outcome or objective of practice - A specific description of the effect a practice is intended to have. ▪ What does the practice entail? Specify elements that are considered to enhance the likelihood of achieving the target outcome for the practice and details related to implementation that would result in relatively uniform and comparable practices and outcomes across implementing entities. ▪ For what encounters/patient populations is the practice indicated? ▪ Who should perform the practice? ▪ Theoretical or clinical rationale for the effect of the practice on target outcome.
Readiness	▪ What technology/tools are necessary to perform the practice? ▪ What personnel qualifications are required to perform the practice? ➢ What additional staff training is required for implementation?

No criteria exist for how to describe the key components of a PSP intervention. Thus, we developed general principles of a PSP description based on theory and existing frameworks. We also created two specific examples based on expert opinion from the project team. Key concepts for describing a PSP and additional elements that can be helpful are included in Box 1. In describing these concepts, we recognize that the borders between implementation and context are not sharp, and that the delineations may evolve as our understanding of important contextual factors grows. Currently, the line between what is context and what is PSP is often not clear. Likewise, the PSP and its implementation are usually too intertwined to tease them apart. The criteria we present here are limited by the lack of evidence to state what should be described about a PSP, so they should be considered as general concepts that appear to be important based on principles of behavior change, other guidelines, and the overall finding of this project.

Box 1 - Key concepts in describing a patient safety practice

Target: Patient safety problem practice is intended to address:

- Key elements of the intervention.
- Sufficient detail of the implementation process to allow relatively comparable adaptation to another entity.
- Population or settings where practice applies; is the intervention intended to apply to a single level of the organization or multiple levels, and if so, how many?
- Health care professionals and administrators that were involved.
- Personnel qualifications and additional staff training.
- Required technology or tools.

Implementation detail should include:

- Initial plans for intervention.
- Step-by-step explanation or diagram of intervention.
- Description of elements key to association with improved outcomes.
- Whether the PSP implementation requires education for executive leaders, team leaders or staff.
- How the study or implementation team ensured that the executive leaders, team leaders, and staff executed the PSP implementation
- How the study or implementation team evaluated whether the actions of executive leaders, team leaders, and staff made a difference.
- Other barriers and facilitators of change and how these were addressed.
- Steps for ensuring that the PSP was implemented as designed (or measuring how it was actually implemented and why changes were made).

Other possible elements to include:

- Factors that contributed to choice of the specific intervention (e.g., analysis of causes of dysfunction; other interventions that might have been considered).
- Resources used and which were required; feasibility.
- Likelihood that observed gains may weaken over time.
- Plans, if any, for monitoring and maintaining improvement (or state that such planning was not done).
- Plans for moving from study to maintenance.
- Ease of incorporating PSP into clinical practice or systems.

- Generalizability of the elements of the intervention to other settings: unique features of the organization or external environment that may have influenced impact of the PSP.
- Risk for unintended consequences.

To supplement these generic criteria for describing the PSP, we provide here more specific criteria for describing two of our five target PSPs: catheter-related bloodstream infection prevention and CPOE. We solicited input from two members of the project, Peter Pronovost and David Bates, who are international experts in designing, implementing, and evaluating these two PSPs. As such, these criteria are based mostly on expert opinion. Nevertheless, in the absence of strong empirical evidence, the opinions of recognized experts can be a valuable source of guidance for implementers and evaluators. Criteria for describing interventions to prevent catheter-related bloodstream infections are presented in Box 2, and criteria for describing CPOE interventions are presented in Box 3.

Box 2 - Case example: Key elements in describing PSP to prevent catheter-related bloodstream infections (CRBI)

- Items on the checklist supported by strong evidence from the Centers for Disease Control and Prevention and professional societies.
- Efforts to ensure patients receive checklist items.
- Who is intended to do what task when the PSP is implemented?
- Staff education and training.
- Internal incentives.
- Whether there was local tailoring or an iterative process in the intervention implementation.
- Selection and involvement of leadership (unit and executive).
- Improving patient safety culture and teamwork.
- Providing evidence summaries and standardized measures.
- Identifying local barriers.
- Removing barriers to comply with checklist.
    - Creating central line carts that store all needed supplies.
    - Asking hospital leaders to purchase central line kits that have chlorhexadine.
    - Improving culture and teamwork through the Comprehensive Unit-based Safety Program.

Box 3 – Case example: Key elements in describing studies of CPOE

- To what extent clinical and operational leadership were involved in building support.
- Staff education and training over time.
- Extent of tailoring in implementation.
- Pace of implementation.
- Project management during implementation.
- Response time of application.
- Level of clinical decision support implemented.
- What type of process was put in place to identify issues with the application and correct them?
- Measurement of alert frequency and responses to alerts.

Our Proposed Framework for a PSP Classification System

Through a process of synthesizing existing conceptual frameworks and an expert panel consensus process, we developed a conceptual framework for describing the dimensions of PSPs. The framework includes 11 dimensions, as shown in Table 4. The process of how this was developed is described in Appendix 2.

Table 4. Classification dimensions for patient safety practices

Dimension	Definition and examples
Regulatory versus voluntary	Whether required by external entity, such as the Joint Commission
Setting	Hospital, nursing home, ambulatory
Feasibility	Ability to implement PSP in a variety of settings, even in small facilities[12]
Individual activity vs. organizational change	Whether the target of the PSP is individual providers' behavior (e.g., handwashing) or the structure of the organization
Temporal (one-time vs. repeated/long-term)	Structural change (e.g., switch to antibiotic-impregnated catheters) or change that requires regular maintenance (e.g., hand hygiene education)
Pervasive in setting vs. targeted to specific units or providers	Whether the PSP addresses a safety issue that applies to all patients in a unit or setting (universal protocol would apply to all surgeries, but fall prevention would be targeted to at-risk patients)
Common vs. rare event as target	Whether the patient safety event that the PSP is intended to address is relatively common (e.g., medication errors) or rare (e.g., wrong-site surgery)
PSP maturity, established vs. newer	Whether the PSP has been well-studied, and implementation needs are well-known
Degree of controversy or conflicting evidence (or both)	Whether the PSP is widely accepted; whether examples of ineffective PSPs exist
Degree of behavioral change required for implementation	How much the PSP implementation involves human factors issues (e.g., an institutional policy switching to use of chlorhexidine would not depend on provider behavior)
Sensitivity to context	Whether the success of PSP implementation depends on issues such as leadership, patient safety culture, or teamwork

Reference for Chapter 8

1. Davidoff F, Batalden P, Stevens D, et al. SQUIRE Development Group. Publication guidelines for improvement studies in health care: Evolution of the SQUIRE Project. *Ann Intern Med*. 2008; 149(9):670-6.

Chapter 9. Description of Ideal Evaluation Methods: Selecting Key Domains of Context

As previously noted, we lack a universally agreed-upon definition of what constitutes "context." Context can be conceptualized as consisting of a discrete number of known constructs (e.g., organizational complexity, patient safety culture, etc.) all the way to everything that is currently unexplained or unknown about why a patient safety practice (PSP) implementation succeeds or fails. In our discussion with the technical expert panel (TEP), we constantly found ourselves asking (when considering a particular construct), "Is this context or is it part of the intervention?" Consequently, we determined that trying to reach agreement on what constitutes the boundaries of context would not be as fruitful as concentrating on a limited group of constructs that all agreed were important and could be considered contextual variables. Hence, in this report there is no overarching definition of what context "is," but rather thee is a determination and discussion of variables believed to be important in understanding PSP implementation and effectiveness that currently do not receive the attention they deserve.

To begin this process, we used existing published papers regarding our five representative PSPs and our own knowledge to come up with a long list of potential influences on PSP effectiveness that might be considered context. We next surveyed the TEP in June 2009, and asked them to rate the importance of each contextual feature for each of the five PSPs. Based on the results of the survey, we developed a scheme for classifying and selecting PSP contexts for the next phase of the project. We attempted to take into account things like mutability (from the organization's perspective), whether PSPs were tactical or not (in terms of tactics that might be used to enhance implementation success), their measurement ability, and the evidence base supporting their importance. With input from the TEP, which included an Internet survey and a long teleconference discussion, we shortened the long list to a number of "high priority" contextual variables, which we then organized into four domains:

1. External factors. These were all rated separately, were rated high in the Internet survey, and are related. Also, none are mutable from the organization's perspective but could be mutable by policymakers;

2. Structural organizational characteristics, such as size, complexity, and financial status or strength. These are not mutable and might have a bi-directional effect on PSP implementation, with increasing size and complexity facilitating some PSP implementations but making others more difficult;

3. Teamwork, leadership, and patient safety culture. These were all rated separately and rated high in the Internet survey. Although it is unclear how independent they are, they all somewhat address the social or cultural aspects of a PSP implementation context;

4. Management/implementation tools, including training resources, internal organizational incentives, audit and feedback, and collaboration with QI consultants. These are all factors that researchers can influence while implementing the intervention.

After reviewing the available literature regarding these contexts and our five representative PSPs (see Appendix A, Part 1, section on contexts), we then discussed these contextual variables in more detail with the TEP. As a result of this process, we separated an assessment of contexts into "important for describing context" and "important in assessing the effect of context on implementation success." The former was judged to be important so that health care organizations could better assess the applicability of a PSP implementation to their own institution. We then conducted a second Internet-based survey of the entire TEP to determine which of the 32 contexts the TEP thought had a high priority for data collection. We asked the TEP to consider this question for each of the five PSPs when either assessing the effect of context on this PSP implementation or describing context in papers. The results of that survey are summarized in Table 5, which shows the contexts that respondents voted as "high priority" when assessing the effect of context or describing the context (full results of the survey are in Appendix D).

Table 5. Results of Survey of High Priority Contexts

Context	Infection Checklist	Universal Protocol	CPOE/DSS	Medication Reconciliation	Falls
1. Structural Organizational Characteristics					
Size	✓		✓	✓	✓
Location					
Financial status			✓		
Academic status				✓	
Organizational complexity		✓	✓	✓	
Date of study					
Volume	✓				
Existing quality/safety infrastructure		✓	✓	✓	✓
Space/physical environment					
Past experience with IT			✓		
Physician ownership			✓		
2. External Factors					
Regulatory requirement	✓	✓	✓	✓	✓
Payments or penalties	✓	✓	✓	✓	✓
Local sentinel event		✓			✓
Marketplace competition					
Competing demands					
3. Patient Safety Culture, Teamwork, Leadership					
Patient safety culture – org. level				✓	✓
Patient safety culture – unit level	✓	✓		✓	✓
Teamwork – org. level					
Teamwork – unit level	✓	✓		✓	✓
Leadership – org. level			✓	✓	✓
Leadership – unit level	✓	✓		✓	✓
4. Implementation and Management Tools					
Staff education and training	✓	✓	✓	✓	✓
Designated staff time to implement	✓	✓	✓	✓	✓
Use of audit and feedback	✓	✓		✓	✓

Context	Infection Checklist	Universal Protocol	CPOE/DSS	Medication Reconciliation	Falls
Internal or external person responsible for implementation	✓	✓	✓	✓	✓
Internal incentives	✓		✓		✓
Local tailoring or iterative process	✓	✓	✓	✓	✓
Helpdesk support			✓		
Extent of project management			✓		
Timeline of implementation			✓		
Implementation process – one unit at a time or all at once			✓		

Note: IT = information technology; org = organization; PSP = patient safety practice; CPOE = computerized provider order entry; DSS = decision support system
Universal Protocol = for preventing wrong procedure, wrong site, wrong person surgery;
Bloodstream Infection Checklist = catheter related

Tools for Measuring Key Domains of Context

One of the goals of the project is to suggest ways to measure contexts. Many of these contextual features have not posed a measurement problem in prior studies (i.e., size, location, academic status, regulatory requirement, use of audit and feedback, etc.). Other contexts do pose a measurement challenge, such as teamwork, leadership, patient safety culture, and organizational complexity. We concentrated on these four in our efforts to determine ways to measure context. To help guide a discussion of how these contexts might be measured, we did an extensive literature search in the health care peer-reviewed and "gray" literatures for measures. The measures we found for teamwork, leadership, and patient safety culture were sufficiently on target to include in subsequent activities. However, we did not find many measures of organizational complexity, even after expanding our search to the organization and business literature, such that we could not assess the relative strengths of measures of this context. Development of organizational complexity measures relevant for PSP evaluation remains an area for future development. All measures we found for the four contexts are listed in Appendix E.

Given that there are multiple measures, and no one measure is superlative in all aspects, expert judgment is needed to help select the measures that might be more appropriate to use. As such, in late November 2009, we surveyed the TEP using another Internet-based survey to determine their opinions on some of the measures we found for teamwork, patient safety culture, and leadership (we did not survey the TEP on the measures of organizational complexity because there were too few). As in our prior survey about contexts, some TEP members abstained on the grounds that they were not expert in this area. We also heard from several TEP members that they did not think the field was sufficiently advanced to recommend specific measures. For example, one TEP member said: "I think this is beyond the scope (of what we can do)," and another TEP member said "this is futile, there are too many to choose from, and the choice of instrument would depend on the nature of the work being done." Other panelists, while acknowledging that the evidence is too thin to support any one particular instrument, argued that

providing an expert opinion-based recommendation was still useful, since evaluators and researchers have choices that need to be made about which instruments to use, and the guidance of experts is better than no guidance at all. One expert put it this way: "Expert opinion is often the best we have at a moment in time, and making use of expert opinion is not, in any logical sense, tantamount to accepting or endorsing its validity. Nor does it preclude further and different work."

In the end, we received between 11 and 14 responses (depending on the context; out of 22 possible participants) to our questions about which measures to use. This means that no measure could have received the 15 votes necessary for us to consider it the TEP recommendation. Furthermore, even when assessed as a proportion of actual respondents, no measure received endorsement above the 75 percent threshold that would constitute sufficient agreement for a recommendation from those who did respond.

For these reasons, our conclusion is that the evidence base is too thin and agreement among experts insufficient to make strong recommendations about which measures are preferred for assessments of patient safety culture, teamwork, and leadership, suggesting the need for ongoing dialogue among researchers. However, for patient safety culture, the most support was given to the various AHRQ surveys relevant to this topic, plus the Safety Climate Scale[1] and the related Safety Climate Survey.[2] For teamwork, the most support was given to the ICU Nurse-Physician Questionnaire;[3] no other measure received more than half the votes of respondents. Finally, for leadership, the measures receiving the most support were the ICU Nurse-Physician Questionnaire,[3] the Leadership Practice Inventory, [4] and the Practice Environment Scale.[5] No other measure received more than half the votes of respondents. The full results of the survey are presented in Appendix F.

References for Chapter 9

1. Pronovost P, Weast B, et al. (2003). Evaluation of the culture of safety: Survey of clinicians and managers in an academic medical center, *Qual Saf Health Care*. 12:405-10.
2. Sexton, et al. [as found in paper by Kho M., Carbone J, et al.] Safety Climate Survey: Reliability of results from a multicenter ICU survey. *Qual Saf Health Care*. 2005; 14:273-8
3. Shortell SM, Rousseau DM, Gillies RR, et al. Organizational assessment in intensive care units (ICUs): Construct development, reliability, and validity of the ICU nurse-physician questionnaire. *Med Care*. 1991; 29(8):709-26.
4. Tourangeau AE, McGilton K. Measuring leadership practices of nurses using the Leadership Practices Inventory. *Nurs Res*. 2004; 53(3):182-9.
5. Lake ET. Development of the practice environment scale of the nursing work index. *Res Nurs Health*. 2002; 25(3):176-88.

Chapter 10. Description of Ideal Evaluation Methods: Measuring and Describing the Implementation Context

Special contribution from Gery Ryan, Ph.D., RAND, Santa Monica, CA

Clearly any attempt to replicate a patient safety practice (PSP) intervention will vary from context to context. If we wish to compare PSPs across contexts, then ideally we would like to be able to describe systematically each context to determine how it is similar to and different from each other context. We also need to determine to what degree (if any) these similarities and differences might have an effect on the effectiveness of the PSP intervention being studied. Measuring both the implementation process and how context influences the process are part of an ideal and rigorous evaluation.

The context in which an intervention is being implemented can logically be divided into a description of two main categories: (a) the intervention and how it was operationalized and (b) the physical and organizational context in which the intervention was embedded.

The Intervention Context

All interventions can be described as someone doing something to someone else for a particular purpose. So, at a minimum, we need to clearly understand the following:

3. Who are the interveners?
    a. How were they selected?
    b. What role do they play in the organization?
    c. What is their relationship with the intended intervenees (i.e., the targets of the interventions)?
4. Who are the intended intervenees?
    a. How were they selected (if selected)?
    b. What role do they play in the organization?
    c. What is their role vis-à-vis patients?
5. What specifically are the interveners doing to the intervenees?
    a. How consistent is the interveners' behavior across intervenees (fidelity)?
6. What (if any) new technology or changes to physical plant, organizational structures, or policies and procedures were introduced?
    a. To what degree do intended intervenees vary in their exposure to these changes?
7. How is the intervention expected to influence the behavior of the intervenee(s)?

Consider the example of an education-based intervention to train front-line staff to wash their hands before and after contact with a patient. It would be ideal to know who was conducting the training, who they trained, what kind of training was provided, in what format, for how long, and how the training was expected to affect the specific behaviors of those trained. In some cases, the intervener may not seem obvious, for instance, administrators may change policies or may introduce new technologies or facilities. It is important to know if it was the quality control department that introduced new sinks outside of exam rooms on their own, or if it was the facilities department that made the decision because of new State regulations.

Often interventions are made up of multiple components. Each component should be treated as a separate intervention and described in the manner above, although the action of the components may not be independent of one another.

Knowing how each intervention component was operationalized is also important. At a minimum, we need to understand clearly the following:

1. To what degree and how were expectations of the intervention made explicit to the intended interveners and intervenees?
2. What (if any) kinds of positive and negative incentives (e.g., monetary, prestige, in-kind incentives, reprimands, or other disincentives) were used to motivate interveners or intervenees?
3. How (if at all) was the performance of the interveners or intervenees monitored?
4. What kinds of feedback or consequences (if any) did interveners or intervenees experience for meeting or not meeting what was expected of them?

These latter four questions essentially describe the degree to which the expectations, incentives, monitoring procedures, and resulting consequences of an intervention (or an intervention component) have been made explicit to all the players involved. We find that evaluators often overlook these four topics in describing PSP interventions, but they do play a significant role in the intensity and success of a program.

The Physical and Organizational Environmental Context

To describe the physical and organizational context in which a PSP is embedded, implementers and evaluators first need to describe clearly the patient safety behavior that they are trying to improve and the range of people who are involved. Four types of players or organizational units are important: (1) the people directly responsible for ensuring that the patient safety behaviors are carried out; (2) the people who are responsible for initiating and carrying out the patient safety interventions; (3) the unit(s) within the organization where the patient safety behavior of interest is located; and (4) policymakers (e.g., at the State, Joint Commission, etc.).

1. What is the patient safety behavior of interest? [Note: patient safety behavior can refer to changes in individual, organizational, or policy-level actions.]

2. Who are all the players responsible for this behavior?

a. Who is responsible in the organization for establishing the standards and clear expectations regarding this particular patient safety behavior? For example, is this PSP something that is driven primarily from upper levels of the administration, or is it something that is championed primarily at the clinic level?

b. What role does each player have in ensuring that the patient is not harmed? For example, in PSPs that involve information flow: Who is responsible for generating information that may or may not harm the patient? Who acts as a conduit for passing along such information? Who is responsible for ensuring that such information is accurate and remains accurate throughout the process? Note that some players may have important roles for each of these activities. The roles are somewhat different (and therefore have to be measured differently) for PSPs that are more behaviorally focused, such as hand-washing. Here we want to know: Who is engaged in the behavior of interest, and who is responsible for monitoring (through direct or indirect means) that such appropriate behaviors are indeed being carried out?

c. How are these players affected (if at all) by the intervention? This should include the players who are directly involved as part of the intervention; others that the players, in turn, are expected to influence; and the players who may be affected inadvertently. Take, for example, an intervention that provides guidance to nurses on how they can help monitor doctors' hand-washing behavior. The intervention directly affects nurses, and intentionally, but indirectly, affects doctors. At the same time, the intervention may inadvertently affect nurse's aides (in a positive manner) or may inadvertently affect the work load of administrators should tensions between nurses and doctors increase.

d. What consequences are there for the players if they do not adequately perform their expected role vis-à-vis the patient safety behavior?

e. Where are each of these players located in terms of the organizational structure?

f. How does their performance on this patient safety issue affect others in their unit, division, organization?

3. Who are the players responsible for initiating and carrying out the intervention? In any description of an intervention it is important to note explicitly who is doing what to whom. Some PSP interventions are commissioned by administrators and implemented by outsiders. Other interventions are championed, initiated, and carried out by insiders, and there are many combinations in between.

a. Where are the initiators and implementers located in the organization?

b. What role do they have in the organization?

c. What motivates them to participate?

d. To what degree do they motivate others?

4. In what unit within the organizational structure is the intervention located?

a. How important is this patient safety issue to the leadership of this unit? There are three fundamental ways to measure how important a PSP is to a unit. First, we can ask units to compare this PSP directly to other issues they may be addressing. The second and third approaches are a bit more indirect but much more empirically grounded. Here we can

describe the incentives and disincentives of high and low performance for the unit itself, as well as the incentives and disincentives the unit imposes on its members. For example, we could ask directly if the unit has an explicit list of priorities it wants to address and if so, where does this PSP fall on that list? More indirectly, we can first ask how (if at all) the unit is incentivized to report and improve their performance on this PSP? For instance, is the unit required to monitor their performance and report the results to upper-level administrators? Is the unit's performance compared with other units? Is the unit rewarded in any way for improvement? We can also ask how (if at all) the unit tries to monitor or incentivize its members to improve or achieve high performance. For instance, how regularly are unit members monitored? Does PSP performance affect a unit members' career, salary, status, or prestige in any way?

b. What (if any) consequences are there for the unit as a result of the success or failure of the intervention? Here, we want to know more about the stakes that surround the intervention itself. For example, is this an intervention being watched by trustees and hospital administrators? Is this an intervention that uses scarce resources that was chosen over other important priorities such that failure might breed ill will? Or is this one of multiple interventions being tried to improve PSP within the unit or hospital?

c. What kind of resources (e.g., financial, labor, etc.), if any, has the unit contributed? For example, for interventions focused on training and education, it would be useful to know how much time was spent by the trainers in preparing and presenting the materials, and how much staff time was required by the trainees? Further, was this staff time part of the regular work cycle, or was it considered extra work? If the latter, was this time compensated in any way? For interventions that require the acquisition of any new equipment or materials (e.g., sinks for washing hands, video cameras for monitoring performance, carts for wheeling around equipment and supplies), it would be useful to know the initial costs for purchasing and installing such equipment, as well as the cost of maintaining the equipment over time.

After outlining the key components of an intervention's context, the next step is to ask how each component should be reported or measured, and to what degree these reports or measures should be standardized.

Although it is clear that having standardized, close-ended instruments would facilitate comparisons across cases and therefore make it easier to conduct meta-analyses, currently there are few (if any) validated instruments for measuring specific context components. Also, extreme caution is warranted to ensure that whatever standardized instruments are ultimately selected, they can be appropriately applied across the full range of PSP contexts. Taking instruments that have been developed for other purposes and simply applying them to intervention contexts is a risky venture. Picking inappropriate instruments (e.g., ones that are overly simplistic or complicated, lack face validity, are unreliable, or fail to capture the full range of issues) will make it more difficult (not easier) for researchers and decisionmakers alike to fully understand the context in which an intervention occurred.

The most practical way to standardize context is to use a staged approach that moves from an exploratory, open-ended approach of reporting context components, to a more systemized and close-ended approach to reporting the context.

The first stage would standardize what particular context components were to be reported but not standardized specific instruments for how a particular content component was to be measured. In this stage, investigators would be presented with a list of key context components (such as the questions above) and asked to provide a description that is as detailed and honest as possible for each. Here, researchers would describe each context component in their own words, drawing on specific examples as appropriate. For example, consider the question "How important is this patient safety issue to the leadership of this unit (or units)?" We could imagine one case reporting that the intervention was one of the key PSP projects championed by the unit's leadership—one they held up as a quality improvement example and one in which the director of the unit was personally involved and engaged. We could imagine another case reporting that intervention was recognized by unit leadership as one of many quality improvement practices being implemented, and that the unit leadership took notice and provided additional support once it became apparent that the intervention was generating noticeable results.

By examining such descriptions across a range of different settings and different kinds of PSP interventions, we would begin to understand the range of ways in which "importance to leadership" could be potentially measured as a context effect. When we can combine these empirical results with what is found in the literature on leadership effects, we could begin to develop pilot, close-ended instruments for measuring different aspects of context.

The second stage would then add such pilot instruments to the open-ended questions used in the first stage. Combining the open- and closed-end reporting styles would allow us to see to what degree the closed-end instruments capture the important nuances of the context, as well as to test their reliability across contexts. Only after the new closed-end instruments could be shown to be valid and reliable, should they be used exclusively as the standardized instrument for a particular context component.

Chapter 11. Description of Ideal Evaluation Methods: Assessing for Possible Harms

All interventions can cause harms. Often these harms are unexpected adverse events. The methods of detecting unexpected adverse events are much less well developed than are the methods for detecting benefits, even for traditional assessments of efficacy or effectiveness of pharmaceuticals or devices. Long after having been studied for efficacy and receiving approval for use, unexpected associations have been discovered between COX-2 inhibitors and myocardial infarction, atypical antipsychotics and death in elderly patients, certain bisphosphonates and jaw osteonecrosis, and drug-eluting coronary stents and late restenosis. Therefore, our TEP judged an assessment of possible unintended adverse events to be an important criterion for an ideal evaluation of a PSP. They offered the following suggestions for how this might be accomplished:

1. Before PSP implementation, spending designated time with the developers, the implementers, and organizational staff to brainstorm about what could go wrong is key (this exercise is sometimes known as a "pre-mortem" discussion). This can include a careful "walk through" of the logic model, assessing where problems may arise. More formal methods for prospective risk assessment could be used. For instance, failure modes and effects analysis has been used to anticipate changes (including what could go wrong) related to technology implementation.[1-2] For potential adverse events foreseen by this process, evaluators can incorporate into the evaluation measurements of these potential harmful processes and outcomes and then include such information in their quantitative analysis.

2. For adverse events that are truly unexpected, meaning that no amount of pre-implementation planning could account for them, direct interviews with organizational staff responsible for implementing the PSP would be one way to assess for their occurrence. Educators can then retrospectively ask questions such as, “How did this really work?” or “Did anything go wrong?” and then follow up appropriately. Another method is for the implementers to keep a log or diary of issues related to implementation and use of the PSP. The list of issues can then be analyzed to identify those issues that were anticipated and those that were not.

References for Chapter 11

1. Wetterneck TB, Skibinski KA, Roberts TL, et al. Using failure modes and effects analysis to plan implementation of smart intravenous pump technology. *Am J Health Syst Pharm*. 2006; 63:1528-38.
2. Bonnabry P, Despont-Gros C, Grauser D, et al. A risk analysis method to evaluate the impact of a computerized provider order entry system on patient safety. *JAMIA*. 2008; 15(4):453-60.

Chapter 12. Description of Ideal Evaluation Methods: Quantitative Approaches to Context Heterogeneity

Special contribution from Naihua Duan, Ph.D., Columbia University, New York, NY

Introduction

Context often moderates intervention effectiveness; i.e., the effectiveness of an intervention might vary from site to site, depending on the contextual factors present at each site.[1,2] This phenomenon is what we have termed context heterogeneity. This moderation effect is usually formulated statistically through the "intervention × context" interaction:

$$(1a) \quad Yi = b_0 + b_1 \times Ti + b_2 \times Ci + b_{12} \times Ti \times Ci + \varepsilon i,$$

where i denotes the unit of analysis (usually the various sites in the study, but it can also be dyads of sites in matched comparisons), Yi denotes the outcome measure, Ti denotes the intervention status (Ti=1 for intervention, Ti=0 for control), Ci denotes the contextual factor, Ti × Ci denotes the "intervention × context" interaction, εi denotes random error, b_0 denotes the intercept for the model, b_1 denotes the main effect for the intervention, b_2 denotes the main effect for the contextual factor, and b_{12} denotes the moderation effect for the contextual factor, i.e., the influence of the contextual factor on intervention effectiveness.

As an example, consider a dichotomous contextual factor, say, C=1 denotes a teaching hospital and C=0 denotes a non-teaching hospital. According to model (1a), the intervention effect for a non-teaching hospital is given by b_1, while the intervention effect for a teaching hospital is given by $b_1 + b_{12}$. If the moderation effect is absent ($b_{12}=0$), the intervention effect does not vary between teaching and non-teaching hospitals. If the moderation effect is present ($b_{12}\neq 0$), the intervention effect does vary between teaching and non-teaching hospitals, the difference being the moderation effect b_{12}.

Model (1a) presents the "intervention × context" interaction" for a single contextual factor. The model can be generalized in a straightforward manner to accommodate multiple contextual factors:

$$(1b) \quad Yi = b_0 + b_1 \times Ti + b_2 \times C_{2i} + b_3 \times C_{3i} + \ldots + b_k \times C_{ki} + b_{12} \times Ti \times C_{2i} + b_{13} \times Ti \times C_{3i} + \ldots + b_{1k} \times Ti \times C_{ki} + \varepsilon i,$$

where C_{2i}, C_{3i},…, C_{ki} denote multiple contextual factors, and b_{12}, b_{13},…, b_{1k} denote the respective moderation effects.

The assessment of the moderation effect depends on the methodology used to assess the intervention effect. We discuss here eight scenarios.

Pre-Post Comparisons

One option that could be used to assess the intervention effect is to compare the outcome measures pre- and post-intervention, without concurrent control sites. Under the assumption that the outcome measure is stable over time if no intervention were provided, any change observed would be attributed to the effect of the intervention. In particular, Model (1a) takes the following form:

(2a) $Y_{0i} = b_0 + b_2 \times C_i + \varepsilon_{0i}$,

(2b) $Y_{1i} = b_0 + b_1 + b_2 \times C_i + b_{12} \times C_i + \varepsilon_{1i}$,

(2c) $Y_{1i} - Y_{0i} = b_1 + b_{12} \times C_i + (\varepsilon_{1i} - \varepsilon_{0i})$,

where the subscript $_i$ denotes the i-th site in the study, Y_{0i} denotes the pre-intervention outcome measure at the i-th site, Y_{1i} denotes the post-intervention outcome measure at the i-th site; C_i denotes the contextual factor at the i-th site; ε_{0i} and ε_{1i} denote the respective error terms for the pre- and post-intervention outcome measures. Compared to Model (1a), the terms $b_1 \times T$ and $b_{12} \times T \times C$ are absent in submodel (2a) because the intervention status T assumes the value T=0 under the control condition. Similarly, in submodel (2b) compared to Model (1a), the intervention status T assumes the value T=1 under the intervention condition, therefore the terms $b_1 \times T$ and $b_{12} \times T \times C$ are given as b_1 and $b_{12} \times C_i$. Submodel (2c) compares submodels (2a) and (2b): the term $Y_{1i} - Y_{0i}$ denotes the pre-post change, which measures the intervention effect at the i-th site.

The moderation effect (b_{12} in Model (2c)) can be assessed by regressing the intervention effect at the i-th site, $Y_{1i} - Y_{0i}$, on the contextual factor C_i in model (2c). For continuous contextual factors, this regression analysis estimates the rate of change for the intervention effect at the i-th site, $Y_{1i} - Y_{0i}$, when the contextual factor C_i changes by one unit. For dichotomous contextual factors, this regression analysis simplified to a two-sample comparison, comparing the average of the intervention effect among sites with the contextual factor C_i=1 (such as teaching hospitals) versus the average of the intervention effect among sites with the contextual factor C_i=0 (such as non-teaching hospitals).

The validity of pre-post comparisons depends on the validity of the assumption that the outcome measure is stable over time. If this assumption is questionable, e.g., if there is a possibility of a secular trend in the outcome measures, the validity of pre-post comparisons is questionable both for the assessment of intervention effect per se, and for the assessment of the effect of moderation for the contextual factors.

Longitudinal Comparisons

An important extension of pre-post comparisons is longitudinal comparisons of repeated measurements of the outcome measures over time, without concurrent control sites. Under the assumption that the outcome measure is stable over time if no intervention were provided, any

change over time that is observed would be attributed to the effect of the intervention. In particular, Model (1a) takes the following form:

(3a) $Y_{0i} = b_0 + b_2 \times C_i + \varepsilon_{0i}$,

(3b) $Y_{ti} = b_0 + b_1 \times t + b_2 \times C_i + b_{12} \times C_i \times t + \varepsilon_{ti}$,

(3c) $R_i = b_1 + b_{12} \times C_i + \delta_i$,

where the subscript $_i$ denotes the i-th site in the study, Y_{0i} denotes the pre-intervention outcome measure at the i-th site, Y_{ti} denotes the outcome measure at time t for the i-th site; R_i denotes the rate of change for the outcome measure for the i-th site; C_i denotes the contextual factor at the i-th site; ε_{0i} and ε_{ti} denote the respective error terms for the outcome measures; δ_i denotes the error term for the rate of change. We assume here that the trajectory of the outcome measure is linear over time, therefore the influence of time on the outcome measure in Model (3b) can be expressed as linear functions in time, t. Furthermore, the linearity assumption allows us to summarize the trajectory using the rate of change, R, in Model (3c): the rate of change for the i-th site, R_i, can be estimated by regressing the outcome measures, Y_{ti}'s, on time, t, within the i-th site. It is of course possible to extend the model beyond linear trajectories and allow non-linear trajectories.

In Model (3c), b_1 measures the intervention effect for sites with null values for the contextual factor, such as non-teaching hospitals; for these sites, the outcome measures improve at the rate of b_1 per unit time. For sites with C=1, say, teaching hospitals, the intervention effect is $b_1 + b_{12}$ – the term b_{12} measures the moderation effect for the contextual factor C.

The moderation effect (b_{12} in Model (3c)) can be assessed by regressing the rate of change, R_i, on the contextual factor C_i in model (3c).

The validity of longitudinal comparisons depends on the validity of the assumption that the outcome measure is stable over time. If this assumption is questionable, e.g., if there is a possibility of a secular trend in the outcome measures, the validity of longitudinal comparisons is questionable both for the assessment of intervention effect per se, and for the assessment of the effect of moderation for the contextual factors.

Matched Comparisons for Post-intervention Outcome Measures

Another strategy that could be used to assess the intervention effect is to include concurrent control sites matched individually to the intervention sites and compare the post-intervention outcome measures across sites.[1] We assume that the sites are matched on the contextual factor. Under the assumption that the matched sites differ only in the intervention status, the intervention effect can be assessed by the difference in the post-intervention outcome measures

[1] This strategy can be combined with pre-post comparisons, to be discussed in the following section. For now we assume that the pre-intervention outcome measures are not available.

for each dyad of matched sites. In particular, Model (1a) takes the following form under this approach:

(4a) $Y_{0i} = b_0 + b_2 \times C_i + \varepsilon_{0i}$,

(4b) $Y_{1i} = b_0 + b_1 + b_2 \times C_i + b_{12} \times C_i + \varepsilon_{1i}$,

(4c) $Y_{1i} - Y_{0i} = b_1 + b_{12} \times C_i + (\varepsilon_{1i} - \varepsilon_{0i})$,

where the subscript $_i$ denotes the i-th dyad of matched sites in the study, Y_{0i} denotes the post-intervention outcome measure at the control site in the i-th dyad, Y_{1i} denotes the post-intervention outcome measure at the intervention site in the i-th dyad; C_i denotes the contextual factor for both sites in the i-th dyads; ε_{0i} and ε_{1i} denote the respective error terms for the post-intervention outcome measures. Compared to Model (1a), the terms $b_1 \times T$ and $b_{12} \times T \times C$ are absent in submodel (4a) because the intervention status T assumes the value T=0 for the control site. Similarly, in submodel (4b) compared to Model (1a), the intervention status T assumes the value T=1 for the intervention site, therefore the terms $b_1 \times T$ and $b_{12} \times T \times C$ are given as b_1 and $b_{12} \times C_i$. Submodel (4c) compares submodels (4a) and (4b): the term $Y_{1i} - Y_{0i}$ denotes the difference between the intervention and control sites in the i-th dyad, which measures the intervention effect in the i-th dyad.

The moderation effect (b_{12} in Model (4c)) can be assessed by regressing the intervention effect in the i-th dyad, $Y_{1i} - Y_{0i}$, on the contextual factor C_i in model (4c). For continuous contextual factors, this regression analysis estimates the rate of change for the intervention effect in the i-th dyad, $Y_{1i} - Y_{0i}$, when the contextual factor C_i changes by one unit. For dichotomous contextual factors, this regression analysis simplified to a two sample comparison, comparing the average of the intervention effect among dyads with the contextual factor C_i=1 (such as teaching hospitals) versus the average of the intervention effect among dyads with the contextual factor C_i=0 (such as non-teaching hospitals).

The validity of matched comparisons depends on the validity of the assumption that the matched sites differ only in the intervention status. If this assumption is questionable, e.g., if there are important prognostic factors that differ between matched sites in the same dyad, the validity of matched comparisons is questionable, both for the assessment of intervention effect per se, and for the assessment of the effect of moderation for the contextual factors. If the unmatched prognostic factors are observed, it is possible to adjust for them using analysis of covariance (ANCOVA) models for post-intervention outcomes, or propensity scores analyses, to be discussed in Section F below.

Matched Comparisons for Pre-Post-Intervention Changes in Outcome Measures

Another strategy that can be used to assess the intervention effect is to combine strategies (A) and (B) and obtain both pre- and post-intervention outcomes measures for both intervention and control sites. The pre-post changes in outcome measures are compared across intervention and control sites to assess intervention effects. By combining pre-post and matched site comparisons,

this approach can be applied under weaker assumptions than the assumptions required for either strategy discussed in Sections A and B. In particular, this combined approach no longer requires the rather strong assumption in Section A that there is no secular trend. Instead, this approach only requires that any secular trend that might be present be the same between intervention and control sites in the same dyad. This assumption is also weaker than the rather strong assumption in Section B that the matched sites in the same dyad differ only in the intervention status. Instead, this approach allows the sites to differ in their pre-intervention status as long as these differences do not affect the pre-post change. In particular, Model (1a) takes the following form under this approach:

(5a) $D_{0i} = b_0 + b_2 \times C_i + \varepsilon_{0i}$,

(5b) $D_{1i} = b_0 + b_1 + b_2 \times C_i + b_{12} \times C_i + \varepsilon_{1i}$,

(5c) $D_{1i} - D_{0i} = b_1 + b_{12} \times C_i + (\varepsilon_{1i} - \varepsilon_{0i})$,

where the subscript $_i$ denotes the i-th dyad of matched sites in the study, C_i denotes the contextual factor for both sites in the i-th dyads. In submodel (5a), D_{0i} denotes the pre-post change in the outcome measure at the control site in the i-th dyad, which measures the secular trend in the i-th dyad. Here we allow the secular trend to depend on the contextual factor. In submodel (5b), D_{1i} denotes the pre-post change in the outcome measure at the intervention site in the i-th dyad. Submodel (5c) compares submodels (5a) and (5b): the term $D_{1i} - D_{0i}$ denotes the difference in the pre-post change between the intervention and control sites in the i-th dyad, which measures the intervention effect in the i-th dyad.

The moderation effect (b_{12} in Model (5c)) can be assessed by regressing the intervention effect in the i-th dyad, $D_{1i} - D_{0i}$, on the contextual factor C_i in model (5c). For continuous contextual factors, this regression analysis estimates the rate of change for the intervention effect in the i-th dyad, $D_{1i} - D_{0i}$, when the contextual factor C_i changes by one unit. For dichotomous contextual factors, this regression analysis simplified to a two-sample comparison, comparing the average of the intervention effect among dyads with the contextual factor C_i=1 (such as teaching hospitals) versus the average of the intervention effect among dyads with the contextual factor C_i=0 (such as non-teaching hospitals).

The validity of matched comparisons depends on the validity of the assumption that any secular trend that might be present be the same between intervention and control sites in the same dyad. If this assumption is questionable, the validity of matched comparisons is questionable both for the assessment of intervention effect per se, and for the assessment of the effect of moderation for the contextual factors. If the unmatched prognostic factors associated with the secular trend are observed, it is possible to adjust for them using analysis of covariance (ANCOVA) models for pre-post changes, or propensity scores analyses, to be discussed below.

Matched Comparisons for Longitudinal Rates of Change in Outcome Measures

A combined strategy similar to matched comparison of pre-post change, discussed in Section D above, is to combine strategies (A) and (C) and assess longitudinally the rate of change for outcomes measures for both intervention and control sites. The rate of change is compared across intervention and control sites to assess intervention effects. By combining longitudinal and matched sites comparisons, this approach can be applied under weaker assumptions than the assumptions required for either option discussed in Sections A and C. In particular, Model (1a) takes the following form under this approach:

(6a) $R_{0i} = b_0 + b_2 \times C_i + \delta_{0i}$,

(6b) $R_{1i} = b_0 + b_1 + b_2 \times C_i + b_{12} \times C_i + \delta_{1i}$,

(6c) $R_{1i} - R_{0i} = b_1 + b_{12} \times C_i + (\delta_{1i} - \delta_{0i})$,

where the subscript $_i$ denotes the i-th dyad of matched sites in the study, C_i denotes the contextual factor for both sites in the i-th dyads. In submodel (6a), R_{0i} denotes the longitudinal rate of change in the outcome measure at the control site in the i-th dyad, which measures the secular trend in the i-th dyad. Here we allow the secular trend to depend on the contextual factor. In submodel (5b), R_{1i} denotes the longitudinal change in the outcome measure at the intervention site in the i-th dyad. Submodel (6c) compares submodels (6a) and (6b): the term $R_{1i} - R_{0i}$ denotes the difference in the longitudinal rate of change between the intervention and control sites in the i-th dyad, which measures the intervention effect in the i-th dyad.

The moderation effect (b_{12} in Model (6c)) can be assessed by regressing the intervention effect in the i-th dyad, $R_{1i} - R_{0i}$, on the contextual factor C_i in model (6c).

The validity of matched comparisons depends on the validity of the assumption that any secular trend that might be present be the same between intervention and control sites in the same dyad. If this assumption is questionable, the validity of matched comparisons is questionable both for the assessment of intervention effect per se, and for the assessment of the effect of moderation for the contextual factors. If the unmatched prognostic factors associated with secular trend are observed, it is possible to adjust for them using analysis of covariance (ANCOVA) models for longitudinal rate of change, or propensity scores analyses, discussed below.

Adjusted Comparisons for Post-intervention Outcome Measures

Matched comparisons discussed in Sections (B)-(E) above assume that the intervention and control sites can be matched to the degree required under each strategy. In practical applications, this usually is not a realistic assumption. Therefore, adjustment for covariates is usually important, both for studies in which matching is attempted and for studies in which matching is not attempted. The adjustments can made either using the analysis of covariance (ANCOVA) model, or the propensity scores analysis.[3-6]

With ANCOVA, Model (1a) takes the following form:

(7a) $Y_i = b_0 + b_1 \times T_i + b_2 \times C_i + b_3 \times W_i + b_{12} \times T_i \times C_i + \varepsilon_i$,

where the subscript $_i$ denotes the i-th site in the study, Y_i denotes the post-intervention outcome measure for the i-th site, T_i denotes the intervention condition for the i-th site (T=1 if intervention, T=0 if control), C_i denotes the contextual factor for the i-th site, W_i denotes the covariates for the i-th site, and ε_i denotes the error term. The inclusion of the term $b_3 \times W_i$ adjusts for the imbalance in the covariates, W, that might be present between the intervention vs. control sites.

The coefficient b_1 denotes the intervention effect for sites with null values of the contextual factor (C=0), such as non-teaching hospitals; the intervention effect for sites with the value off the contextual factor C=1, such as teaching hospitals, is given by $b_1 + b_{12}$. The coefficient b12 denotes the moderation effect, e.g., how the intervention effect differs between teaching hospitals and non-teaching hospitals.

The moderation effect (b_{12} in Model (7)) can be assessed (along with the other coefficients in the model) by regressing the post-intervention outcome measure, Y_i, on the intervention status T_i, the contextual factor C_i, the covariates, W_i, and the interaction term, $T_i \times C_i$, in model (7).

With propensity score analysis, we first model the propensity for the i-th site to be an intervention site:

(7b) $\pi_i = \text{logit}(P(T_i=1)) = g0 + g1 \times C_i + g2 \times W_i$.

The propensity model (7) is usually specified and fitted as a logistic regression of intervention status (T) on the contextual factor (C) and covariates (W). The fitted model is then applied to all sites in the study to derive the propensity score, π, for each site to be an intervention site. The propensity scores can then be used in several ways to adjust for the imbalance between the intervention and control sites in the sample. One option that is particularly suitable for the assessment of the moderation effect for the context factor, C, is the following ANCOVA model that uses the propensity score π instead of the covariates W in model (7a):

(7c) $Y_i = b_0 + b_1 \times T_i + b_2 \times C_i + b_3 \times \pi_i + b_{12} \times T_i \times C_i + \varepsilon_i$.

Alternative ways that can be used to implement the propensity score analysis include matching, stratification, and weighting.

The validity of adjusted comparisons depends on the success of the adjustment to remove all imbalances between intervention and control sites in the sample. With either ANCOVA or propensity scores analysis, it is necessary to assume that all relevant covariates are observed, i.e., there are no hidden confounders.

Adjusted Comparisons for Pre-post-intervention Changes in Outcome Measures

Another strategy is to combine strategies (A) and (F), and apply ANCOVA or propensity score analysis to the pre-post changes in outcome measures.

With ANCOVA, Model (1a) takes the following form:

$$(8a) \quad D_i = b_0 + b_1 \times T_i + b_2 \times C_i + b_3 \times W_i + b_{12} \times T_i \times C_i + \varepsilon_i,$$

where D_i denotes the pre-post change for the i-th site. The rest of the model is identical to Model (7a) discussed in Section F above.

With propensity score analysis, the same propensity model (7b) is used to assess the propensity scores π_i. The fitted propensity scores are then used in the following model:

$$(8c) \quad D_i = b_0 + b_1 \times T_i + b_2 \times C_i + b_3 \times \pi_i + b_{12} \times T_i \times C_i + \varepsilon_i.$$

Adjusted Comparisons for Longitudinal Rates of Change in Outcome Measures

A combined strategy similar to adjusted comparison of pre-post change, discussed above, is to combine strategies (A) and (E) and compare longitudinally the rate of change for outcomes measures between intervention and control sites, adjusted for covariates that might be imbalanced, using either ANCOVA or propensity scores analysis.

With ANCOVA, model (1a) takes the following form:

$$(9a) \quad R_i = b_0 + b_1 \times T_i + b_2 \times C_i + b_3 \times W_i + b_{12} \times T_i \times C_i + \varepsilon_i,$$

where R_i denotes the pre-post change for the i-th site. The rest of the model is identical to Model (7a) discussed above.

With propensity score analysis, the same propensity model (7b) is used to assess the propensity scores π_i. The fitted propensity scores are then used in the following model:

$$(9c) \quad R_i = b_0 + b_1 \times T_i + b_2 \times C_i + b_3 \times \pi_i + b_{12} \times T_i \times C_i + \varepsilon_i.$$

The choice of analytic strategies for the assessment of intervention effect and the corresponding strategies for the assessment of the moderation effect for contextual factors depend on the design of the study. Strategies that are based on pre-post changes or longitudinal rates of changes can be applied only to studies that obtain pre-post measures or repeated measures of outcomes. Strategies that are based on matched comparisons can be applied only to studies designed with matched sites. In order to allow more flexibility in the analytic strategies, it would be advantageous to design the studies to include these features (either pre-post measures of outcome, or, more preferably, repeated measures; and matched sites).

The strategies discussed above are not exhaustive. Some of the strategies can be expanded, e.g., the linearity assumption in the longitudinal models can be relaxed to allow for non-linear trajectories over time. In addition, strategies such as instrumental variables analysis[7] and causal sensitivity analysis[6,8] can be used to address hidden bias, i.e., unobserved factors that are imbalanced between intervention and control sites. However, the eight strategies discussed above are probably the most practical methods and most commonly applied.

References for Chapter 12

1. Kravitz R, Duan N, Braslow JT. Evidence based medicine: Heterogeneity of treatment effects, and the trouble with averages. *Milbank Q*. 2004; 82(4):661–87. Erratum in: *Milbank Q*. 2006; 84(4):759-60.
2. Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: Implications for guidelines, payment, and quality assessment. *Am J Med*. 2007; 120(4 Suppl 1):S3-9.
3. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41-55.
4. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984; 79:516-24.
5. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997; 127(8 Pt 2):757-63.
6. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society, Series B. 1983b; 45:212-218.
7. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with Discussions). *J Am Stat Assoc*. 1996: 91(434):444-72.
8. Rosenbaum PR. Observational studies (2nd Edition). New York:Springer Verlag; 2002.

Chapter 13. Description of Ideal Evaluation Methods: Assessing the Strength of Evidence Across Studies of Patient Safety Practices

A key step when conducting a systematic review is assessing the strength of the evidence across the studies of a particular topic. An extended discussion of this is included in Appendix H.

One of the most widely used methods is that developed by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (www.gradeworkinggroup.org).[1] GRADE has tools for grading the quality of evidence and the strength of practice guideline recommendations. These tools or related ones are already in widespread use by the American College of Physicians, the British Medical Journal's *Clinical Evidence*, the Society of Critical Care Medicine, the Scottish Intercollegiate Guidelines Network, and more than 35 other organizations. An adaptation of GRADE has been published for diagnostic tests.[2]

AHRQ's Evidence-based Practice Center (EPC) program has developed its own method for assessing the strength of evidence, which started with GRADE but was adapted for the particular needs of the EPC program. The two methods share much in common, but they differ in the names they use for this construct as "quality of evidence" versus "strength of evidence" and in the labels and descriptors for the levels of evidence (Table 6). Also, GRADE suggests explicit weights for determining the level of evidence, while the EPC approach says that other methods, in addition to the GRADE weights, are acceptable as long as the method is transparent.

The rationale for developing an adaptation of GRADE or the AHRQ EPC system for patient safety practices (PSPs) is that there are issues about PSP interventions (as detailed in this report) that differ sufficiently from the kinds of interventions that the existing GRADE or EPC system are most commonly used for (drugs, surgery, etc.), such that a modification may be more relevant to stakeholders than trying to apply the existing GRADE or EPC criteria.

Table 6. Two methods for assessing the strength of evidence

GRADE	AHRQ EPC Program
High = Further research is very unlikely to change our confidence on the estimate of effect.	**High** = **High confidence that the evidence reflects the true effect**. Further research is very unlikely to change our confidence in the estimate of effect.
Moderate = Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.	**Moderate** = **Moderate confidence that the evidence reflects the true effect**. Further research may change our confidence in the estimate of effect and may change the estimate.
Low = Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.	**Low** = **Low confidence that the evidence reflects the true effect**. Further research is likely to change our confidence in the estimate of effect and is likely to change the estimate.
Very Low = Any estimate of effect is very uncertain.	**Insufficient =** Evidence either is unavailable or does not permit a conclusion.

In an adaptation for PSPs, we propose using descriptive categories similar to these. Using the GRADE and AHRQ EPC tools as a starting point, a tool to assess the strength of evidence across studies of PSPs might look like Table 7. This uses the EPC labels and the GRADE system of weights (+1, -1, etc.) and domains from both GRADE and the EPC schemes, plus adds key domains we identified during this project as relevant to evaluations of PSPs.

Table 7: Criteria for assigning strength of evidence

Does this evidence help me decide whether I can implement this PSP in my organization and get a similar result?

Type of evidence

Randomized trial = high

Decrease score if:

- No explanation of why the PSP might work, either in terms of theory, logic models, or prior success in other fields or in pilot studies (-1).
- No reporting of contexts, including at least structural organizational characteristics, external factors, patient safety culture, leadership, teamwork, or implementation tools (-1).
- PSP not described in sufficient detail to permit replication (-1).
- No reporting of the implementation process, assessment of unplanned events, or changes to workflow (-1).
- No assessment of the effect of contexts on implementation effectiveness (-1).

Observational study = low

Increase score if:

- Consistent results obtained in multiple studies (+2).
- Use of observational study designs of stronger internal validity (controlled before-and after, time series, statistical process control) (+1).
- Very strong effect (+1).
- Use of theory/logic models, assessment of contexts, reporting of implementation process, and fidelity of implementation (+1).

Any other evidence = insufficient

Across all study types, decrease score if:

- Serious (-1) or very serious (-2) limitation to study quality.
- Important inconsistency across studies (-1).
- Imprecise or sparse data (-1).
- High probability of reporting bias (-1).

This approach takes into account many of the points made by this project. For example, RCT evaluations about a PSP that lack reporting of theory, context, implementation, etc. decrease the strength of evidence to moderate or even low. Likewise, a body of evidence about a PSP that

comes entirely from studies that are not RCTs can be considered high quality evidence if the studies use observational designs of stronger internal validity (such as statistical process control or controlled before-and-after); if they inform theory and measure and report contexts; or if they have very strong effects or consistent results are obtained in many studies. Our suggestion here is preliminary and would benefit from refinement from a varied group of PSP stakeholders. For example, one concept not yet incorporated into this scheme that deserves discussion is the concept of proportionality, meaning that interventions that are low cost and low risk (e.g., hand-washing) may be accepted with a lower strength of evidence than interventions that have higher cost or risks (e.g., CPOE/DSS).

References for Chapter 13

1. Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions. Agency for Healthcare Research and Quality and the Effective Health Care Program. *J Clin Epidemiol*. 2009 Jul 10. [Epub ahead of print]. Also available as: Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions. In: Methods guide for comparative effectiveness reviews [posted July 2009]. Rockville, MD: Agency for Healthcare Research and Quality; 2009. Available at: http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60.
2. Schünemann HJ, Oxman AD, Brozek J, et al., GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Br Med J*. 2008; 336(7653):1106-10

Chapter 14. Results: Needs for Future Development

As described in our methods, we developed a list of 13 domains for future research. We then surveyed our technical expert panel (TEP) and queried which of these domains could be considered high priorities for future research. All 22 TEP members responded. Six domains received majority support as high priorities for future research (as bolded in Table 8).

Table 8. Priorities for Future Research

Domain	Is a high priority for future research	Not a high priority at this time	Skipped question
Developing/validating measures of leadership	10	11	1
Developing/validating measures of patient safety culture	13	9	0
Developing/validating measures of teamwork	10	11	1
Developing criteria/recommendations for what constitutes "reporting the intervention in sufficient detail that it can be replicated"	14	8	0
Understanding the important items to measure and report in implementation	13	9	0
Developing detailed methods for assessing potential unintended consequences	7	14	1
Developing evaluation methods for interventions that have very rare events as patient outcomes	6	15	1
Identifying critical aspects of context that influence PSPs not already considered in this project	5	15	2
Developing a theory-based taxonomy with which to describe and evaluate key elements of interventions, contexts, and targeted behaviors	12	9	1
Refining a framework for assessing the body of evidence about a PSP	12	8	2
Identifying barriers to conducting high quality research in patient safety	4	17	1
Conducting empirical research to show that the named factors from this project do make a difference	14	7	1
Evaluating the usefulness of another forum for communicating organizational experiences, other than peer-reviewed journals	9	12	1

We now expand on these survey results with a summary of the TEP discussion of these topics.

1. Developing and validating measures of patient safety culture. Discussion at the panel meetings indicated that several technical experts considered patient safety culture to be the overarching important construct. This view may explain why patient safety culture received majority support as a high priority for future research, whereas research on leadership and teamwork measures did not. Specific suggestions for future research included:

a. Developing validated measures of cultural adaptability to change.
b. Assessing the potential distinction between a culture of safety, a culture of excellence, and organizational culture.
c. Establishing connections between aspects of patient safety culture and patient outcomes or processes of care.
d. Assessing correlations between measures.

Additional comments that we received can be summarized as “we think teamwork and leadership are important," "several measures are currently available," and "the most important thing at this point is for people to use them so we can start building some evidence about this construct.”

2. Developing criteria and recommendations for what constitutes "reporting the intervention in sufficient detail that it can be replicated." More precise criteria for how PSP interventions should be described warrant additional research. In particular, the guidance described here, along with that provided by Standards for Quality Improvement Reporting Excellence (SQUIRE) and the National Quality Forum (NQF), needs to be evaluated. Doing so will help determine which PSP elements are necessary to describe in order to evaluate whether the PSP is truly effective. This also will help maximize the possibility of successful PSP replication with similar outcomes. Further research could also evaluate the effect of applying these draft criteria regarding PSP descriptions on the quality of PSP projects and published articles. Clearly, thoroughly describing PSPs also can help readers determine the relevance of an evaluation study to other PSPs or other contexts. For example, if a PSP requires an individual behavior change such as hand-washing, then knowing in detail what the intervention is may help readers of the study assess whether the given results are relevant only to hand-washing interventions or if they could be applied to other types of PSPs requiring individual behavior change. Knowing the details of the intervention also could help readers of the study determine how much the success of the PSP implementation depended on contextual issues (e.g., organization or teamwork).

3. Understanding the important items to measure and report on for implementation. Experts consider having comprehensive information about implementation key to being able to replicate a PSP. However, little empirical evidence exists about what makes a description of the PSP adequate for reporting. It is critical to assess what implementers need to know if they are to be able to implement or adapt an intervention in their own settings. Most experts considered "understanding the important items to measure and report on for implementation" to be related to or even the same as "reporting the intervention in sufficient detail that it can be replicated." This view suggests that the distinction between “the intervention” and “the implementation” may be an arbitrary line, and that ideal evaluations of PSP interventions need to consider the implementation as part of this intervention.

4. Developing a theory-based taxonomy or framework with which to describe and evaluate key elements of interventions, contexts, and targeted behaviors. Although the current project made a promising start on meeting this need, progress in this area will require additional development to produce a taxonomy that is both sufficiently broad based and flexible enough to be widely useful. Issues to be considered include whether a taxonomy is the preferable way to proceed, or whether a more useful strategy might be to create an explicit methodology that researchers could apply to specific problems and contexts. Yet another approach might be to devise an "assessment framework." Some experts sounded cautionary notes on this topic. They reported that outpatient PSP research may be too new to apply a taxonomy at this stage. They also reported that a single "unified" taxonomy may not be sufficiently flexible for diverse PSPs, and multiple taxonomies may be needed in any case. The countervailing view to these cautionary notes was that the field would not be well-served by having a proliferation of taxonomies. Instead, they reported, what is needed is a coherent, sufficiently comprehensive taxonomy that can accommodate the challenges of the subject.

5. Refining a framework for assessing the strength of a body of evidence. We did developmental work on an adaptation of the GRADE and EPC systems for assessing the strength of evidence across studies of a PSP. This work warrants further development.

6. Generating empirical evidence that the contextual factors identified in this project influence the success of the PSP. We acknowledge that most of the recommendations in the report have a thin empirical evidence base, which simply reflects the relatively immature state of research in this still relatively young field. Building a stronger evidence base will help future efforts at refining the recommendations presented here.

Additionally, the TEP acknowledged unintended consequences and rare events are important but not a high priority at this time. The issue of how to assess PSPs for context-sensitive beneficial outcomes in situations where the outcome is not rare was considered to be the more important priority.

Continuing the TEP's Efforts

We additionally queried the panelists about whether they would be willing to continue participating should AHRQ decide to continue this kind of methodological development work. All 22 panelists responded affirmatively, and many volunteered enthusiastic and laudatory comments about the process.

Of the list of "high priority" items, those we judge most likely to be fruitful for additional work using this TEP are:

- Developing a theory-based taxonomy with which to describe and evaluate key elements of interventions, contexts, and targeted behaviors.
- Developing criteria/recommendations for what constitutes " reporting the intervention in sufficient detail that it can be replicated" and the related topic of understanding the important items to measure and report in implementation.
- Refining a framework for assessing the body of evidence about a PSP.

All three of these topics were discussed as part of this project, and this report contains the results of those discussions. But each of these, alone, could have been the focus of an entire 2-day TEP meeting, rather than the 1 or 2 hour allotment they received. Often we found we had to “move on” from a rich and insightful discussion in order to try and discuss all of the topics on this project’s agenda. Additional development of the preliminary ideas is needed, and this TEP is both an appropriate group to pursue this and already engaged.

Chapter 15. Discussion

In this ambitious 1-year project we assembled a Technical Expert Panel (TEP) of patient safety experts, methods experts, and other stakeholders; met with the TEP three times; performed numerous literature reviews; conducted five Internet surveys; and achieved consensus on the following items:

1. The following five patient safety practices (PSPs) represent a diversity of important domains, including setting, regulation, target of the PSP (in terms of individual clinician versus organizational change), a more common versus a more rare patient safety event, among others:
   a. Checklist to prevent catheter-related bloodstream infection.
   b. The Universal Protocol to prevent wrong procedure, wrong site, wrong person surgery.
   c. Computerized order entry/decision support system.
   d. Medication reconciliation.
   e. Interventions to prevent in-facility falls.

   Interpretation and significance: Subsequent efforts examining PSPs, by AHRQ and others, may wish to use this diverse and representative list of PSPs to help focus their work.

2. Important evaluation questions for these PSPs are:
   a. What is the effectiveness of the PSP?
   b. What is the implementation experience of the PSP at individual institutions?
   c. What is the success of widespread adoption, spread, and sustainability of the PSP?

   Interpretation and significance: Evaluations of PSPs should explicitly consider these three questions. Journals should consider asking researchers to report on them separately. Also, implementers will want to assess their experience across all three questions.

3. High-priority contexts for assessing context-sensitive effectiveness at individual institutions are:
   a. Structural organizational characteristics (such as size, location, financial status, and existing quality and safety infrastructure).
   b. External factors (such as regulatory requirements, the presence in the external environment of payments or penalties such as pay-for-performance or public reporting, national patient safety campaigns or collaboratives, or local sentinel patient safety events).
   c. Patient safety culture (not to be confused with the larger organizational culture), teamwork, and leadership at the level of the unit.

d. Availability of implementation and management tools (such as staff education and training, presence of dedicated time for training, use of internal audit-and-feedback, presence of internal or external individuals responsible for the implementation, or degree of local tailoring of any intervention).

Interpretation and significance: Context is considered important in determining the outcomes of PSPs. The study investigators and the TEP judged these four domains as the most salient areas of context. This recommendation has broad implications for a variety of audiences. Researchers should be encouraged to measure and report on these contexts when describing a study of a PSP. Consumers of research will want to look for such reports, which will influence their interpretation of the study results and affect the applicability of the PSP to their setting. Accreditors and regulators should be reluctant to mandate adoption of a given PSP if it appears to be very dependent on context. In that case, they should also provide guidance on how that PSP might need to be modified depending on local contexts.

4. There is insufficient evidence and expert opinion to recommend particular measures for patient safety culture, teamwork, or leadership. Given the plethora of existing measurement tools we identified and reviewed, our recommendation is to use whichever method seems most appropriate for the particular PSP being evaluated.

   a. For patient safety culture, the measurement methods with the most support were the AHRQ Patient Safety Culture surveys, the Safety Climate Scale, and the related Safety Climate Survey.

   b. For teamwork, the most support was given to the ICU Nurse-Physician Questionnaire; no other measure received more than half the votes of respondents.

   c. For leadership, the measures receiving the most support were the ICU Nurse-Physician Questionnaire, the Leadership Practice Inventory, and the Practice Environment Scale.

   Interpretation and significance: Because the four areas of context described under Point 2, above, are judged highest priority, it will be crucial to develop and use valid measures of them in PSP studies. Researchers' use of common validated instruments would better enable readers to evaluate whether published results are applicable to their own settings. The state of the science here is immature, and funders and researchers are encouraged to continue to develop standard measures of the key domains of context.

5. The PSP field would advance by moving past considering studies of effectiveness as being "controlled trials" versus "observational studies." Although controlled trials offer greater control of sources of systematic error, they often are not feasible, either in terms of time or resources. Also, controlled trials often are not possible for PSPs because they require large-scale organizational change or PSPs targeted at very rare events. Hence, strong evidence about the effectiveness and comparative effectiveness of PSPs can be developed using designs other than randomized controlled trials. However, PSP evaluators are to be discouraged from drawing cause-and-effect conclusions from studies with a single pre- and post-intervention measure of outcome. More sophisticated designs (such as a time series or stepped-wedge design), are available and should be used when possible.

Interpretation and significance: Given the major push to improve patient safety and the focus on evidence-based practices (which are rapidly embedded in national standards such as those issued by the National Quality Forum, the Joint Commission, the Institute for Healthcare Improvement, and others), it will be crucial to develop standards for appropriate evaluations to answer key safety-oriented questions. The results above will help journal editors, funders, researchers, and implementers adopt robust study methods for PSPs, methods that most efficiently answer the key questions without undue bias.

6. Regardless of the study design chosen, criteria for reporting on the following items in a PSP evaluation are necessary, both for understanding how the PSP worked in the study site and whether it might work in other sites:
   a. An explicit description of the theory for the chosen intervention components and/or an explicit logic model for "why this PSP should work."
   b. A description of the PSP in sufficient detail that it can be replicated, including the expected change in staff roles.
   c. Measurement of contexts in the four domains described in Point 3, above.
   d. Details of the implementation process, what the actual effects were on staff roles, and how the implementation or the intervention changed over time.
   e. Assessment of the impact of the PSP on outcomes and possible unexpected effects. Including data on costs, when available, is desirable.
   f. For studies with multiple intervention sites, an assessment of the influence of context on intervention and implementation effectiveness (processes and clinical outcomes).

   Interpretation and significance: These criteria (items a-f) are deemed necessary for an understanding of PSP implementation and effectiveness and the degree to which these elements are sensitive to context. Future AHRQ-supported evaluations of PSP implementation should adhere to the criteria developed by this project. Only through repeated assessments and measurements will it be possible to determine the context-sensitivity of PSPs and to build the evidence base for which contexts are most important and how they should be measured and reported.

Limitations

The strengths of our work to arrive at these criteria include the broad-based expertise and viewpoints within the project team and the TEP, the grounding of our work in theory and the practical assessment of literature, and the careful and painstaking process of consensus-building, through formal and informal group judgment processes. Limitations of our work are mainly the limitations of the state of the science: there is no agreed-upon definition of what is "context," the boundaries between context and the intervention are often arbitrary, the intervention and the implementation of the intervention may often be considered to be a single construct, and there are insufficient data or expert opinion to specify in greater operational detail several of our important criteria, such as "description of the intervention in sufficient detail that it can be

replicated" or even what constitutes an adequate description of the use of theory. Furthermore, as already discussed, there is insufficient evidence and opinion to recommend specific measures for patient safety culture, teamwork, and leadership, even though these are three important contexts believed to influence intervention effectiveness. Lastly, our discussions were anchored by consideration of five specific PSPs. While they were chosen specifically to be diverse and representative, it is possible that contextual factors may be different for other PSPs. Our results could also benefit from a critical examination by an even wider-ranging group of patient safety stakeholders.