Finding order in heterogeneity: types of quality-improvement intervention publications

L V Rubenstein,1,2,4 S Hempel,2 M M Farmer,1 S M Asch,1,2,4 E M Yano,1,3 D Dougherty,5 P W Shekelle1,2,4

See Editorial, p 394

ABSTRACT

Background: Stakeholders in quality improvement agree on the need for augmenting and synthesising the scientific literature supporting it. The diversity of perspectives, approaches, and contexts critical to advancing quality improvement science, however, creates challenges. The paper explores the heterogeneity in clinical quality improvement intervention (QII) publications.

Methods: A preliminary classification framework was developed for QII articles, aiming for categories homogeneous enough to support coherent scientific discussion on QII reporting standards and facilitate systematic review. QII experts were asked to identify articles important to QII science. The framework was tested and revised by applying it to the article set. The final framework screened articles into (1) empirical literature on development and testing of QIIs; (2) QII stories, theories, and frameworks; (3) QII literature syntheses and meta-analyses; or (4) development and testing of QII-related tools. To achieve homogeneity, category (1) required division into (1a) development of QIIs; (1b) history, documentation, or description of QIIs; or (1c) success, effectiveness or impact of QIIs.

Results: By discussing unique issues and established standards relevant to each category, QII stakeholders can advance QII practice and science, including the scope and conduct of systematic literature reviews.

Quality of care in the United States is of uneven quality and often even unsafe.1 Clinical quality improvement intervention (QII) practitioners and evaluators seek to improve this situation by engaging organisations and their providers in developing and/or implementing better approaches to care delivery. In an era of cost containment, QIIs that aim to advance science as well as local care quality are urgently needed. Yet both publications, medicine’s mainstay for advancing science, and the synthesis of published work through systematic review have been problematic for the QII field. We aimed to develop a framework that recognised the diversity of types of articles essential to understanding and replicating QIIs, and that could be used for screening QII literature into more homogeneous categories for literature evaluation and synthesis.

An important challenge confronting stakeholders in promoting QII publication and synthesis has been the lack of agreed-upon standards for evaluating QII research. While agreement on standards is developing,2,3 there is as yet no method for systematically matching QII articles to relevant standards. In addition, there may be important types of QII articles that are excluded from most systematic reviews, yet could shed light on critical QII issues. To gain a comprehensive view of important article types, we asked experts to identify articles they considered important to advancing QII science. We then refined and tested our framework based on the resulting article set.

QIIs seek to improve the quality of care delivered by target organisations or organisational units such as practices, intensive care units, or community groups. QIIs can include, for example, changes in internal or external organisational policy, staffing, resources or any of the components of the chronic illness care model, such as links to external resources, decision support, informatics, care design or self-management support. In the work presented here, we focused on QIIs directed at improving evidence-based clinical practices such as those identified in guidelines. QIIs address both the method(s) for encouraging change (eg, continuous quality improvement, expert consultation or reimbursement changes) and the intended changes themselves. QIIs change structure or organisation in order to charge the process and outcomes of care. By nature, QIIs are carried out in diverse, often complex settings.4-7 Documentation and analysis of, for example, QII settings, encouragement methods, and direct and indirect outcomes may require a variety of approaches.

In addition, many QIIs employ theories, designs and methods that have not been extensively used in the clinical literature, with its focus on randomised controlled trials. Many of the approaches used are also new to classical health-services research, which relies heavily on epidemiologically based methodologies. Healthy development of the QII field may require studies focused on theories, terminology development, intervention development and hypothesis generation to a greater extent than traditional clinical or health services research. Recognising diverse types of studies and their goals, in addition to traditional hypothesis testing studies, may encourage more rapid, efficient progress toward effective QIIs. Our classification framework aims to highlight this diversity, while sorting QII publications into categories homogeneous enough to support scientific discussion of methodological and reporting standards potentially applicable to each category.

We report here on the development and preliminary testing of our classification framework. We assessed whether reviewers could apply the framework to screen articles suggested by experts into categories, and whether it comprehensively classified the full expert-generated article set.
METHODS

For this project, we approached experts with extensive exposure to scientific approaches to clinical quality improvement who were part of a scientific group for a research and evaluation design and methods conference in 2005.14 The conference was directed at a broad group of about 100 stakeholders including journal editors, researchers, funders, and lay and professional quality-improvement practitioners and advocates. The planning group experts were health services and public-health researchers, many of whom had specific programmatic responsibility for developing QII within their organisations, that is the Agency for Healthcare Quality and Research, the Centers for Disease Control, the Veterans Administration, the National Institutes for Health and the Robert Wood Johnson Foundation. We asked them to identify examples of publications important to advancing QII science. We sought publications involving either healthcare or community organisations or organisational units, and that concerned the implementation of consensus guidelines or evidence-based clinical practices for a population of patients.

Experts suggested articles during regularly scheduled conference planning calls, one face-to-face planning meeting, or by email over a 5-month period beginning in 9/04, and conducted a final review of our article list in 2/05. The core planning group initially included eight people from participating organisations but gradually expanded to 22 plus five additional intermittent consultants. Of the final 27 total, 12 either contributed articles or commented on the final article list.

We developed a classification framework using an iterative process. Our goal was to achieve categories that were sensitive (included a high proportion of the expert-identified articles), could be reliably applied by reviewers and had face validity as a starting-point for applying the right methodological standards to identified articles. Groups of two to five reviewers active in QII research (three senior physician health services researchers (LR, SA, FS), one senior PhD health services researcher (EY), and one junior PhD health services researcher (MF) scrutinised identified publications assessing similarities and differences between publications. The evolving categories were further discussed regarding their potential for specifying types of articles to which coherent methodological norms or standards could be applied. We applied categories through a decision tree operationalised as a standardised screening form, and sequentially improved the form through expanding cycles of review.

Two reviewers (LR and MF) independently applied the screening form to the full set of identified publications. Disagreements were resolved through consensus. As a quality check, three further reviewers independently subjected selected articles to the screening form and compared results. Two additional authors (SH and DD) then reviewed the final article set and categories to check final article assignment.

RESULTS

The set of 80 publications identified by the expert panel is documented in the Appendix (online). The iterative review process revealed these publication types: (I) Empirical literature on development and testing of QIIs; (II) QII stories, theories, and frameworks; (III) QII literature synthesis and meta-analysis; (IV) Development and testing of QII-related tools, and a miscellaneous category for books and unclassifiable articles. Table 1 provides an overview of the general type of publication, the underlying goal, a broad definition and the number of identified studies within the expert selection. The table also summarises the resulting methodological issues within the framework of QII research that each type of publication may represent or may be able to address.

Twenty-five articles15–39 reported data collected to develop or test QIIs (Category I). These studies aimed to improve quality of care provided by a specific organisational unit or units, practice or community for a defined population.

Thirty articles40–65 discussed theories, frameworks or “stories” relevant to QIIs (Category II). This type included articles that did not report on formal qualitative or quantitative data collection carried out for the purpose of the subject article, other than expert panel-type data, and that were not formal literature syntheses or meta-analyses.

Ten articles66–75 were formal literature syntheses or meta-analyses focused on QII research (Category III).

Nine articles76–84 dealt with QII tools (Category IV). Tools are tests or methods to assess and document any QII. These publications report data concerning the development, testing or refinement of QII methods or tools in naturalistic settings (eg, sensitivity, specificity, reliability, validity, usefulness, acceptability).

Four of the publications11–15 85 suggested by the experts were books on quality and evaluation methods, and could not be classified in the four main categories. Two publications86–88 focused on consent procedures and a Cochrane group description of non-randomised methods.

The review process revealed that Group I, the empirical literature on development and testing of QIIs, should be further distinguished into the subcategories: (Ia) Development of QIIs; (Ib) History, documentation, or description of QIIs; (ic) Success, effectiveness or impact of QIIs. Table 2 outlines the goals, definitions and number of identified studies, and summarises further methodological issues and conceptual considerations.

Eight of the 25 category I articles23 24 but not 22 focused on the preintervention phase and reported data collected for the purpose of developing QIIs (Category Ia).

Two articles focused on the postintervention phase (Category Ib) but did not report empirical data on the success, effectiveness or impact of a single QII study and were therefore deemed to represent a distinct subtype.25 24

Fifteen articles reported data on QII success, effectiveness or impacts (Category Ic).26–39 This subtype of publication included randomised trials as well as a variety of more naturalistic designs.

DISCUSSION

This project aimed to explore the heterogeneity of QII publications by determining what types of articles experts would identify as germane to the field of QII research. We discovered a rich and heterogeneous literature. Within this heterogeneity, however, we identified several relatively homogeneous types of publications identified by experts as important to clinical QII research. These types of publications covered nearly all suggested articles, had face validity and could be reliably identified by independent reviewers using a decision tree. This broad categorisation appears suitable for promoting coherent discussion around standards for conduct and publication directed at any one of the specific types, and as the basis for initial screening of articles for systematic reviews.

Three of the categories included articles with a focus on reporting data on QIIs, while one included articles without such a focus. The first type encompassed a diverse set of data-based articles on QIIs spanning all phases from development through evaluation. Within this category, a number of important sources of methodological heterogeneity can be identified that merit
**Developing research and practice**

### Table 1  Categories of quality-improvement intervention (QII) research

<table>
<thead>
<tr>
<th>General type</th>
<th>Goals</th>
<th>Definition</th>
<th>Number</th>
<th>Further issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Empirical literature on development and testing of QIIs</td>
<td>Assess barriers to quality improvement</td>
<td>Reports data collected to develop or test the QII</td>
<td>25–39</td>
<td>What subtypes among the diverse investigations in this category will be most useful?</td>
</tr>
<tr>
<td></td>
<td>Discover effective quality improvement or provider behaviour change methods</td>
<td>Aims to improve quality of care for a defined population of individuals cared for/targeted by a specific organisational unit or units, practice, or community</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Implement guidelines or best practices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II. QII stories, theories, and frameworks</td>
<td>Advance understanding or theory related to QII development or testing</td>
<td>No formal qualitative or quantitative data collection other than expert panel-related data</td>
<td>30^4–6^64–62</td>
<td>How should articles in this category be distinguished from each other? Should the field develop quality standards for this kind of research, such as assessing the extent to which information is validated, systematic or original?</td>
</tr>
<tr>
<td></td>
<td>Develop consensus</td>
<td>Includes one or more case descriptions, illustrative stories, theories or frameworks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III. QII literature synthesis and meta-analysis</td>
<td>Develop the evidence base on methods for quality improvement or provider behaviour change</td>
<td>Reports literature synthesis or meta-analysis data</td>
<td>10^14–23</td>
<td>Should standards for evaluating multicomponent interventions across studies be developed?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assesses the effectiveness or characteristics of a QII</td>
<td></td>
<td>Are current quality standards for literature synthesis and meta-analysis adequate for behaviour-change interventions?</td>
</tr>
<tr>
<td>IV. Development and testing of QII-related tools</td>
<td>Discover effective technologies for use in quality improvement</td>
<td>Reports data collected to develop or test a QII tool</td>
<td>9^14–62</td>
<td>Are standards for identifying naturalistic settings and conditions needed?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The method or tool is developed or tested in a naturalistic setting, under routine conditions</td>
<td></td>
<td>Are standards for assessing QII tool success needed?</td>
</tr>
</tbody>
</table>

See Appendix (online) for an annotated review list.

Further analysis and discussion. A challenging but typical aspect of publications in this category was that the full set of information about a particular study was often presented in multiple and methodologically diverse articles. Judgements of data-based QII article quality thus may require, for example, consideration of qualitative research standards and randomised trial standards as applicable across relevant article sets. Another challenging aspect of articles in this category was based on the multiyear, multiphase, evolutionary nature of many QIIs, resulting in different study goals at different phases of scientific learning.

We distinguished three subcategories of data-based QIIs that took account of different study goals. Knowledge based on each subcategory appeared crucial for understanding the QII’s context as required by realistic evaluation frameworks. Subtype la studies were concerned with the development of QIIs. These raised the issue of the role of a systematic development process as part of QII projects in general. Does a development process that begins with evidence review, plan–do–study–act cycles and systematic engagement of stakeholders in design produce a higher-quality QII? Subtype lb studies addressed the history, documentation or description of QIIs. While the other subtypes represented studies also commonly found in other fields, subtype lb studies appeared more distinct. These articles did not report new data on whether the QII achieved the intended changes in clinical process or outcomes, but reported on, for example, challenges or critical lessons learnt across QII stages or a series of QIIs.

### Table 2 Group I empirical studies subtypes

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Goals</th>
<th>Definition</th>
<th>Number</th>
<th>Further issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia. Development of QIIs</td>
<td>Create new QII approach</td>
<td>Reports data collected for the purpose of developing a QII</td>
<td>8^16–22</td>
<td>Should documentation of development process be a standard for all QIIs?</td>
</tr>
<tr>
<td></td>
<td>Locally adapt previously developed QIIs</td>
<td>No data on QII success, effectiveness, or impacts</td>
<td></td>
<td>Are there minimum standards for QII development documentation?</td>
</tr>
<tr>
<td>Ib. History, documentation, or description of QIIs</td>
<td>Enable replication or spread of QIIs</td>
<td>Reports data on the conduct of a QII</td>
<td>2^15–24</td>
<td>Can/should criteria for the validity and reliability of descriptive information on QIIs be developed?</td>
</tr>
<tr>
<td></td>
<td>Enable in-depth case analysis of implementation</td>
<td>No data on QII success, effectiveness, or impacts</td>
<td></td>
<td>Should evaluations of QIIs be required to reference or incorporate this kind of information?</td>
</tr>
<tr>
<td>Ic. QII success, effectiveness or impact</td>
<td>Determine effectiveness, effects, and predictors of success of QIIs</td>
<td>Reports qualitative or quantitative data on the success, effectiveness, or impacts of a QII</td>
<td>15^25–39</td>
<td>What kinds of studies are used to rigorously evaluate QIIs?</td>
</tr>
</tbody>
</table>

See Appendix (online) for an annotated review list.

QII, quality-improvement intervention.
Developing research and practice

no articles describing an individual QII were suggested by our experts, we envision this category as including articles that focus, for example, on data showing whether an intervention was implemented as planned. In addition to their importance for understanding QII evaluation results, such articles may set the stage for subsequent rigorous approaches such as preplanned meta-analyses across a set of implementation studies or other kinds of crosscutting research aimed at understanding the impacts of whole initiatives.

Type Ic studies evaluated the success, effectiveness or impact of QIIs—a type of article one would immediately think of as integral to QII research, yet one that represented only a minority among articles our experts considered to be important. In addition to reflecting the diversity of important work in the QII field, this may reflect a relative dearth of high-quality empirical publications evaluating QIIs. The lack of standard expectations for what should be included in evaluating and reporting on QIIs may make it particularly difficult for authors and reviewers to achieve common ground on papers, further diminishing the number of QIIs entering the literature. The many existing quality-improvement initiatives in clinical practice are thus not currently well reflected in the literature. Further exploration of studies in this category may yield additional insights about how researchers are currently handling the challenges inherent in QII research, and eventually greater consensus about which approaches fit which purposes. Although the verdict on suitable study designs for evaluating QII success, effectiveness or impact may still be out, many reporting and methodological standards are available.5 99 We used an article classification decision tree to assign articles to our categories. Existing standards could be applied to appropriate articles or article sections based on further development of the decision tree past the initial screening stage reported on here.

The second group (Category II) focused on articles that did not focus on data. These included stories, theories and frameworks related to QIIs. Clearly, theory is essential to the development of a deeper understanding of the mechanisms of action of QIIs, and hence to our ability to predict the usefulness of particular improvement approaches under particular conditions. Deming, for example, recognised the importance of taking account of “profound knowledge” such as psychological theories in quality improvement work.91 Our reasoning on understanding causality in QII work, where context is important, requires scrutiny.92 Theoretical development in the literature sample focused both on developing theories about QIIs and on applying existing social science theory to QIIs. Articles on frameworks addressed the underlying concepts used in implementing and evaluating QIIs and were used to generate, organise or promote hypotheses. Articles on stories provided a chronological description of what happened as part of implementing a QII, but were not based on systematic data collection. Learning theory and practical application demonstrate the power of stories in communicating about quality improvement.93–95 We did not attempt to subdivide Category II articles; further discussion of these articles, however, might have been fruitful. For example, articles that “tell the story” of a QII might be considered to be of a higher quality if they represented first-hand knowledge and/or were reviewed by multiple QII participants. Conceptual frameworks might be rated as of a higher quality if they have undergone expert panel review. In addition, subdivisions of Category II might consider, for example Tilly’s four categories of explanatory reasons (conventions, formulas, stories and technical accounts),96 or other theoretical frameworks. Focusing on this category of publications will encourage literature synthesis that includes rather than excludes important conceptual articles and that thus enables systematic development of the theoretical basis for quality improvement.

The third type of article focused on developing the QII evidence base through data from literature syntheses and meta-analyses. The importance of accumulating knowledge across studies about quality-improvement interventions is self-evident. Discussions of methodological issues in this category, such as how best to assess multicomponent and organisational interventions through meta-analysis, are likely to be fruitful. Methodological quality standards that can be applied by reviews dealing specifically with QIIs, such as those in progress through the SQUIRE group initiative,97 will be critical for developing this type of research.

Finally, the fourth category may seem surprising as a QII research category at first glance, as this category dealt with data on the development or testing of QII tools in routine care or other naturalistic settings. These articles often used classical methods of measure development such as verification, validation, user acceptability or human-factors testing.98 On further investigation, however, it is apparent that a focus on QII measurement is critical. Measures developed and tested only within a research context may not be valid or reliable when applied by usual personnel in actual care sites, and may be impractical or unfeasible. On the other hand, measures developed solely for the purposes of a QII may not have undergone testing at all, and may be highly misleading. Whatever their origins, the use of measures that have not been specifically tested in naturalistic settings can have serious negative organisational and clinical consequences in a QII context, resulting in erroneous clinical decisions, excessive costs or invalid evaluation. Quality standards for this type of article would involve those related to instrument development, but with attention to issues related to the QII context. Measurement work in the context of QII development and evaluation should be a focus of further discussion and investigation.

This study has limitations. First, the article sample, while representative of the opinions of a diverse group of experts, was essentially a convenience sample, albeit larger than samples used in some prior published studies using similar methodology.5 97 Readers should therefore not infer that, for example, the proportions of articles in each category are representative of the full QII literature. Second, this article focuses on initial classification of articles into broad categories. A variety of social science34 35 and QII34 54 56 58 97 theorists have developed frameworks that may prove fruitful in understanding and improving QII research. In our initial empirically based article screening strategy, however, these theories, as well as other study dimensions such as outcomes would be considered at a later stage of review. Finally, this study sought articles whose goals were to produce organisational or structural change either in healthcare settings or in the community. While this aim excludes the large number of studies whose goal is to test the efficacy or effectiveness of an intervention independently of organisational context, we see this focus as productive in identifying articles directly aimed at improving the quality of routine healthcare and health promotion.

In conclusion, we found both heterogeneity and coherence in the QII literature we reviewed. By considering each important type of literature separately, and building on existing investigative approaches to understand and refine methods, QII