OPEN ACCESS

Development of the Quality Improvement Minimum Quality Criteria Set (QI-MQCS): a tool for critical appraisal of quality improvement intervention publications

Susanne Hempel,[1] Paul G Shekelle,[1,2] Jodi L Liu,[1] Margie Sherwood Danz,[1,3] Robbie Foy,[4] Yee-Wei Lim,[5] Aneesa Motala,[1] Lisa V Rubenstein[1,3,6]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2014-003151).

For numbered affiliations see end of article.

Correspondence to
Dr Susanne Hempel, RAND Health, RAND Corporation, 1776 Main Street, Santa Monica, CA 90407, USA; susanne_hempel@rand.org

Received 23 April 2014
Revised 15 April 2015
Accepted 22 April 2015
Published Online First 26 August 2015

To cite: Hempel S, Shekelle PG, Liu JL, *et al*. *BMJ Qual Saf* 2015;**24**:796–804.

ABSTRACT

Objective Valid, reliable critical appraisal tools advance quality improvement (QI) intervention impacts by helping stakeholders identify higher quality studies. QI approaches are diverse and differ from clinical interventions. Widely used critical appraisal instruments do not take unique QI features into account and existing QI tools (eg, Standards for QI Reporting Excellence) are intended for publication guidance rather than critical appraisal. This study developed and psychometrically tested a critical appraisal instrument, the QI Minimum Quality Criteria Set (QI-MQCS) for assessing QI-specific features of QI publications.

Methods Approaches to developing the tool and ensuring validity included a literature review, in-person and online survey expert panel input, and application to empirical examples. We investigated psychometric properties in a set of diverse QI publications (N=54) by analysing reliability measures and item endorsement rates and explored sources of disagreement between reviewers.

Results The QI-MQCS includes 16 content domains to evaluate QI intervention publications: Organisational Motivation, Intervention Rationale, Intervention Description, Organisational Characteristics, Implementation, Study Design, Comparator Description, Data Sources, Timing, Adherence/Fidelity, Health Outcomes, Organisational Readiness, Penetration/Reach, Sustainability, Spread and Limitations. Median inter-rater agreement for QI-MQCS items was κ 0.57 (83% agreement). Item statistics indicated sufficient ability to differentiate between publications (median quality criteria met 67%). Internal consistency measures indicated coherence without excessive conceptual overlap (absolute mean interitem correlation=0.19). The critical appraisal instrument is accompanied by a user manual detailing *What to consider*, *Where to look* and *How to rate*.

Conclusions We developed a ready-to-use, valid and reliable critical appraisal instrument applicable to healthcare QI intervention publications, but recognise scope for continuing refinement.

INTRODUCTION

Quality improvement (QI) interventions account for substantial investments by organisations aiming to improve healthcare quality, and a large volume of literature documents these efforts.[1] QI research necessarily reflects work with organisational context and local environments. QI interventions tend to be complex, multicomponent, often uniquely tailored to settings, and may evolve over time.[1 2] Intervention details, context and information on the QI process are critical to evaluate the success of QI interventions.

To address the unique requirements of QI research, the Standards for QI Reporting Excellence (SQUIRE) group has developed detailed guidance for reporting evaluations of QI interventions.[3] The reporting guideline helps authors describe QI interventions so that they can be identified as such in electronic databases. It aims

to ensure readers can understand and appraise the intervention and its evaluation by identifying for authors the details they need to report. However, tools are also needed to guide the critical appraisal of published QI studies. Critical appraisal assesses the quality of publications, informs decisions about applicability of results, and aims to identify high-quality published studies. While reporting guidelines can be aspirational and comprehensive because they are designed for future publications, critical appraisal tools must be applicable to the wide range of completed studies and concentrate on key assessment domains if they are to be useful in practice.

Researchers have frequently questioned the methodological quality of QI studies.[4] However, tools widely used for the critical appraisal of clinical interventions, such as the Cochrane Risk of Bias tool,[5] may not encompass the domains most relevant to QI research. The lack of a QI-specific focus can limit the ability of researchers, practitioners and policy makers to identify—and learn from—higher quality QI studies.

We have developed the QI Minimum Quality Criteria Set (QI-MQCS) to appraise the quality of QI-specific aspects of QI publications. The QI-MQCS is intended as a resource for reviewers, assisting in synthesising the vast available evidence on QI interventions, and providing a framework for critical appraisal in this complex research area. This article describes the development and evaluation of the QI-MQCS.

METHODS

Our international workgroup of QI and systematic review experts (subsequently called 'workgroup') followed a structured process to develop and evaluate the QI-MQCS. We used a broad and inclusive definition of QI interventions to ensure the QI-MQCS applies to a variety of efforts to change/improve the clinical structure, process and/or outcomes of care by means of an organisational or structural change.

The QI-MQCS reflects core domains developed through literature review, inputs from QI experts and stakeholders, and item development through iterative application to empirical studies. Formal reliability testing and reviewer guidance were used to enable consistent and replicable scoring. We designed the QI-MQCS items to be modest in number to ensure scoring feasibility, have strong face validity with QI stakeholders, meet psychometric standards to enable reliable assessment, avoid repeating internal validity items from study-design specific appraisal tools and applicable to a wide range of QI publications.

The following describes the development of the domains (the content the QI-MQCS aims to cover), the operationalisation as QI-MQCS items (the concrete appraisal questions and scoring criteria), the available tools and resources (the QI-MQCS form and manual) and the psychometric evaluation of the QI-MQCS.

Domain development

To ensure that the QI-MQCS represents the breadth of relevant domains,[6] we first reviewed a wide range of existing tools. We assessed widely endorsed general[5 7 8] and specific critical appraisal tools;[9] reporting guidelines for QI and[3 10 11] behaviour change interventions;[12] study design-specific guidelines;[13 14] relevant frameworks such as Reach Effectiveness Adoption Implementation Maintenance;[15] and the Medical Research Council (MRC) Guidance for Complex Interventions.[16] Relevant resources were identified through a PubMed literature search for critical appraisal and QI; screening EQUATOR-network.com; critical appraisal resources provided by the Center for Reviews and Dissemination, the Evidence-based Practice Center programme of the Agency for Healthcare Research and Quality, the Oxford Centre for Evidence Based Medicine, the National Institute for Health and Care Excellence, and the Cochrane Effective Practice and Organisation of Care (EPOC) Review Group; and existing systematic reviews of critical appraisal and evidence level hierarchies.[17–19] In addition, workgroup members assessed all 57 SQUIRE items for their relevance to a critical appraisal instrument. They rated 22 items as important or very important, for example 'Describes the intervention and its component parts in sufficient detail that others could reproduce it' and 'Identifies the study design chosen for measuring impact of the intervention on primary and secondary outcomes', but rated many other aspects of the reporting guideline as less important (eg, 'Title states the specific aim of the intervention', 'Discussion relates results to other evidence').

A consensus panel of international technical experts and key stakeholders in QI interventions, informed by the literature review and SQUIRE survey results, established the QI-MQCS domains.[20] We elicited the input of this technical expert panel (TEP) through online surveys and in person meetings.[21] The project aim was to establish a feasible instrument that covers core QI domains rather than compiling an exhaustive list of potentially relevant or intervention-specific elements. An overarching conclusion of the content discussions was that the QI-MQCS should address domains that complement, rather than replace, instruments addressing the internal validity of study designs.[5]

Operationalisation

The workgroup operationalised the domains as a critical appraisal instrument. Items were iteratively developed to capture the content of the domains and to enable reliable scoring of published articles. We included a domain description (eg, 'Rationale linking the intervention to its expected effects'), guide (eg, 'Consider citations of theories, logic models, or existing empirical evidence that link the intervention to its expected effects') and minimum standard for each quality criterion ('Names or describes a rationale linking at least one central intervention component to

intended effects'). This process involved translating conceptual constructs (eg, 'Penetration/Reach') and phrases open to interpretation (eg, 'Intervention and its component parts described in sufficient detail that others could reproduce it') into practical scoring rules ('Describes the proportion of all eligible units that actually participated'; 'Describes at least one specific change in detail including the personnel executing the intervention'). We sought to avoid conceptual overlap, so that scoring of one domain would not influence other domain scores. We refined the criteria by applying them to empirical examples of the literature.

Throughout the process, we held discussions with key informants and drew upon examples of empirical literature to define the domains and standards for published QI evaluations. We sought input from QI researchers, QI practitioners and systematic reviewers experienced in QI literature syntheses. We applied all suggested critical appraisal domains, reviewer guidance and scoring criteria to empirical examples of existing QI publications to establish the QI-MQCS.

Tools and resources

We designed a form that translates the established domains into critical appraisal items with a dichotomous answer mode and scoring criteria to help reviewers decide whether a minimum quality standard is met. In addition, we adopted the Appraisal of Guidelines for Research and Evaluation structure[22] to provide detailed guidance for QI-MQCS users. The *Description* defines the domain, *What to Consider* lists aspects relevant to the domain, *Where to Look* directs users to where the information is typically found in publications and *How to Rate* guides item scoring. The guidance provides illustrative article excerpts relevant to each domain.

Psychometric evaluation

To test the psychometric properties of the QI-MQCS, we used a validated QI search strategy to identify an empirical sample of diverse published QI and continuous QI intervention studies indexed in PubMed.[23] The strategy combined QI and continuous QI, QI intervention components and EPOC-eligible interventions search terms. We screened the search output to identify publications evaluating the effects of QI interventions. We applied our working definition of QI[24 25] by using four broad criteria to select relevant studies: healthcare delivery organisation context; reporting data on the effectiveness, impacts or success of an intervention; reporting patient, caregiver, provider behaviour, or process of care outcomes; and interventions aiming to change how delivery of care is routinely structured. The interventions in the 54 studies included in the QI-MQCS evaluation data set focused on restructuring of departments and teams, checklists or audit and feedback to increase preventive services and performance indicators, shared medical appointments, pain management programmes, fall management and restraint prevention programmes, staff training and education restructuring, hospital care and diagnostic procedure redesigns, clinical guidelines, medication management models, incentive programmes to increase patient access, computerised registers, discharge planning, antenatal care restructuring, and telehealth.

Two reviewers agreed on the main intervention and outcome for each publication prior to quality appraisal; if publications referred to additional publications on the same study, we obtained them as well. Studies were reviewed by two independent reviewers in batches of nine and then reconciled, mirroring a systematic review process that uses independent reviewers and reviewer reconciliation to reduce reviewer errors and bias. In cases where we had to revise items to incorporate additional guidance, we discarded previous ratings. Psychometric results shown below reflect the final version of the QI-MQCS.

We analysed the answer frequency for each item (item endorsement rate: number of publications meeting the criterion in the sample) based on ratings reconciled across two reviewers.[26] We measured two aspects of reliability: rater agreement and internal consistency.[27] Agreement was measured through Cohen's κ and the per cent agreement of two independent reviewers before reconciliation. We assessed internal consistency and conceptual overlap across the QI-MQCS domains through interitem correlations across the 16 assessed items, across all assessed publications, and based on reconciled reviewer ratings (correlating each item score with all other item scores to quantify the empirical associations between individual items). Finally we identified sources of disagreement for each of the assessed publications.

RESULTS

QI-MQCS content

The QI-MQCS addresses the following domains: Organisational Motivation, Intervention Rationale, Intervention Description, Organisational Characteristics, Implementation, Study Design, Comparator, Data Source, Timing, Adherence/Fidelity, Health Outcomes, Organisational Readiness, Penetration/Reach, Sustainability, Spread and Limitations. Table 1 describes each domain and table 2 shows the TEP's ratings of the importance of each domain (face validity).

Organisational Motivation assesses whether the motivational context of the organisation in which the intervention was introduced was described; for example to convey whether a given quality problem—such as shortcomings in quality of care indicators—was being addressed. *Intervention Rationale* assesses whether a rationale was given that suggests why the intervention may produce improvements in the outcome (empirical evidence, theories or logic models).

Table 1 Quality Improvement Minimum Quality Criteria Set (QI-MQCS) domains

Domain	Description
1. Organisational motivation	Organisational problem, reason or motivation for the intervention
2. Intervention rationale	Rationale linking the intervention to its expected effects
3. Intervention description	Change in organisational or provider behaviour
4. Organisational characteristics	Demographics or basic characteristics of the organisation
5. Implementation	Temporary activities used to introduce potentially enduring changes
6. Study design	Study design and comparator
7. Comparator	Information about comparator care processes
8. Data source	Data sources and outcome definition
9. Timing	Timing of intervention and evaluation
10. Adherence/fidelity	Adherence to the intervention
11. Health outcomes	Patient health-related outcomes
12. Organisational readiness	Barriers and facilitators to readiness
13. Penetration/reach	Penetration/reach of the intervention
14. Sustainability	Sustainability of the intervention
15. Spread	Ability to be spread or replicated
16. Limitations	Interpretation of the evaluation

Intervention Description requires a detailed description of the change in the structure or organisation of healthcare, including personnel involved. QI interventions are diverse and may address changes in care processes (eg, use of care managers) or strategies aiming to change provider behaviour (eg, electronic reminders), and the content (eg, avoiding catheter-related blood stream infections), and the means to achieve the goal (eg, audit and feedback) are often intertwined. We restricted the definition to permanent structural or organisational changes, not temporary activities aiming to develop or introduce the change. This domain had the highest rating in the assessment of the domain importance shown in table 2.

Organisational Characteristics assesses whether key demographics of the setting are described to provide information that enables readers to assess the generalisability to their organisation.

Implementation addresses temporary activities used to introduce the permanent change, for example, staff education to introduce a new care protocol. The QI-MQCS focuses here on the introduction of the intervention into clinical practice, not its development.

Study Design assesses whether the evaluation design to determine whether the intervention was successful was identified. Acknowledging that different questions require different study designs, the quality emphasis is on outlining the evaluation approach, not on specific designs or features (eg, randomisation).

Comparator assesses the control condition to which the intervention is compared, for example, routine care before the intervention was introduced. We added this item, most prominently described in the Workgroup for Intervention Development and Evaluation Research (WIDER) criteria,[12] in response to TEP discussions and empirical evidence.[28] Given that healthcare contexts are continually evolving, it is important to know whether the comparison group comprised current 'state-of-the-art' or

Table 2 Technical expert panel (TEP) ratings of included QI domains and per cent criterion met

#	Domain	Panel item	Mean rating*	% Criterion met†
1	Organisational motivation	Description of the organisational problem/reason or motivation for intervention	2.78	64
2	Intervention rationale	Description of rationale linking the intervention to expected effects	2.78	67
3	Intervention	Description of specific changes in healthcare delivery organisation/structure	3.00	93
4	Organisational characteristics	Description of organisational demographics and basic characteristics	2.89	89
5	Implementation	Description of the approach to designing and/or introducing organisational changes	2.89	92
6	Study design	Description of study design	2.89	44
7	Comparator	n/a	n/a	67
8	Data source	n/a	n/a	67
9	Timing	Description of timing (intervention components introduction and evaluation)	2.78	56
10	Adherence/fidelity	Description of intervention adherence/fidelity	2.78	47
11	Health outcomes	Description of health-related outcomes	2.33	58
12	Organisational readiness	Description of organisational readiness for the studied intervention	2.00	84
13	Penetration/reach	Description of intervention penetration/reach	2.56	85
14	Sustainability	Description of potential for intervention maintenance or sustainability	2.22	83
15	Spread	Description of ability to be spread or replicated	2.11	89
16	Limitations	Quality of the interpretation of findings	2.56	64

*Members (N=9) of an international TEP assessed independently whether the domain should (score=3), should maybe (score=2) or should not (score=1) be part of the Quality Improvement Minimum Quality Criteria Set (QI-MQCS). The respondents were instructed that the goal was to identify a minimum number of core domains; n/a: not applicable, the items were developed as a response to panel input.
†Percentage of publications meeting the criterion in psychometric evaluation sample (total N=54 publications, number of observations ranged from 18 to 45 as only the final item version was included in the analysis).

poor quality care. *Data Source* considers how data were obtained for the evaluation and whether the primary outcome was defined; conveying what exactly was measured should avoid a 'false implicit understanding' of terms and definitions[24] and is independent from the study design selected for the evaluation.

Timing addresses the clarity of the timeline in relation to the evaluation of the intervention, for example, when a complex change was fully implemented and when evaluated, in order to determine the follow-up period. *Adherence/fidelity* addresses compliance with the intervention. QI interventions can be introduced with enthusiasm, but whether personnel actually adhere to them (eg, a new assessment tool) in busy routine clinical practice is another matter. Readers need to be able to judge whether any intervention failure was attributable to the intervention itself, suboptimal translation in clinical practice, or a combination of both. Any information on adherence (including the lack thereof) is acknowledged in assessing this domain.

Health Outcomes considers whether patient health outcomes are part of the evaluation. Although an intervention may result in changes in healthcare processes (eg, tests ordered), they may not necessarily improve patient outcomes. The QI-MQCS acknowledges studies that assess this crucial patient-centered question. *Organisational Readiness* refers to the QI culture and resources present in the organisation, which helps to assess the transferability of results. The TEP did not express strong unanimous support for including this item (table 2).

Penetration/reach assesses what proportion of eligible units participated. This domain requires a denominator; stating the number of participating sites without also reporting how many sites were initially approached or were eligible is not sufficient. *Sustainability* addresses whether information on the sustainability of the intervention is available; including positive evidence (eg, an extended intervention period) or acknowledgment that the intervention may be maintained only with additional resources.

Spread addresses the ability of the intervention to be spread to or replicated in other settings. The minimum quality standard is met if the potential or unsuccessful attempts at spread or positive evidence of spread (eg, large-scale rollouts) are presented. *Limitation* refers to disclosed limitations of the evaluation of the intervention.

Online supplementary appendix 1 shows the QI-MQCS, a ready-to-use form for critical appraisal. Online supplementary appendix 2, a user manual developed for the QI-MQCS, provides detailed information on each domain and scoring criteria, including *What to consider*, *Where to look* and *How to rate*.

Psychometric properties

The item endorsement rates (criterion met) ranged between 44% and 93% (table 2) with a median rate of 67% indicating that the QI-MQCS items were able to differentiate between high and low quality studies in an empirical sample of QI publications. Two items were endorsed in more than 90% of assessed QI publications (*Intervention Description* and *Implementation*).

The median inter-rater agreement between two independent reviewers across all items was $\kappa=0.52$ and 82% agreement (table 3). Coefficients ranged from $\kappa=0.09$ (*Adherence/fidelity*) to $\kappa=0.82$ (*Sustainability*) with corresponding per cent agreement values of 56% and 74%. Agreement for 81% of items was fair to good; the items *Timing*, *Adherence/fidelity* and *Spread* were below $\kappa=0.40$. Sources of disagreements between reviewers are documented in table 4 and encompassed omissions (ie, a reviewer overlooked reported information), the interpretation of the reported information (eg, associated with disagreements in *Adherence/fidelity*) and the interpretation of criteria (ie, sufficient to meet the criterion).

The mean interitem correlation across all QI-MQCS items in the empirical sample of QI publications was 0.08 (mean absolute interitem correlation 0.19) and all individual interitem correlations were below 0.67. Results indicated conceptual independence between criteria (discriminant validity); items showed some coherence but not identity of assessed domains. Correlations of 0.61 to 0.66 were found for the domains *Intervention Description* and *Data source*, *Implementation* and *Organisational Readiness*, and *Data Source* correlated with *Penetration/Reach* as well as *Limitations*.

Table 3 Inter-rater agreement Quality Improvement Minimum Quality Criteria Set (QI-MQCS)

#	Domain	n	κ (95% CI)	% agreement
1	Organisational motivation	45	0.46 (0.19 to 0.73)	0.76
2	Intervention rationale	18	0.61 (0.21 to 1.00)	0.83
3	Intervention	27	0.65 (0.02 to 1.28)	0.96
4	Organisational characteristics	45	0.49 (0.17 to 0.82)	0.84
5	Implementation	36	0.62 (0.23 to 1.01)	0.92
6	Study design	45	0.73 (0.53 to 0.93)	0.87
7	Comparator description	54	0.40 (0.14 to 0.65)	0.72
8	Data source	18	0.87 (0.62 to 1.12)	0.94
9	Timing	54	0.39 (0.15 to 0.63)	0.70
10	Adherence/fidelity	36	0.09 (−0.22 to 0.40)	0.56
11	Health-related outcomes	45	0.64 (0.42 to 0.87)	0.82
12	Organisational readiness	45	0.45 (0.14 to 0.76)	0.82
13	Penetration/reach	27	0.52 (0.18 to 0.85)	0.81
14	Sustainability	18	0.82 (0.49 to 1.15)	0.94
15	Spread	27	0.13 (−0.23 to 0.48)	0.67
16	Limitations	45	0.77 (0.58 to 0.96)	0.89

κ, Cohen's κ; n, Number of assessed publications.

Table 4 Sources of reviewer disagreements

Source of disagreement	Source description	Literature examples
Omissions	Some disagreements were associated with simple reviewer mistakes, that is, one reviewer overlooking reported information	Several disagreements were simply due to one reviewer overlooking reported information and did not seem to follow any pattern (random errors). However, the low agreement in the Spread domain seemed to have, in parts, to do with information being 'buried' in the discussion section Omission-based disagreement was also encountered repeatedly for the domain Organisational characteristics, due to information not being reported in the main manuscript text but elsewhere, for example in the author's biography[32]
Interpretation of reported information	Some disagreements were associated with the interpretation of the information that was reported in the publication	The low agreement in the domain Adherence/fidelity was to some extent associated with publications where adherence was the main outcome or the outcome and the intervention were identical (eg, guideline implementation to improve adherence to evidence-based practices)[33] A further example was whether reviewers considered a state-wide initiative sufficient to infer the motivation to participate for all included hospitals.[34] Multiple site studies often do not provide information on individual facilities[35] and studies in low-income countries may have had an initiating body that was not a healthcare delivery organisation[36] and reviewers disagreed to which extent they extrapolated from the presented information to individual organisations Disagreements in the Health Outcome domain were associated with the type of outcome and how systematically data were collected in order to be recognised as a health outcome/data[37]
Interpretation of criteria	Despite the careful, iterative development of the tool, some disagreements were associated with the interpretation of the scoring criteria. Given the large scope of interventions included in the test set, some ambiguities could not be resolved	Identified disagreement in the domain Intervention Rationale was associated with publications where only highly selective intervention components were linked to existing empirical literature and reviewers disagreed whether the specific aspect was sufficient to meet the criterion[38] Disagreements in the Comparator domain were associated with the question of how much detail was considered sufficient to meet the quality criterion, for example, if only a component of the usual care was described[34] Disagreements also occurred when publications described a structural change without information on the uptake, for example, an installation of a comfort room for patients—but whether the room was used in clinical practice was not reported; hence reviewers had to decide whether the intervention was the installation of the room or the use of the room[39]

Examples taken from validation sample (N=54 publications), rater agreement is documented in table 3.
Mistakes (omissions) as well as remaining ambiguity (interpretation of reported information and interpretation of criteria) were sources of disagreement between literature reviewers. A qualitative analysis of the disagreements pointed to some systematic, rather than random, reviewer errors.

DISCUSSION

The QI-MQCS is a critical appraisal instrument that assesses 16 expert-endorsed QI domains applicable to a wide range of QI studies. Its scoring guidance facilitates use by different raters with known psychometric properties. A structured critical appraisal instrument development process ensured feasibility, validity and reliability.

The QI-MQCS development included a comprehensive literature search to ensure content validity and iterative development of the operationalisation of domains applied to existing, published QI literature to ensure construct validity. The empirical test of the QI-MQCS shows sufficient ability to discriminate between studies, indicating that the QI-MQCS avoids items representing unattainable standards but includes items that discriminate quality across an empirical sample of publications. Furthermore, the QI-MQCS does not show excessive conceptual overlap across domains, and none of the items shows redundancy with content already captured through other items. Agreement between two independent reviewers was fair to good in a diverse sample of a complex research field.

Despite the careful, iterative development of items and scoring criteria, some domains showed limited rater agreement, such as adherence/fidelity and spread. Future work is warranted to test the reliability in a narrower set of interventions, for example, those included in typical systematic reviews, or to develop the criteria further in order to achieve better consensus. However, the QI-MQCS compared favourably to some other commonly used tools, such as the Cochrane Risk of Bias tool.[29] Plus, few published

quality assessment tools have been tested for their psychometric properties.[17] Reviewer disagreements may be easier to anticipate and to avoid in a more restricted sample, for example, one that is limited to a set of selected QI interventions.

QI stakeholders agree on the pressing need for better research and better literature synthesis methods. The QI-MQCS was developed to support evidence synthesis by providing a critical appraisal tool to identify high quality QI studies, for example, in the context of a systematic review. It is designed to be applicable to a wide range of QI studies. Developing critical appraisal criteria for QI publications is challenging due to the diversity of QI interventions, interdisciplinary language and study designs. Consequently, the QI-MQCS assesses, for example, whether the rationale specified for the intervention links to the study's main outcome, without dictating which type of rationale (eg, which evidence-based intervention or theory) may be superior, given that this determination may depend on the specific interventions in this particular field of research. To ensure wide applicability, we purposefully applied the QI-MQCS to a diverse set of QI publications in the psychometric evaluation and did not limit the sample to specific clinical conditions, QI interventions, outcomes or study designs.

The QI-MQCS targets the informational value of the QI study, giving credit to publications that assess and provide information on crucial variables. Thus, for example, a publication that reports limited adherence to an intervention or describes that the spread of the intervention was unsuccessful receives credit for reporting on adherence and spread. Reviewers may want to highlight positive expressions of the domain, for example, evidence of adherence indicating that the intervention took place as outlined. In this case, the QI-MQCS can be used as a framework for a more refined assessment. The specific standards will depend on the individual field of application.

We designed the QI-MQCS to determine the minimum quality threshold of core QI domains. QI experts selected and prioritised domains in order to establish a feasible critical appraisal instrument. Furthermore, we developed detailed scoring criteria in an iterative process to ensure reliability. The assessment must rely on the information presented in the publication, and reliable scoring requires clear guidance that cannot be based on guessing or inside knowledge of individual reviewers. Nevertheless, reporting shortcomings may not necessarily indicate the absence of the process in the conduct of the study (eg, the publication's word limits may have precluded a full description of the methods) and the psychometric evaluation distinguished only whether the domain criteria were met or not. Using the QI-MQCS and its assessment domains as a framework may allow reviewers to further differentiate study quality by creating response options for partially met criteria; by differentiating unmet criteria into 'unclear' and 'low quality,' or by defining criteria for exceptionally high quality studies. Further differentiation and moving away from the dichotomy of minimum criteria met or not may also provide a resolution for some of the described disagreements between reviewers. Additional or alternative criteria, for example criteria capturing other aspects of QI interventions[3] building on the QI-MQCS may be important in specific research contexts.

The QI-MQCS was explicitly designed to complement, not to replace, critical appraisal instruments focusing on the internal validity of study designs. Other tools that may be helpful to reviewers are the EPOC group criteria for randomised controlled trials, controlled trials and controlled before-after studies;[30] the quality criteria for programme evaluations,[31] and a published critical appraisal instrument for Plan-Do-Study-Act QI.[9] Fan *et al*[4] provide a hierarchy of methodological strength to evaluate a body of evidence for QI interventions.

The QI-MQCS facilitates access to the vast available literature on QI interventions by identifying high quality studies, for example in the context of a systematic review aiming to synthesise the available evidence for specific interventions or outcomes, and provides a framework for critical appraisal in this complex research area. It is accompanied by a ready-to-use standardised quality assessment form and a detailed user manual. However, we have deliberately titled the tool V.1.0, expecting that its use will lead to further refinement and improvements.

CONCLUSIONS

We developed a ready-to-use, valid and reliable critical appraisal instrument applicable to a wide range of healthcare QI intervention evaluation publications, but recognise scope for continuing refinement.

Author affiliations
[1]RAND Corporation, Santa Monica, California, USA
[2]Veterans Affairs West Los Angeles Medical Center, Los Angeles, California, USA
[3]Veterans Affairs Greater Los Angeles, North Hills, California, USA
[4]University of Leeds, Leeds Institute of Health Sciences, Leeds, UK
[5]National University of Singapore, Saw Swee Hock School of Public Health, Singapore
[6]University of California, Department of Medicine, Los Angeles, California, USA

Acknowledgements The authors thank the members of the international expert panel that guided the selection of the domains: David Atkins, Department of Veterans Affairs; Frank Davidoff, Institute for Healthcare Improvement; Martin Eccles, Newcastle University Institute of Health and Society; Robert Lloyd, Institute for Healthcare Improvement; Vin McLoughlin, The Health Foundation; Shirley Moore, Case Western Reserve University; Drummond Rennie, University of California, San Francisco; Susanne Salem-Schatz, Independent Consultant; David P Stevens, Dartmouth Institute; Edward H Wagner, Group Health Center for Health Studies; Brian Mittman,

Department of Veterans Affairs; and Greg Ogrinc, Dartmouth Institute. The authors thank Sean O'Neill, Denise Dougherty, Judy Sangl and Larry Kleinman for valuable contributions to the workgroup, Roberta Shanman for the literature searches, Jeremy Miles and Marika Suttorp for assistance with the statistical analysis, Breanne Johnsen for project assistance and Sydne Newberry for editorial assistance.

Contributors SH is the guarantor of the work and had full access to the data. LVR and PGS obtained funding; SH, JL, MSD, MA, RF and Y-WL contributed to the data acquisition; JL and SH to the data analysis; LVR, SH, PGS, JL, Y-WL and RF to the interpretation of the data. SH drafted the manuscript, all authors provided critical revisions and approved the final version of the manuscript.

Funding The project was funded by a grant from the Robert Wood Johnson (RWJ) Foundation (ID 65113), the Veterans Affairs (VA) Greater Los Angeles Healthcare System, the RAND Corporation, and the Agency for Healthcare Research and Quality (AHRQ, 1R13HS018139-01). The funding agencies had no role in the study design, collection, analysis and interpretation of the data; the writing and the decision to submit the manuscript for publication. The findings and conclusions are those of the authors; the content of the manuscript should not be construed as the official position of the RWJ Foundation, the Department of Veteran Affairs, the RAND Corporation, or AHRQ.

Competing interests None declared.

Ethics approval The study was reviewed by the Human Subject Protection Committee (HSPC) of the RAND Corporation and determined to be exempt (ID 2009-0071).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The raw data can be obtained from the authors.

REFERENCES

1 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.
2 Rubenstein L, Khodyakov D, Hempel S, *et al*. How can we recognize continuous quality improvement? *Int J Qual Health Care* 2014;26:6–15.
3 Ogrinc G, Mooney SE, Estrada C, *et al*. The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: explanation and elaboration. *Qual Saf Health Care* 2008;17(Suppl 1):i13–32.
4 Fan E, Laupacis A, Pronovost PJ, *et al*. How to use an article about quality improvement. *JAMA* 2010;304:2279–87.
5 Higgins JP, Altman DG, Gotzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
6 Hempel S. Content validity. In: Davey G, ed. *Encyclopaedic dictionary of psychology*. London: Hodder Arnold, 2005: Statistic section (Section Ed: A. Field) p. 421.
7 Whiting P, Rutjes AW, Reitsma JB, *et al*. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
8 Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1–12.
9 Speroff T, James BC, Nelson EC, *et al*. Guidelines for appraisal and publication of PDSA quality improvement. *Qual Manag Health Care* 2004;13:33–9.
10 Moss F, Thompson R. A new structure for quality improvement reports. *Qual Health Care* 1999;8:76.
11 Moss F, Thomson R. A new structure for quality improvement reports. *Qual Saf Health Care* 2004;13:6–7.
12 Abraham C, Albarracín D, Araújo-Soares V, *et al*. WIDER Recommendations to Improve Reporting of the Content of Behaviour Change Interventions.
13 Strengthening the reporting of observational epidemiological studies. STROBE Statement. Checklist of Essential Items. Secondary Strengthening the reporting of observational epidemiological studies. STROBE Statement. Checklist of Essential Items. http://www.strobe-statement.org/Checkliste.html
14 Consolidated Standards of Reporting Trials (CONSORT). Secondary Consolidated Standards of Reporting Trials (CONSORT). http://www.consort-statement.org/
15 Reach Effectiveness Adoption Implementation Maintenance (RE-AIM). Secondary Reach Effectiveness Adoption Implementation Maintenance (RE-AIM). http://www.re-aim.org/
16 Medical Research Council. Developing and evaluating complex interventions: new guidance. Secondary Developing and evaluating complex interventions: new guidance. http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC004871
17 West S, King V, Carey TS, *et al*. Systems to rate the strength of scientific evidence. *Evid Rep Technol Assess (Summ)* 2002;47:1–11.
18 Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;36:666–76.
19 Deeks JJ, Dinnes J, D'Amico R, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;7:iii–x, 1–173.
20 Rubenstein L, Danz M, Hempel S, *et al*. *Advancing the science of continuous quality improvement: A framework for identifying, classifying and evaluating continuous quality improvement studies*. Final project report. June 2010.
21 Rubenstein L, Hempel S, Danz M, *et al*. *Final Progress Report Expert Panel Meeting Advancing the Science of Continuous Quality Improvement. Grant Award Number: 1R13HS018139-01*. Prepared for AHRQ,May 2011.
22 AGREE: Advancing the science of practicing guidelines. Secondary AGREE: Advancing the science of practicing guidelines. http://www.agreetrust.org/
23 Hempel S, Rubenstein LV, Shanman RM, *et al*. Identifying quality improvement intervention publications—a comparison of electronic search strategies. *Implement Sci* 2011;6:85.
24 Danz MS, Rubenstein LV, Hempel S, *et al*. Identifying quality improvement intervention evaluations: is consensus achievable? *Qual Saf Health Care* 2010;19:279–83.
25 Hempel S, Shetty KD, Shekelle PG, *et al*. *Machine learning methods in systematic reviews: identifying quality improvement intervention evaluations*. Research White Paper (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 12-EHC125-EF. Rockville, MD: Agency for Healthcare Research and Quality, September 2012. http://www.effectivehealthcare.ahrq.gov/ehc/products/478/1270/WhitePaper_MachineLearningSR_FinalReport_20120920.pdf

26 Hempel S, Suttorp MJ, Miles JNV, *et al*. *Empirical Evidence of Associations Between Trial Quality and Effect Sizes*. Methods Research Report. (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 11-EHC045-EF. Rockville, MD: Agency for Healthcare Research and Quality, April 2011. http://effectivehealthcare.ahrq.gov

27 Hempel S. Reliability. In: Miles J, Gilbert P, eds. *A handbook of research methods for clinical & health psychology*. Oxford: Oxford University Press, 2005, pp. 193–204.

28 Hempel S, Newberry S, Wang Z, *et al*. Hospital fall prevention: a systematic review of implementation, components, adherence, and effectiveness. *J Am Geriatr Soc* 2013;61:483–94.

29 Hartling L, Hamm MP, Milne A, *et al*. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J Clin Epidemiol* 2013;66:973–81.

30 Cochrane Effective Practice and Organisation of Care Group. Suggested risk of bias criteria for EPOC reviews. Secondary Suggested risk of bias criteria for EPOC reviews 12 August 2013. 2013. http://epoc.cochrane.org/sites/epoc.cochrane.org/files/uploads/14%20Suggested%20risk%20of%20bias%20criteria%20for%20EPOC%20reviews%202013%2008%2012.pdf

31 Ovretveit J, Gustafson D. Evaluation of quality improvement programmes. *Qual Saf Health Care* 2002;11:270–5.

32 Hensing JA. The quest for upper-quartile performance at Banner Health. *J Healthc Qual* 2008;30:18–24.

33 Panaretto KS, Mitchell MR, Anderson L, *et al*. Sustainable antenatal care services in an urban Indigenous community: the Townsville experience. *Med J Aust* 2007;187:18–22.

34 Pronovost P, Needham D, Berenholtz S, *et al*. An intervention to decrease catheter-related bloodstream infections in the ICU. *N Engl J Med* 2006;355:2725–32.

35 Peterson A, Carlhed R, Lindahl B, *et al*. Improving guideline adherence through intensive quality improvement and the use of a National Quality Register in Sweden for acute myocardial infarction. *Qual Manag Health Care* 2007;16:25–37.

36 Kanara N, Cain KP, Laserson KF, *et al*. Using program evaluation to improve the performance of a TB-HIV project in Banteay Meanchey, Cambodia. *Int J Tuberc Lung Dis* 2008;12 (3 Suppl 1):44–50.

37 Stajduhar KI, Bidgood D, Norgrove L, *et al*. Using quality improvement to enhance research readiness in palliative care. *J Healthc Qual* 2006;28:22–8.

38 Liu SK, Homa K, Butterly JR, *et al*. Improving the simple, complicated and complex realities of community-acquired pneumonia. *Qual Saf Health Care* 2009;18:93–8.

39 Barton SA, Johnson MR, Price LV. Achieving restraint-free on an inpatient behavioral health unit. *J Psychosoc Nurs Ment Health Serv* 2009;47:34–40.