Identifying quality improvement intervention evaluations: is consensus achievable?

M S Danz,[1,2] L V Rubenstein,[1,2] S Hempel,[1] R Foy,[3] M Suttorp,[1] M M Farmer,[2] P G Shekelle[1,2]

[1]RAND Corporation, Santa Monica, California, USA
[2]Veterans Affairs Greater Los Angeles Healthcare System, North Hills, California, USA
[3]Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

Correspondence to
Dr M S Danz, RAND Corporation, Santa Monica, CA 90407, USA;
mjsdanz@gmail.com

Accepted 18 January 2010
Published Online First 14 June 2010

ABSTRACT

Background The diversity of quality improvement interventions (QIIs) has impeded the use of evidence review to advance quality improvement activities. An agreed-upon framework for identifying QII articles would facilitate evidence review and consensus around best practices.

Aim To adapt and test evidence review methods for identifying empirical QII evaluations that would be suitable for assessing QII effectiveness, impact or success.

Design Literature search with measurement of multilevel inter-rater agreement and review of disagreement.

Methods Ten journals (2005-2007) were searched electronically and the output was screened based on title and abstract. Three pairs of reviewers then independently rated 22 articles, randomly selected from the screened list. Kappa statistics and percentage agreement were assessed. 12 stakeholders in quality improvement, including QII experts and journal editors, rated and discussed publications about which reviewers disagreed.

Results The level of agreement among reviewers for identifying empirical evaluations of QII development, implementation or results was 73% (with a paradoxically low kappa of 0.041). Discussion by raters and stakeholders regarding how to improve agreement focused on three controversial article selection issues: no data on patient health, provider behaviour or process of care outcomes; no evidence for adaptation of an intervention to a local context; and a design using only observational methods, as correlational analyses, with no comparison group.

Conclusion The level of reviewer agreement was only moderate. Reliable identification of relevant articles is an initial step in assessing published evidence. Advancement in quality improvement will depend on the theory- and consensus-based development and testing of a generalizable framework for identifying QII evaluations.

International interest in learning about how best to improve quality of care is growing. This growth is occurring in tandem with urgent demands to improve the everyday care delivered by healthcare organisations. As a result, large numbers of quality improvement interventions (QIIs) are being carried out by these organisations, often with significant use of organisational resources.[1] The methodological approaches and outcomes of these QIIs are highly variable.

Evidence reviews of published scientific literature have been at the core of effectiveness and comparative effectiveness assessment in other areas of healthcare. The diversity of approaches to carrying out, evaluating and publishing on QIIs, however, has impeded the usefulness of evidence review for advancing the effectiveness of quality improvement activities. An initial step toward improving QII evidence review capabilities is the development of consensus on approaches for identifying and classifying relevant QII studies. In the absence of such approaches, literature searches and syntheses may yield haphazard results, and consensus around QII best practices will remain difficult to achieve.

Our study aimed to adapt and test evidence review methods for reliably identifying empirical QII evaluations that would be suitable for assessing QII effectiveness, impact or success. This paper describes and evaluates application of an electronic search strategy, primary title and abstract screening, and secondary screening based on a full text review to identify these articles. We examined the reliability with which we could identify empirical QII evaluations (ie, those reporting on development, implementation, or outcomes of a QII). We then analysed the studies that generated disagreement among reviewers in detail (including assessment by experts from the USA and the UK) and conceptualised a strategy to improve inter-rater agreement in identifying empirical QII evaluations suitable for assessing QII effectiveness, impact or success.

METHODS

We used standard evidence review strategies to search electronically for QII publications, to carry out initial title and abstract screening for relevance, and to review complete articles for final inclusion as QII evaluations. Our group is composed of physicians (LVR, PGS, RF, MSD) and health services researchers (SH, MMF, MS) with expertise in quality improvement and evidence synthesis. (See the appendix for a summary of the article review process.)

Electronic search

We applied a simple and inclusive text word search strategy in PubMed to identify QII studies published in 10 core journals over 3 years (2005–2007): five principal general medical interest journals (Annals of Internal Medicine, BMJ, JAMA, Lancet and New England Journal of Medicine) and five key specialty journals (American Journal of Managed Care, Health Services Research, Joint Commission on Quality and Patient Safety, Medical Care and Quality & Safety in Health Care). This yielded 183 publications. In a related study, this search strategy had sufficient sensitivity to identify 43% of articles that a panel of experts had

considered important to the field of quality improvement (Hempel *et al*, in preparation). Although the sensitivity of the electronic search strategy was not ideal, it provided an unbiased and realistic sample of articles.

Primary title and abstract search

Two reviewers (LVR, PGS) screened titles and abstracts of the search output to select those articles reporting on empirical studies on the development, implementation, or impact of a QII.[2] We included articles selected by either reviewer as potentially relevant (74, or 40% of the 183 publications).

Secondary full article screen and reliability testing

We developed a working definition of a QII (figure 1) based on prior work.[2–8] We used our definition as the basis for a secondary screening tool with guidelines for use. Reviewers applied the secondary screening tool, through a full text review, to a random sample of the 74 publications identified through the primary title and abstract search.

Reliability of the secondary screen

Six reviewers (authors MSD, LVR, SH, RF, MMF and PGS) worked in pairs, comprising three teams of two reviewers each. Physicians, quality improvement experts and experienced systematic reviewers were represented in each pairing. Each reviewer independently applied the secondary screener to 22 randomly selected articles from among the 74 identified as potentially relevant based on title and abstract review. The two reviewers in each of the three reviewer pairs then compared their assessments and resolved any disagreements with respect to identifying QII evaluations. Reliability analyses compared the three resolved sets of ratings.

Analysis of disagreements

We identified the articles generating differences of opinion among reviewer pairs. We then surveyed an expert panel of 12 stakeholders, including QII experts and journal editors, on whether the identified articles represented QII evaluations that were suitable for evidence review on QII effectiveness, impacts or success (figure 2). The survey briefly described each article and provided the stakeholders with the following five-point rating scale: definitely (5), probably (4), no preference (3), probably not (2), definitely not (1). Panelists subsequently discussed their ratings as a group. As a final step, study investigators qualitatively identified the issues underlying disagreements.

Working Definition for QII

QII: "An effort to change/improve the clinical structure, process, and/or outcomes of care by means of an organizational or structural change."

- **Structure of care**: Context within which care processes are delivered
  Context includes, for example, internal policies, standard operating procedures, resource use, inter-professional collaboration, educational requirements, or aspects of the chronic illness care model (decision support, clinical information systems, delivery system design, self-management support, or use of community resources).
- **Process of care**: Interaction between a provider and a patient
- **Outcomes of care**: Effects on patient prognosis, symptoms, functional status, or other aspects of health-related quality of life, patient satisfaction with care, and patient economic consequences; as well as organizational outcomes such as cost

Figure 1 Working definition for quality-improvement interventions.

Working Definition for Effectiveness, Impacts, or Success

- **Effectiveness**: The comparative effectiveness of the intervention relative to an alternative intervention or usual care
- **Impacts**: The degree to which the intervention results in changes over time in relevant outcomes for the patients and organizations involved
- **Success**: The degree to which the intervention achieves its goals relative to 1) achieving benchmarks or targets for clinical care, acceptability, adoption, implementation, spread, or sustainability/ maintenance; and 2) the logic model for the intervention.

Figure 2 Working definition for effectiveness, impacts or success.

Statistical analysis

We measured levels of agreement among the three teams using both the absolute percentage agreement and the three-way κ statistic. Kappa measures agreement correcting for chance. Twenty-two articles generate an approximate 95% CI bound of ±0.1 on the three-way κ statistic as a measure of agreement. To assess stakeholder panelist ratings, we calculated response frequencies, medians and means across the 12 stakeholders. We adjusted ratings for reviewer effect.[9]

RESULTS

Level of agreement

The agreement across the three reviewer pairs, each of which had already resolved internal disagreements on whether articles reported on development, implementation or evaluation of a QII, was 73% (but with a very low κ of 0.041 due to imbalances in marginal distributions)(table 1).

Reviewer pairs disagreed on six of the 22 articles. To precipitate further discussion on how to improve inter-rater agreement, we surveyed our stakeholder expert panel on whether these six articles were suitable for assessing effectiveness, impact or success of a QII. All 12 experts completed the survey. We found that the experts did not agree regarding this question for any of the six articles (table 2). In every case, responses ranged over at least four of the five points on the scale. The means and medians fell in the 'probably not' to 'no preference' range.

Areas of disagreement

Discussion among expert panelists and rereview by investigators regarding suitability of articles for assessing QII effectiveness, impact or success focused on the following three issues:

1. The QII evaluations lacked data on patient health, provider behaviour, and process of care outcomes (eg, reported only on provider knowledge or attitudes, or addressed care giver health or satisfaction). In one article, a specific organisation was targeted (a general practice), almost all providers and administrative personnel participated, and there was a definite intent to incorporate the results of the study into routine practice and policy.[10] However, the study focused on changes in satisfaction and knowledge of participating general practitioners only without measuring impacts on patient care.
   In a second article, the study aimed to improve provider reporting of adverse drug reactions through a 1 h educational session.[11] The study focused on changes in provider knowledge but did not directly impact the process of care. The authors themselves stated '…we cannot tell from this study the effect that any of these had on clinical care.'
   In a third example, the intervention and evaluation measures focused on care giver rather than on patient health outcomes.[12]

Table 1 Inter-rater agreement on inclusion of publications as quality improvement intervention evaluations

Publication feature assessed through secondary screener	Three-way kappa across three reviewer pairs	No (%) of publications on which three reviewer pairs agreed		No (%) of publications on which reviewers disagreed on the presence or absence of the feature
		Feature was present	Feature was absent	
Did the article report on development, implementation, or evaluation of a quality improvement intervention?	0.041	16/22 (73%)	0/22	6/22 (27%)
Quality improvement intervention refers to: an effort to change/improve the clinical structure, process, and/or outcomes of care by means of an organisational or structural change				

2. The study intervention focused on an aspect of structure/organisation, but there was no evidence that a tested intervention (ie, tested through Plan–Do–Study–Act (PDSA) cycles or another quality improvement methodology) had been adapted to a local context. In the first relevant article, the intervention (sputum submission education for patients by a health worker) took place in a specific organisation (an outpatient tuberculosis hospital) and was administered to a representative sample of the hospital's patients.[13] There was no mention, however, of integrating the change into routine practice, of locally implementing prior research showing effectiveness of the intervention or of ongoing or prior PDSA cycles for developing the intervention in the local context.
   In another example, the study targeted a specific organisational unit (a cardiothoracic surgery clinic) and included all patient–care giver dyads meeting broad criteria.[14] There was no evidence of intent to incorporate this intervention into routine care, however, and no mention of local adaptation of the intervention. The authors stated that '…the aim of the study was to examine whether PC-ACP [patient-centred advance care planning] would be superior to usual care….'
3. Only observational methods, such as correlational analyses, with no pre–post or other comparison group were used to evaluate the intervention. In this example, the QII was a set of diverse quality initiatives not under the control of the authors.[15] The evaluation used a cross-sectional design across multiple hospitals and included data from hospital quality management directors and registries. The study assessed correlations between features of the involved hospitals and quality initiatives and post-MI β-blocker hospital prescription rates. The study found that β-blocker use was associated with physician leadership and a supportive administration.

DISCUSSION

Quality improvement studies have been broadly described as 'the combined and unceasing efforts of everyone—healthcare professionals, patients and their families, researchers, payers, planners and educators— to make the changes that will lead to better patient outcomes (health), better system performance (care) and better professional development (learning)'.[16] We tested an approach to identifying a homogeneous group of articles that reported on the results of QII implementation. This approach consisted of an electronic search strategy, initial screening by title and abstract, and classification by full text review. We aimed to identify empirical evaluations that would be suitable for assessing the effectiveness, impact, or success of QIIs, while recognising that a much broader set of literature is relevant to the scientific development of the QI field as a whole.[2] We defined QIIs as 'an effort to change/improve the clinical structure, process and/or outcomes of care by means of an organizational or structural change.' This definition included interventions such as provider reminders, academic detailing, provider performance reports, and patient or provider education, provided that the interventions were implemented or tested using standard operating procedures. For example, if provider performance reports were delivered as part of routine care, we considered that to be an organisational or structural change. If reports were developed and delivered by outside researchers, for example, that was not an organisational or structural change by our definition. Since QIIs may utilise a variety of study designs to achieve their goals, ranging from classic or cluster-randomised controlled trials to pre–post or post-only assessments, we did not include study design in our definition.

We found that the level of agreement across three reviewer pairs, each of which had already resolved internal disagreements on whether articles reported on development, implementation or evaluation of a QII, was only moderate. The κ value associated with the reviewer ratings was very low, even though the percentage agreement was moderate, a situation known as the 'high agreement–low κ paradox.' This occurs when, as in the case of the articles studied here, marginal distributions are very unbalanced.[17] In our analyses, reviewers showed agreement on the presence of the feature, but there were no articles in which there was agreement on the absence of the feature.

To address disagreement and to enhance inter-rater reliability, we used feedback from an expert panel to develop article selection priorities for subsequent reviews. First, the study team decided that, for inclusion in our evidence review of the effectiveness, impacts or success of a QII, the evaluation should report on effects on patient health, or on care processes or care giver burden known to impact patient health. We would consider evaluations focusing only on financial savings or on changes in provider knowledge or attitudes as a secondary priority in assessing the benefits of a QII.

Second, the study team decided that, for our subsequent evidence review targeting empirical evaluations that would be suitable for assessing effectiveness, impacts or success of a QII, we would include articles reporting on the subset of studies that focus on changing the ongoing structure or organisation of care (eg, policies, procedures, involvement of non-research personnel) within a particular local environment. For example, interventions in which the aim was to change how a relevant practice, hospital or hospital unit, nursing home, public health or community organisation functioned over time would be included. A study of an organisational intervention carried out independently of ongoing routine care structure or context (eg, a narrowly defined intervention carried out primarily by research personnel) would be excluded. Based on similar reasoning, evaluations of a single clinical or public health intervention not incorporated into routine activities at local sites (eg, a one-time educational intervention for providers) would also be excluded.

Table 2 Stakeholder assessment of intervention summaries as quality improvement interventions

Article	Summary for rating	Should the article be classified as a study of the effectiveness, impacts or success of a QII? Ratings adjusted for reviewer effect Scale = 1–5; where 5 = definitely; 4 = probably; 3 = no preference; 2 = probably not; 1 = definitely not		No of panelists endorsing each rating level				
		Mean±SD	Median	1	2	3	4	5
Spurling and Mansfield[10]	'This study aimed to evaluate the interactions between pharmaceutical sales representatives and GPs in an Australian general practice, and develop and evaluate a policy to guide the interaction….Doctors' prescribing, diaries, practice promotional material and samples were audited and a staff survey undertaken. After receiving feedback, the staff voted on practice policy options.'	2.3±1.4	1.8	3	6	0	1	2
Khan *et al*[13]	'…a pragmatic randomized controlled trial to assess the effect of sputum-submission instructions on female patients…Patients in the intervention group were referred to a designated room where they received guidance as to how to produce a good sputum sample from a female health worker who had been trained by the researcher and a senior tuberculosis control officer as to how to provide sputum samples '	2.5±1.4	2.3	3	4	0	4	1
Belle *et al*[12]	'The intervention addressed care giver depression, burden, self-care, and social support and care recipient problem behaviours through 12 in-home and telephone sessions over 6 months…The intervention involved a range of strategies: provision of information, didactic instruction, role playing, problem solving, skills training, stress management techniques, and telephone support groups to reduce risk in the study's five target areas…'	3.0±1.4	3.4	2	3	0	6	1
Figueiras *et al*[11]	'[Objective:] To evaluate the effectiveness of educational outreach visits for improving adverse drug reaction (ADR) reporting by physicians….[Intervention:] One-hour educational outreach visits tailored to training needs identified in a previous study.'	3.3±1.2	3.7	2	0	2	7	1
Song *et al*[14]	'[Objective:] …to evaluate short-term effects of Patient-Centered Advance Care Planning (PC-ACP)…The PC-ACP interview was delivered by a trained nurse facilitator and lasted from 20 to 45 min.'	2.8±1.2	2.9	2	2	3	5	0
Bradley *et al*[15]	'[Objective:] …to identify quality improvement efforts that were associated with hospitals β-blocker prescription rates after acute myocardial infarction (AMI)… This was a cross-sectional study using data from a telephone survey of quality management directors at participating hospitals linked with patient-level data from the National Registry of Myocardial Infarction (NRMI) during the study period….'	3.5±0.77	3.6	0	2	2	7	1

Third, the study team decided that, for an evidence review of QII effectiveness, impacts or success, our focus should be on studies using direct comparisons (eg, pre–post or experimental/control) rather than purely cross-sectional approaches.[6 7] There are many articles in the literature that report on the application of regression analysis and other techniques to cross-sectional data. The aim of many of these is to look at variations in care across or within settings and to evaluate whether the presence of an existing QII, usually among other factors, is associated with improved quality. These articles can be extremely valuable for identifying the utility of different intervention approaches, relevant barriers and facilitators, and other contextual factors that may affect interventions. Their use in evaluating an intervention, however, risks errors resulting from endogeneity, and these types of articles should probably be considered exploratory for that purpose.

Since study questions have important implications for choosing the most appropriate study design, we did not include study design in our definition of a QII. Studies that address how well an intervention works as compared with alternate or usual care might appropriately favour randomised trials or quasi-experimental

designs.[6 7] Studies that address questions of organisational performance or intervention transferability might use, on the other hand, a wider range of designs that incorporate trade-offs across multiple indicators of internal and external validity such as those suggested by the RE-AIM framework (reach, effectiveness, adoption, implementation and maintenance).[18]

While this article focuses on the reader or reviewer perspective with regard to identifying relevant quality improvement publications, we expect our work to have implications for authors as well. The field is still developing and authors often do not label their work with terms that signal relevance to quality improvement. The process of developing a common language for what we mean by QIIs will help authors describe, in titles and abstracts, the framework within which their articles should be read, reviewed and used.

CONCLUSIONS

Even among reviewers familiar with the QII literature and the initial classification scheme, identifying and reaching agreement on articles reporting on QII development, implementation or outcomes was challenging. Contrary to our expectations, reconciliation of ratings resulted in only moderate agreement. To move forward, the field of quality improvement needs to develop and test an acceptable and generalizable taxonomy for QII publications and a flow of investigative approaches that guide the investigator from science to practice.

Acknowledgements The authors would like to thank the following individuals, who helped guide and support the project: D Atkins, VA; F Davidoff, Institute for Healthcare Improvement; M Eccles, Newcastle University Institute of Health and Society; R Lloyd, Institute for Healthcare Improvement; V McLoughlin, The Health Foundation; S Moore, CWRU; D Rennie, UCSF; S Salem-Schatz, Independent Consultant; DP Stevens, Dartmouth Institute; EH Wagner, Group Health Center for Health Studies; B Mittman, VA; G Ogrinc, Dartmouth Institute; and B Johnson (research assistant), RAND.

Funding Robert Wood Johnson (RWJ) Foundation under a grant to LVR (grant ID 65113: Advancing the science of continuous quality improvement: A framework for identifying, classifying and evaluating continuous quality improvement studies).

Competing interests David P Stevens is Editor-in-Chief, Greg Ogrinc is an Associate Editor, and Frank Davidoff serves on the Editorial Advisory Board of *Quality & Safety in Healthcare*.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Liu CF,** Rubenstein LV, Kirchner JE, *et al*. Organizational cost of quality improvement for depression care. *Health Serv Res* 2009;**44**:225—44.
2. **Rubenstein LV,** Hempel S, Farmer MM, *et al*. Finding order in heterogeneity: types of quality-improvement intervention publications. *Qual Saf Health Care* 2008;**17**:403—8.
3. **Berwick DM.** Disseminating innovations in health care. *JAMA* 2003;**289**:1969—75.
4. **Berwick DM.** The science of improvement. *JAMA* 2008;**299**:1182—4.
5. **Campbell M,** Fitzpatrick R, Haines A, *et al*. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;**321**:694—6.
6. **Cook TD,** Campbell DT. *Quasi-experimentation: design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company, 1979.
7. **McLaughlin C,** Kaluzny A. *Continuous quality improvement in health care: theory, implementation, and applications*. London, UK: Jones and Bartlett Publishers International, 2004.
8. **Wall RJ,** Ely EW, Elasy TA, *et al*. Using real time process measurements to reduce catheter related bloodstream infections in the intensive care unit. *Qual Saf Health Care* 2005;**14**:295—302.
9. **Rubenstein LV,** Kahn KL, Reinisch EJ, *et al*. Changes in quality of care for five diseases measured by implicit review, 1981 to 1986. *JAMA* 1990;**264**:1974—9.
10. **Spurling G,** Mansfield P. General practitioners and pharmaceutical sales representatives: quality improvement research. *Qual Saf Health Care* 2007;**16**:266—70.
11. **Figueiras A,** Herdeiro MT, Polonia J, *et al*. An educational intervention to improve physician reporting of adverse drug reactions. A cluster-randomized controlled trial. *JAMA* 2006;**296**:1086—93.
12. **Belle SH,** Burgio L, Burns R, *et al*. Enhancing the quality of life of dementia caregivers from different ethnic or racial groups. *Ann Intern Med* 2006;**145**:727—38.
13. **Khan MS,** Dar O, Sismanidis C, *et al*. Improvement of tuberculosis case detection and reduction of discrepancies between men and women by simple sputum-submission instructions: a pragmatic randomised controlled trial. *Lancet* 2007;**369**:1955—60.
14. **Song MK,** Kirchhoff KTA, Douglas J, *et al*. A randomized, controlled trial to improve advance care planning among patients undergoing cardiac surgery. *Med Care* 2005;**43**:1049—53.
15. **Bradley EH,** Herrin J, Mattera JA, *et al*. Quality improvement efforts and hospital performance rates of beta-blocker prescription after acute myocardial infarction. *Med Care* 2005;**43**:282—92.
16. **Batalden PB,** Davidoff F. What is 'quality improvement' and how can it transform healthcare? *Qual Saf Health Care* 2007;**16**:2—3.
17. **Cicchetti DV,** Feinstein AR. High agreement but low kappa:II. Resolving the paradoxes. *J Clin Epidemiol* 1990;**43**:551—8.
18. **Glasgow RE,** Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health* 1999;**89**:1322—7.

APPENDIX 1 SUMMARY OF ARTICLE REVIEW PROCESS