BMJ Open Quality

Quality of locally designed surveys in a quality improvement collaborative: review of survey validity and identification of common errors

Julie E Reed ,[1,2] Julie K Johnson,[3] Robert Zanni,[4] Randy Messier,[5] Fadi Asfour,[6] Marjorie M Godfrey[5]

To cite: Reed JE, Johnson JK, Zanni R, *et al*. Quality of locally designed surveys in a quality improvement collaborative: review of survey validity and identification of common errors. *BMJ Open Quality* 2024;**13**:e002387. doi:10.1136/bmjoq-2023-002387

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjoq-2023-002387).

Received 19 July 2023
Accepted 9 January 2024

[1]Julie Reed Consultancy Ltd, London, UK
[2]Halmstad University School of Health and Welfare, Halmstad, Sweden
[3]Northwestern Quality Improvement, Research, and Education in Surgery, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
[4]Robert Wood Johnson Barnabas Health Medical Group, Monmouth Medical Center, Long Branch, New Jersey, USA
[5]University of New Hampshire, Durham, New Hampshire, USA
[6]UTHSC, Utah, Utah, USA

Correspondence to
Professor Marjorie M Godfrey; margiegodfrey@gmail.com

ABSTRACT

Objective Surveys are a commonly used tool in quality improvement (QI) projects, but little is known about the standards to which they are designed and applied. We aimed to investigate the quality of surveys used within a QI collaborative, and to characterise the common errors made in survey design.

Methods Five reviewers (two research methodology and QI, three clinical and QI experts) independently assessed 20 surveys, comprising 250 survey items, that were developed in a North American cystic fibrosis lung transplant transition collaborative. Content Validity Index (CVI) scores were calculated for each survey. Reviewer consensus discussions decided an overall quality assessment for each survey and survey item (analysed using descriptive statistics) and explored the rationale for scoring (using qualitative thematic analysis).

Results 3/20 surveys scored as high quality (CVI >80%). 19% (n=47) of survey items were recommended by the reviewers, with 35% (n=87) requiring improvements, and 46% (n=116) not recommended. Quality assessment criteria were agreed upon. Types of common errors identified included the ethics and appropriateness of questions and survey format; usefulness of survey items to inform learning or lead to action, and methodological issues with survey questions, survey response options; and overall survey design.

Conclusion Survey development is a task that requires careful consideration, time and expertise. QI teams should consider whether a survey is the most appropriate form for capturing information during the improvement process. There is a need to educate and support QI teams to adhere to good practice and avoid common errors, thereby increasing the value of surveys for evaluation and QI. The methodology, quality assessment criteria and common errors described in this paper can provide a useful resource for this purpose.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Front-line improvement teams frequently default to the use of surveys to gather information to guide improvement activities, however, little is known about the quality of surveys developed by local quality improvement (QI) teams, and poorly designed surveys may misguide improvement activities.

WHAT THIS STUDY ADDS

⇒ This study demonstrates the variable quality of locally developed surveys for QI, indicating survey development requires careful consideration, time and expertise.

⇒ Key lessons are identified highlighting common errors in survey design relating to ethics, appropriateness, usefulness and methodological issues.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study presents a method for peer-reviewing surveys that can be applied by other QI teams and collaboratives.

⇒ The findings and key lessons can inform education about the design and development of surveys for front-line improvement teams.

INTRODUCTION

Measurement plays a central role in all quality improvement (QI) approaches.[1–4] However, research has demonstrated that QI methods and approaches are not always used with high fidelity or scientific rigour[5 6] leading to calls to improve the quality of QI,[7] including areas of measurement and evaluation.[8]

Research exploring the use of quantitative data in improvement efforts has highlighted the challenges of developing useful metrics, including precise measure definitions, reliable collection of high-quality data and appropriate analysis, interpretation and action in response to results.[4 9] While these challenges exist in any nationally or regionally driven improvement effort, such problems are particularly prominent in local QI efforts where well intended teams may not have extensive skills or experience in developing and using effective quantitative measures.[4] While there is a growth in the literature on the challenges of quantitative measures and how to overcome them, the same investigation has not taken place for the development

and use of surveys to gain information as part of the QI process.

Surveys are a popular tool in QI that provide a structured approach to capture information from patient or staff respondents about their experiences, opinions, views and impressions.[10] While surveys can include space for qualitative written responses, their power comes from their ability to translate qualitative information into semi-quantified data amenable to statistical analysis. For example, asking about levels of patient satisfaction using a 5-point Likert scale (ranging from very satisfied to very dissatisfied) translates a qualitative subjective opinion into a data point. This allows descriptive statistics to be performed at (a) population level (eg, 83% of patients were very satisfied with the service), (b) comparative statistics to be performed between discrete populations (eg, more patients were very satisfied at hospital X than hospital Y) or (c) overtime (eg, patient satisfaction increased from 40% to 60%).

Research to date has focused on the use of surveys in QI that have been designed for wide scale use.[11 12] The challenges of developing validated surveys is well recognised, and significant effort and expertise has been invested in areas of survey development in national surveys and by clinical specialist groups, for example, patient experience and outcome measures, and patient safety.[13–15] However, in QI initiatives, surveys are usually developed at a more granular local level to support bespoke investigation, evaluation and improvement efforts. The quality of survey instruments developed in such settings has not been explored.

This study aims to assess the quality of surveys produced by QI teams in a QI collaborative. The quality of surveys is often assessed using professional consensus methods such as the Content Validity Index (CVI).[16 17] Such methods select which survey items are high quality based on a high proportion of favourable opinions among a group of independent assessors with relevant expertise. While such methods provide subjective assessment to identify any problematic survey items they do not provide insights as to why the decisions were made, and therefore an opportunity is lost to draw on the reviewers insights to inform the design of future surveys. Therefore this study also aims to explore the reasons behind the survey assessment scores in order to identify lessons about common errors of survey design to help other QI teams avoid common pitfalls and strengthen local evaluations.

This study conducted primary research within a multisite QI collaborative. Due to the on-going and expanding nature of the collaborative there was an opportunity for the work from this research to inform future work of current sites participating in the collaborative and new sites that join as the collaborative expands. As such this research also sought to make pragmatic contributions to the collaborative in identifying which surveys and survey items were recommended to be used by the teams, and to provide guidance to the teams on how to improve the quality of any bespoke surveys they developed. In addition, we believe the demonstration of a method for reviewing survey quality, and insights as to common errors in survey design, will provide valuable guidance to 'improvers' at the front-line of care delivery developing QI surveys, and those responsible for running QI collaboratives or other large programmes of improvement work.

METHODS

Setting: the cystic fibrosis lung transplant transition learning and leadership collaborative

The Cystic Fibrosis Foundation, based in the USA, has a long tradition of organising to improve care for people with cystic fibrosis (CF) and their families. In 2016, people with advanced CF lung disease who had lung transplant reported after their lung transplantation they felt they were no longer part of the CF family and the referral processes from CF programmes to transplant programmes was 'broken'.[18]

The CF Lung Transplant Transition Learning and Leadership Collaborative (CF LTT LLC, herein referred to as 'the collaborative') launched in 2017 was adapted from twenty years of experience designing improvement collaboratives for people with CF (led by MMG). Collaborative methodology was based on the original Institute for Healthcare Improvement break through series framework[19] and was modified to include the microsystem improvement process including people with CF and family members,[20 21] and team coaching.[16] The original CF LLC programme was adapted for the collaborative to improve not only one microsystem, but two microsystems (CF referral and lung transplant programmes individually) and the mesosystem of CF lung transplantation (CF referral and lung transplant programmes together) with a shared purpose to improve care for people with advanced CF lung disease.

The aim of the collaborative was 'within the context of a learning community, explore, improve and decrease practice variation in the systems and processes of lung transplant referrals and transitions from CF programmes to transplant programmes and then to a model of shared responsibility for the patient's care'. In 2017 the original CF LTT collaborative launched 10 pilot pairs of CF referring and lung transplant improvement teams in the USA and Canada. Following 18 months of the collaborative, a CF LTT regional dissemination network (RDN) was created to share the findings and lessons learned from the pilot programme with new CF referral programmes in each of the regional lung transplant programmes. CF LTT RDN wave 1 was launched in March 2019 engaging 10 new CF referring improvement teams, with wave 2 joining in June 2019 (six new CF referring teams) and wave 3 in September 2019 (seven new CF referring teams). This resulted in 33 CF teams partnering with 10 regionally based lung transplant teams. The surveys produced by these teams were reviewed in this study.

Within the LLC methodology, the teams were encouraged to explore their local microsystem to understand

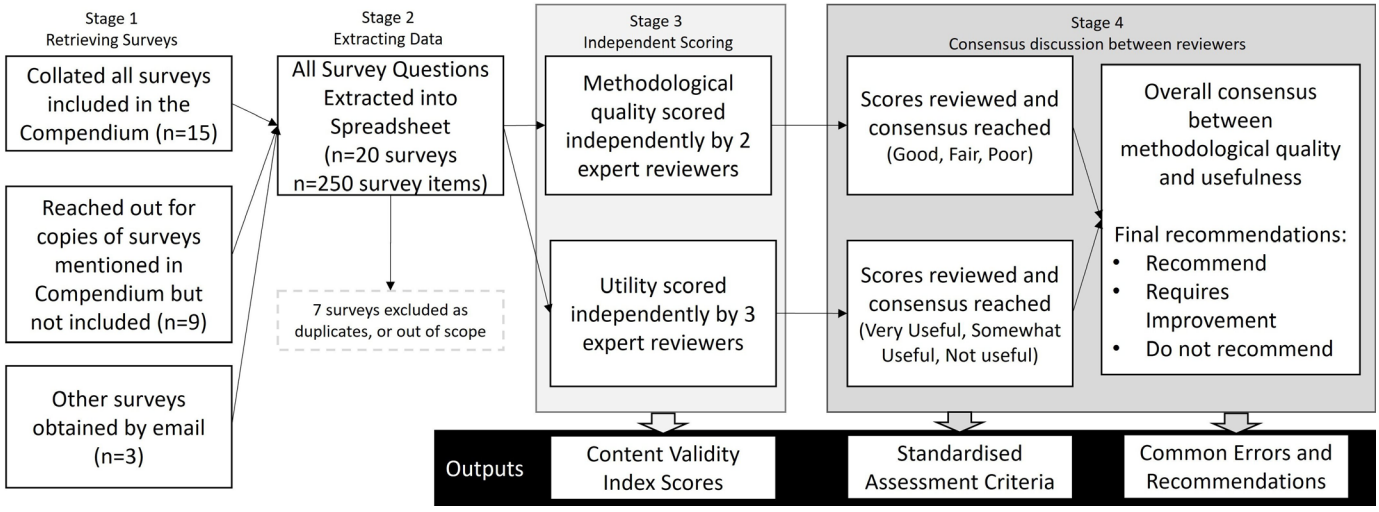

Figure 1 Process of survey collection and analysis.

areas for improvement. Surveys were not a prescribed form of measurement but were suggested as a potential method to understand patient and staff perceptions to inform improvement work.

Review of surveys

An overview of the process of review of surveys is shown in figure 1.

Data collection

Surveys were collated from all CF referring and lung transplant sites. As part of established knowledge sharing procedures, teams shared copies of surveys in the collaborative 'compendium' (a 280-page document containing details of activities and lessons learned of all sites.) This document was searched for surveys that were included in full or mentioned within site reports. Where surveys were mentioned but not included, copies were requested by email and obtained from sites. In addition, an email cascade through CF Quality coaches and clinical teams attempted to elicit any other surveys.

Data from the surveys were extracted into an excel spreadsheet including survey name and each individual survey item comprised a question and response options. This spreadsheet acted as the data collection tool for individual reviewers to add quality scores and comments.

Independent survey scoring

An interprofessional panel of five people reviewed and rated the individual survey item and the overall survey for content validity using a trichotomous rating scale. A trichotomous scale was chosen to capture the range of views held by the review panel, and to inform future action in relation to specific survey items and surveys.

Two of the reviewers had research expertise and focused on methodological quality (JER and JJ) scoring items as 'good', 'fair' or 'poor'. Three of the reviewers had topic specific expertise (relating to the clinical issue of CF Lung Transplant and coaching QI) (RM, RZ, FA) focused on the usefulness of the questions to inform learning, action and improvement, scoring each item or survey as either 'very useful', 'somewhat useful' or 'not useful'. JJ, RM and RZ had previously worked with the study sites in their role as QI coaches.

CVI calculation

The independent reviewer scores were used to calculate a CVI value for each survey.[17 22] The CVI calculates the proportion of items on an instrument that achieved a favourable rating by the reviewers: only a survey item that stands up to scrutiny by multiple reviewers is considered to be of good enough quality to be recommended for inclusion in a survey. The use of a diverse panel of reviewers is preferable for the CVI in order to identify as many 'problems' with a survey item as possible. It is expected that each individual reviewer will identify different types and amounts of problems, thus strengthening the quality of the review.[23–25] As a result, inter-rater reliability is expected to be low across reviewers: what matters is the identification of questions that no one can find fault with.

While the trichotomous scale was valuable in aiding the conduct of the review and for guiding actions in response to the review, a dichotomous scale is needed to calculate CVI. The CVI was therefore calculated by transforming the trichotomous reviewers scale into a dichotomous scale, where 'good' (top methodological score) or 'very useful' (the top utility score) equated to a favourable assessment (scoring 1.0) and 'fair', 'poor', 'somewhat useful' and 'not useful' equated an unfavourable assessment (scoring 0.0). The CVI value for each survey was calculated by averaging the favourability scores for each question across the five reviewers, and then averaging the cumulative questions scores for each survey. A CVI of greater than 0.8 (80%) would demonstrate a high level of agreement that the questions meet the reviewers standards, and a CVI below 0.8 suggests the survey does not adequately meet the reviewers standards and would require substantial further revision before further use.

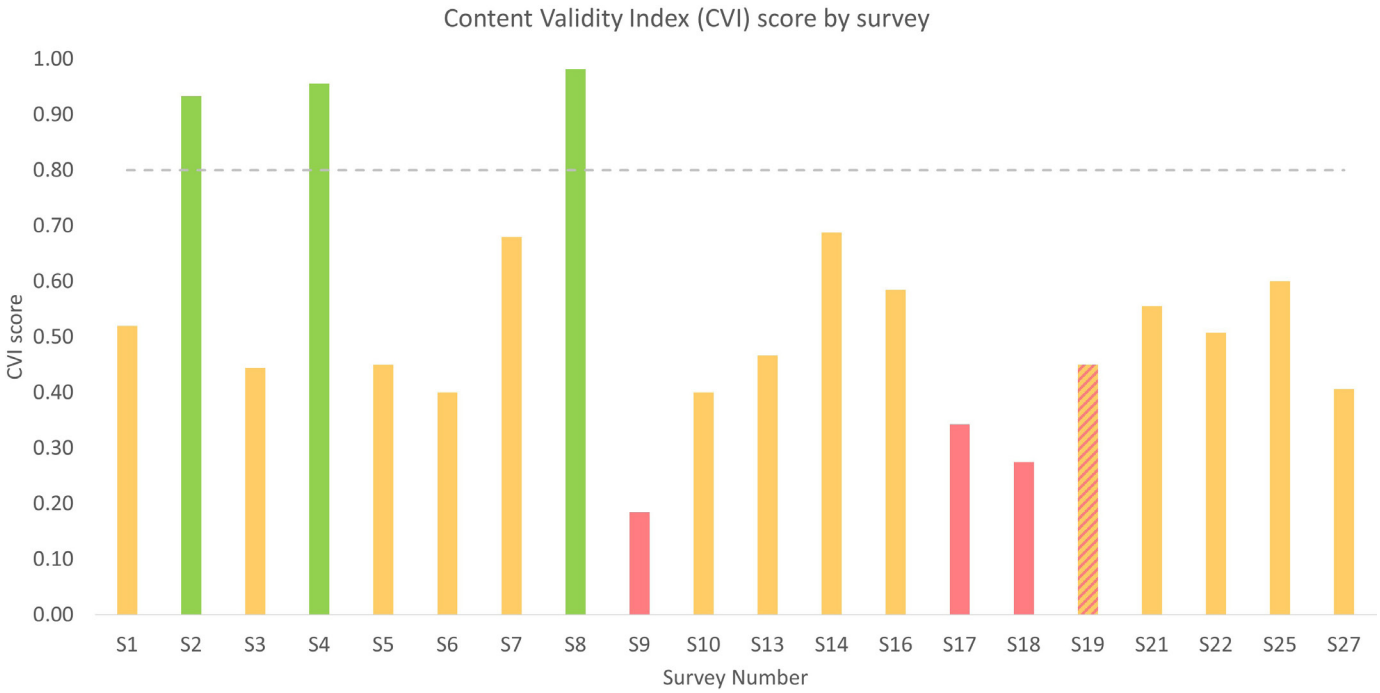

Figure 2 Bar chart shows the Content Validity Index (CVI) score by survey where 1.0 equals consensus on favourability of all survey items, and 0.0 equals consensus on unfavourability of all survey items. The minimum required agreement level of 0.80 is indicated by a grey dashed line. The colours of the bars represent the group agreement on the surveys arrived at after consensus discussion: green—recommend; amber—requires improvement; red—not recommended. Survey S19 was scored as requires improvement (useful topic but problems methodologically). However, a consensus decision was made to amend this score to not recommend due to the existence of a validated questionnaire exploring the same topic, making any potential improvements redundant. Hence, S19 is marked with red and amber stripes to show the agreed change to final recommendation status.

Consensus discussions

Normally CVI assessments stop after independent assessment as the focus is to identify items for which there is a favourable consensus, rather than to understand *why* different reviewers were unfavourable about discarded items. However, in this study we were interested in understanding why the decisions were reached, what could be learnt from the different reviewers' perspectives of the errors they identified. In addition from a practical point of view there was an opportunity to provide specific feedback to the collaborative programme to inform modification to individual survey items, and to inform future survey development.

The conversations in the consensus review meetings (9.5 hours in total) were rich discussions in which different viewpoints were considered and a shared understanding of error types was developed.

Consensus discussions first took place within the two separate review groups to reach consensus on the methodological and usefulness scoring. A final consensus discussion was held between all five reviewers to decide if the surveys and questions were both of good quality, and of high use to the clinical teams. Given a practical consideration was the recommendation of surveys to future sites in the collaborative, the consensus scores were 'recommend', 'requires improvement' and 'do not recommend'. This was a rigorous quality assessment, in that a question or survey could only be considered 'recommended' if it was both very useful and of good methodological quality.

Survey review data analysis

The consensus scores were analysed in Excel using descriptive statistics to analyse how many items had received each score, and to understand the percentage agreement between different reviewers and reviewer groups.

The error types were coded and analysed thematically in Excel, and then discussed and further refined by the authors until agreement was reached.

During these discussions the reviewers articulated and then iteratively developed a description of the survey quality criteria that they had used to assess the surveys, until consensus was achieved. This knowledge had been implicit during the independent scoring, informing each reviewers mental model for the decisions they reached. Through consensus discussion reviewers made these quality criteria explicit and formalised them as descriptions of each scoring quality criteria assessment category.

RESULTS

Surveys

In total, 27 surveys were identified. Seven surveys were removed because they were duplicates or because they were out of scope of the study (eg, surveys not developed

Table 1 Survey quality assessment criteria

	Methodological assessment	Utility assessment	Overall quality assessment
Survey item assessment			
Highest score	Good: the survey item is appropriate and question and response options are methodologically sound	Very useful: this survey item is appropriate to ask in a survey and is likely to provide beneficial learning to any site using the question	Recommended: survey item considered both of good methodological quality and very useful
	Fair: the survey item is appropriate to ask but the question and/or response options have methodological flaws	Somewhat useful: the survey item is appropriate to ask in a survey and may produce useful learning to specific sites or in specific instances	Requires improvement: survey item may have good or fair methodological quality and/or very useful or somewhat useful methodological quality but requires some improvement before it could be recommended
Lowest score	Poor: the survey item is inappropriate to ask in survey form	Not useful: the survey item is either inappropriate to ask in a survey, will not lead to useful learning, or would be better obtained by using another method	Not recommended: survey item is either methodologically and/or not useful or appropriate to ask in a survey format, or validated surveys already exists to assess this topic
Overall survey assessment			
Highest score	Good: the survey is conceptually sound, well designed for completion by respondents, and the majority of questions are scored as good or only require minor modifications	Very useful: the survey has a clear and focused learning purpose and the majority of the questions score as very useful	Recommended: survey considered both of good methodological quality and very useful
	Fair: the survey has some merits but requires significant improvement either to individual questions or to ensure the overall survey is sound and well designed	Somewhat useful: the survey contains some useful or very useful questions but requires substantial improvement either to individual questions or to the clarity or focus of the learning purpose of the survey	Requires improvement: survey may have good or fair methodological quality and/or very useful or somewhat useful methodological quality but requires some improvement before could be recommended
Lowest score	Poor: all questions are marked as poor, or there is a critical flaw in the concept of the overall survey	Not useful: all questions marked not useful	Not recommended: survey is either methodologically poor and/or not useful or appropriate to ask in a survey format, or validated surveys already exists to assess this topic

for the CF LTT collaborative). Twenty surveys were kept for full analysis which contained a total of 250 individual questions (average 12 questions per survey, ranging from 3 to 46 questions). Each of the 10 regions had developed at least one survey, with an average of two per region, and a maximum of four. Fourteen surveys were for patients, and six were for staff. Topics covered by the surveys included patient experience and satisfaction with services, patient preferences and expectations, staff experience, staff education and confidence, and intervention impact assessment (eg, educational training sessions).

Survey review findings

CVI scores

The CVI score is a composite measure of independent assessments of survey item quality.[17 22] The CVI score can range from 0 (poor quality, no survey items received favourable scores from reviewers) to 1 (high quality, all survey items received favourable scores from all reviewers). The average CVI score for the surveys was 0.54, ranging from 0.18 to 0.98. Three of the 20 surveys (15%) scored above the 0.8 quality benchmark. The scores for all three of the surveys were very high (S2=0.93; S4=0.96; S8=0.98) indicating strong agreement of favourability from the reviewers (see figure 2).

Consensus development of survey quality assessment criteria

Consensus was reached on all assessments. However, as expected, the percentage agreement between the independent reviewers prior to the consensus discussions was low.[26–28] The individual question assessment had only 54% agreement between the methodological reviewers, and two-way agreement between each pair of the topic specific reviewers ranged between 47% and 55% (three-way agreement between all topic specific reviewers was 36%). Overall survey assessment was higher with 75% agreement between methodological reviewers, and two way agreement between topic specific reviewers ranging between 40% and 60% (three-way agreement 35%). Prior to consensus discussions there was also very

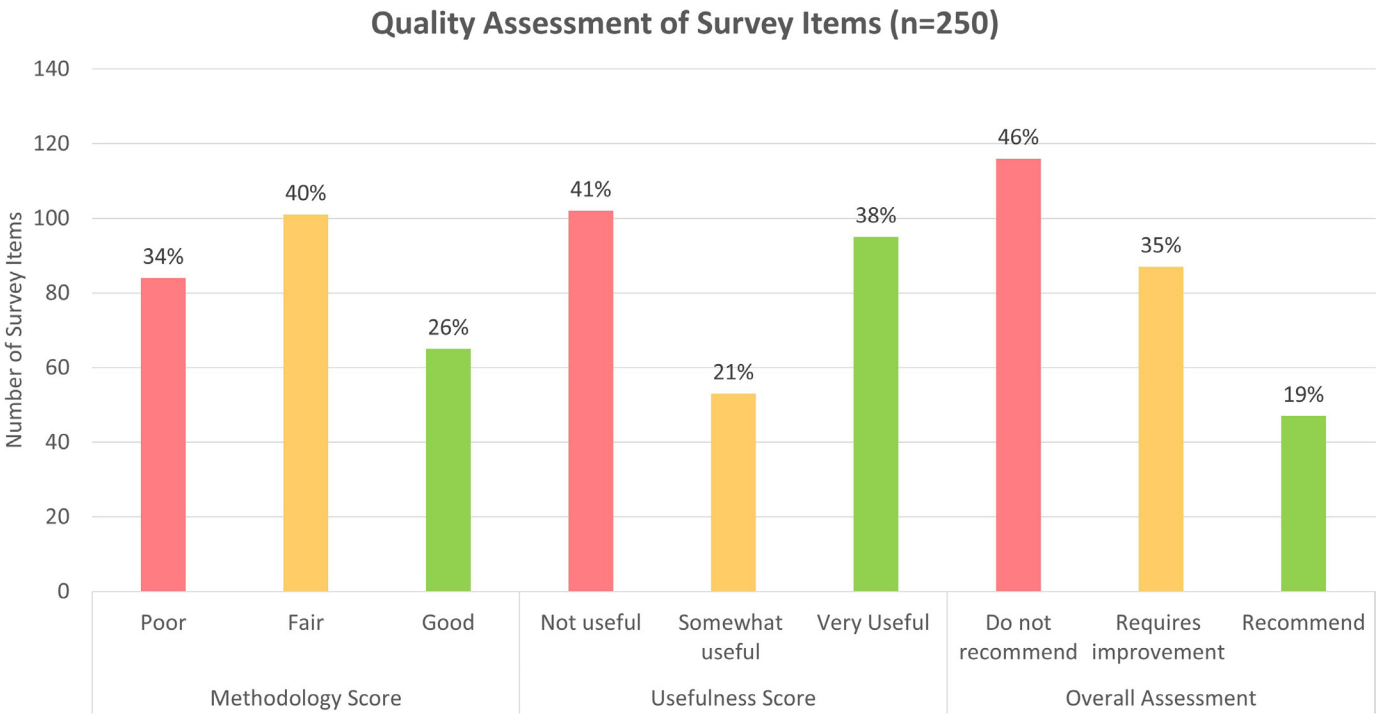

Figure 3 Quality assessment of survey items.

low agreement between reviewers scoring for methodological quality and usefulness (18% agreement between all five reviewers for individual questions, and 25% for overall surveys). This suggests that the reviewers brought a wide range of perspectives about what is problematic in a survey item, therefore strengthening the ability of the panel to exclude problematic questions.

Following completion of consensus discussion there was still only 59% agreement between methodological and usefulness reviewers suggesting that both groups of reviewers were indeed assessing different facets of quality in considering what was of methodological quality versus useful for QI.

As demonstrated by the CVI analysis, there was strong agreement on the assessment on the three surveys that scored above the 0.8 CVI threshold. In this subset of surveys there was 93% agreement between the methodological reviewers, and two-way agreement between the topic specific reviewers ranged between 86% and 97% (three-way agreement 86%).

There was a strong tendency for consensus discussions to downgrade the rating of an item in response to new and valid concerns being raised by the reviewers. For example, of the 114 questions disagreed on by the independent assessments by the methodological reviewers, only 5 (4%) ended up rated as 'good' in consensus discussions, whereas 68 (60%) were rated as fair, and 41 (36%) were rated as 'poor'. This is consistent with expectations of having a panel review survey items in being able to identify high-quality questions via elimination of any problematic questions.

From the consensus discussions the reviewers agreed on standard definitions for assessment criteria (table 1).

Survey quality assessment consensus scoring

Following the consensus discussion, of all of the individual survey items scored, only 26% (n=65) scored as good methodological quality, with 40% (n=101) and 34% (n=84) scoring as fair and poor methodological quality respectively. A higher proportion of the survey items were scored as very useful (38%, n=95), with 21% (n=53) and 41% (n=102) as somewhat useful and not useful, respectively. In the overall quality assessment of survey items (combining methodological score and usefulness score) only 19% (n=47) of survey items were recommended, with 35% (n=87) requiring improvements, and 46% (n=116) not recommended (figure 3).

For the overall survey only 20% (n=4) were scored as good methodologically, with 55% (n=11) scoring fair, and 25% (n=5) scoring poor. In terms of overall survey usefulness 30% (n=6) scored very useful, 50% (n=10) somewhat useful and 20% (n=4) not useful. For the overall quality assessment of surveys only 15% (n=3) were recommended, the large majority required improvement 65%, (n=13) and 20% (n=4) were not recommended (see online supplemental file 1).

Of the three surveys that were 'recommended', 25 of the 29 survey items were rated recommended, and four required improvement (S2: 9 survey items, 6 recommend, 3 required improvement; S4: 9 survey items, 8 recommended, 1 required improvement; S8, all 11 items were recommended).

Common errors

Following consensus discussions, thematic analysis was conducted on the reasons that surveys and survey items were scored unfavourably ('fair' or 'poor', 'somewhat

Table 2 Common error types in survey design

Common error type	Example survey item (source materials are referenced as S#:# to indicate survey number and survey item number)	Explanation (survey item may demonstrate multiple common errors but for the sake of brevity only the specific common error being demonstrated is discussed)
Ethics and appropriateness of questions and survey format		
Potential to cause distress	Question: how much have you thought about whom you would like to make or communicate decisions about your medical care on your behalf, if you are unable to do so yourself? Response: very much; somewhat; not very much *S3:10*	Surveys should be sensitive when asking patients to reflect on potentially distressing topics (eg, questions exploring the patient's mortality). If a patient had not previously considered such issues it could be distressing to first consider them in response to a survey question, and without the ability to share concerns or receive feedback and support from a trained professional. This is particularly important in the CF population as often considering these issues at a young age
Questions with complex answers better suited to discussion	1. Question: what are your worries or concerns about lung transplant? Response: open *S3:16* 2. Question: do you have any specific wishes about the types of medical treatment you want or do not want if you become very ill in the future? Response: yes; no; not sure *S3:11*	1. Such a question is unlikely to be fully understood by a survey method. Asking an open question puts the reliance on the respondent to answer in detail—which is often not the case leading to either reduced response rates or hard to interpret answers, and increases the skills and time required for analysis 2. On the other hand, limiting such questions to discrete options can also be artificial in limiting meaningful responses. Better suited to interview or discussion with patients
Risk of raising false expectations	Question: I would like to talk to a therapist or social worker about my worries and concerns about transplant. Response: strongly agree, agree, not sure, disagree, strongly disagree *S2:8*	This question sets up the expectation that a conversation will be had with a therapist or social worker if the patient states they would like one. This would require immediate follow-up to the survey. Surveys are not typically administered with the intention of immediate patient action—a survey could be completed, sit on someone's desk for a few weeks before the results are entered, and even when results are entered there may not be a clear pathway to action. Equally many surveys are conducted anonymously which would not allow a follow-up care response to be provided. Any question that asks the patients personal care preferences needs to make clear whether the survey will inform clinical care or not, and if so in what time period
Judging patients	1. Question: have you had a chance to read the transplant book? Response: yes (all), yes (some), no (too busy), no (too much), other *S10:4* 2. Question: I will only be seen in the transplant clinic when I am sick Response: true/false *S9:6*	1. This question judges the patient's actions a may trigger a feeling of guilt (if the patient has not done the desired action) which may in turn bias the response rates to demonstrate compliance 2. This question judges patient knowledge. It is not appropriate to make patients feel like their knowledge is being examined. In addition, there is a risk that patients remember false statements as facts, particularly if there is no feedback on whether the answers are correct, which could lead to the patient following incorrect and potentially harmful advice
Data already available	Question: how many months before the actual referral did the CF team first bring up transplantation as an option? Response: 0–6 months, 7–12 months, 13–24 months, >24 months *S22:2*	This question relies on patient recollection which might suffer from recall bias. Data relating to this question could be found in review of patient notes which would be more accurate and reduce the burden of survey response for the patients
Usefulness of survey items to inform learning or lead to action		
Burden of survey not proportionate to value of feedback	Question: what did you like or dislike about this new referral form? Response: open *S17:6*	This survey to staff relates to feedback on a new intervention designed as part of QI efforts. Only a small number of staff would have used the referral form making the value of structured survey responses limited. Informing improvement of the referral form would more likely benefit from a discussion exploring any issues to guide improvements. Reviewers felt it would be quicker, less work and more likely to achieve useful learning if brief discussions among staff were had instead
Judgement and blame	Question: overall, how well do patients appear to understand information communicated to them by the lung transplant team? Response: appear to understand completely; appear to have a partial understanding; appear to understand a little bit; appear to not understand information communicated *S18:5*	Using surveys to collate evidence of who to blame for problems experienced is unlikely to result in useful learning. This question is asking the staff respondent to speculate about a patient's opinion of another staff group. Not only is this unlikely to generate meaningful research data (subjective view about another person's opinion), it also appears to seek evidence to form a judgement about another groups work (they either did or did not do a good job communicating to the patients) which could then be used for blaming the other clinical team, rather than collecting constructive information on what areas could be improved

Continued

Table 2 Continued

Usefulness of survey items to inform learning or lead to action		
Self congratulatory	Question: how helpful was the book? Response: 10-not at all; 25-a little; 50-neutral; 75-moderately; 100-very *S9:3*	Looking for positive affirmation for the materials provided rather than seeking insights for learning and action. Patients may not be comfortable expressing a truthful review of work done by their team
Conflict with existing guidance	Question: would you prefer to wait to talk about lung transplant only if you become sick enough to refer for transplant evaluation? Response: yes; no; other (open) *S5:3*	This question conflicts with established national guidelines for CF that states that lung transplant should be discussed regularly with patients. This question is therefore not useful as (a) lots of research has been conducted on this topic already to inform national guidelines and establish best practice and (b) it potentially sets up patient expectations that they can choose when they think is best when that option may not be available to them
Using service industry questions in wrong context	1. Re Lung Transplant Education Session for patients: Question: I would recommend to others Response: strongly agree, agree, disagree, strongly disagree *S1:10* 2. Question: meeting room and facilities were adequate and comfortable Response: strongly agree, agree, disagree, strongly disagree *S1:6*	These questions were felt not to be useful for QI. 1. It seems inappropriate to ask patients if they would recommend a lung transplant education session—it is a necessary part of care for people who require treatment, not something that is chosen by the customer 2. Likewise, sessions are often delivered in hospital settings and choices of room and facilities are limited Neither question would be likely to lead to action or improvements

CF, cystic fibrosis; QI, quality improvement.

useful' or 'not useful'; n=203 questions and n=17 surveys). Twenty-three error types were identified that were grouped under six themes: ethics and appropriateness of questions and survey format; usefulness of survey items to inform learning or lead to action, and methodological issues with survey questions; survey response options; and overall survey design (see tables 2 and 3 for details of the error types with example survey items that demonstrate that error type).

Many of the survey items had multiple errors. For example, the question: would you ever consider a lung transplant? (response: yes; no; maybe) (*S3:5*) was considered unethical due to its potential to cause distress (eg, if this was the first time a patient was learning about lung transplant as a treatment option), not suitable for survey format due to the complexity of an answer (eg, lung transplant is a life changing decision with risk of mortality and any individual patient decision will be influenced by multiple complex factors) which would therefore be better to discuss in person instead; and given the simplicity of the response options, reviewers felt it was unlikely to generate useful learning that could be acted on by the improvement team.

DISCUSSION

Summary of findings

This study demonstrates the variable quality of surveys developed by local well intended QI teams, with only a small proportion of surveys and survey items being recommended for use by the review panel. These findings echo similar results highlighting the low quality use of quantitative measurement by local QI teams,[4] and suggests that, as with quantitative measurement, developing surveys is a highly technical task that requires time and expertise to develop reliable and meaningful measures.[29]

The consensus discussions of the surveys highlighted the complex, multifaceted and nuanced examination of detail required for rigorous assessment of the methodological quality and usefulness of survey items. While each individual assessor came with relevant expertise the diversity of views created a rich and dynamic discussion—where one reviewer had a favourable opinion of a survey item another reviewer might identify a problem reflecting their particular knowledge, expertise and experience. This emphasises the value of drawing on multiple expert perspectives to identify flaws with face and content validity. If multiple diverse expert opinions are unable to find flaw with a survey item it is likely to be of high validity.[17]

During the review of surveys it became clear that the issue of ethics was an important consideration; whether it was appropriate and sensitive to ask patients specific questions that might cause distress, and particularly in a survey form. This was felt to be of high importance to the CF patient population given the serious morbidity and mortality associated with the disease and the young average age of those being surveyed. These findings resonate with previous calls for the need of ethical oversight of QI activities.[30]

More broadly, reviewers questioned whether survey was the most appropriate method to obtain responses, particularly in instances where there were open ended questions with complex answers which would be better suited to interview. The reviewers concerns resonates with previous studies that have explored the value and limitations of surveys versus interviews in understanding patient

Table 3 Common error types in survey design

Common error type	Example survey item (source materials are referenced as S#:# to indicate survey number and survey item number)	Explanation (survey item may demonstrate multiple common errors but for the sake of brevity only the specific common error being demonstrated is discussed)
Methodological issues: survey questions		
Multiple possible meanings	Question: my CF centre discussed transplant with me *S14:6*	The word 'discussed' could mean different things to respondents—for example, one could consider a brief mention of the topic a discussion, while another could consider a discussion to be an in depth conversation dedicated to the topic
Subjective interpretation	Question: I know enough about lung transplant that I could explain it to others *S13:1*	The question is subject to interpretation by the respondent as to what 'know enough' means (what level of competency) and who 'others' are (eg, a senior professional, a patient, a non-specialist). The ambiguity in the question would make it impossible to interpret any responses as would be down to the subjective interpretation of the question by the respondent
Unspecific	Question: time allotted was sufficient/appropriate *S1:5*	The question does not specify what the time was allotted for, (eg, the entire consultation, or with specific team members) or what it was sufficient/appropriate for (eg, to understand what was to happen next, to have any questions answered)
Jargon	Question: I understand how my non-lung related health issues could affect success or recovery from transplant *S2:1*	'Non-lung' related health is not plain English and should be replace with more commonly used term such as 'overall health'
Leading	Question: after receiving the education folder and speaking with my CF team, I feel like I have a better understanding of the transplant process than I did before *S7:2*	The question is leading in suggesting that the experience was positive. This may bias responses
Double barrelled	Question: at the end of my first appointment at (clinic) I felt like my questions had been answered and there was a plan outlined *S14:12*	Assessing two different and independent items in a single question (whether patients' questions answered, and outlining of a plan) making it difficult to answer and interpret
Methodological issues: survey responses		
Non-standardised scale	Response: very comfortable; somewhat comfortable; not very comfortable; not comfortable at all; not sure *S3:13–14* Response: 10-not at all; 25-a little; 50-neutral; 75-moderately; 100-very *S9:3*	The scale design and anchor choice will influence respondents' ratings. Standardised response scales have gone through testing to ensure they have high validity. New scales are unlikely to be same level of rigour and so it is preferable that standardised response scale are used
No neutral option	Response: strongly agree, agree, disagree, strongly disagree *S1:3–10*	No neutral option is provided, for example, to say neither agree or disagree
Skewed/biased scale	Question: how much, if any, did the ACP meeting reduce anxiety about transplant? Response: none; not much; some; much; a great deal *S6:8*	The response options assume that the intervention, the ACP (advanced care planning) meeting, will only have neutral or positive options. Negative options also need to be included (did the intervention make anxiety worse)
Methodological issues: overall survey (does the overall survey make sense)		
Unnecessary duplication	Questions: this form was easy to complete This form was difficult to complete The instructions were clear and easy to understand *S17:2–4*	Multiple questions all addressing a similar issue may lead to response fatigue
Inconsistent or confusing response options	*S3* included 18Qs, 11 different response types including a 10-point scale, 2 different 5-point scales, 3 different 3-point scales, a 2-point scale, box-check answers and open ended questions	Increases the cognitive burden for the respondent, may reduce response rate or completion rate, and adds challenge to analysing and interpreting data
Complex subject areas addressed by existing validated surveys	Question: when thinking about your upcoming ACP meeting, how much (sic) days over the last 2 weeks have you felt nervous/anxious? Response: not at all, several days, more than half the days, nearly everyday *S6:2*	It is very difficult to measure complex issues such as anxiety accurately. Many existing validated surveys exist to explore such issues. It was also discussed that it would be very difficult for a respondent to separate anxiety caused by the meeting from other causes of anxiety

Continued

Table 3 Continued

Methodological issues: overall survey (does the overall survey make sense)		
First question filter	Question: did you receive the Lung Transplant Education Binder (Owners Manual)? Response: yes, no *S9:1*	The rest of the survey is irrelevant if the respondent has not received the Education Binder—suggests redundant question or inappropriate selection of survey respondents

CF, cystic fibrosis; QI, quality improvement.

experience to inform QI, and suggests the need to better educate and support QI teams to be aware of the variety of methods available, and when and how to best use such methods to inform learning.[3132] The burden of surveys for patients and staff was also considered, especially in relation to obtaining feedback on small Plan-Do-Study-Act tests of change where it was felt a face to face conversation between staff would be a more efficient and effective way of obtaining feedback, or where existing quantitative data could be analysed rather than relying on patient (or staff) recall. These findings echo the call made by Meyer *et al* to streamline the growing volume of metrics used to in QI, and consider the parsimony and burden of metrics across a project, service or organisation.[33]

The common error themes of ethics and appropriateness, and usefulness of survey items to inform learning or lead to action are a unique contribution of this research, reflecting issues of primary concern to the healthcare improvement community. The methodological themes on the other hand reflect well known errors in survey design.[17 34] The large number of methodological issues identified in the surveys suggests more is required to educate and support QI teams in developing quality surveys. All of these findings emphasise the importance of the normal steps of survey development which should include iterative cycles of testing and development to assess and refine survey items. Survey developers should strive to put themselves in the shoes of the patients (or staff) intended to use the survey, ideally engaging representative respondents in the design and development of surveys, and ensure that rigorous peer-review and piloting of the surveys take place. Importantly any surveys should have a clear purpose that directly links to improvement goals, and a clear plan for how any data collected will inform learning and action of the QI initiative.

Developing good quality surveys is a highly technical and time consuming task. Based on the evidence of this study we suggest that teams think carefully before deciding to embark on developing a new survey— in terms of the ethics, the appropriateness of the survey format, the burden on staff and patients, and the usefulness of data collected to directly lead to learning and action to improve quality. Only if teams are satisfied that their needs meet all of these criteria should they proceed, and then with caution ensuing there is sufficient time and resource to properly validate and pilot surveys before using them in practice. Given our experience in conducting this review we would advocate the establishment or oversight groups to ensure the appropriateness and quality of surveys being used within a collaborative which is a time efficient approach to support QI teams.

Methodological considerations and further research

This study conducted a rigorous assessment of the content validity of the collated surveys using multiple expert reviewers. Further research could be conducted on the recommended surveys including cognitive interviewing and survey piloting (including data collection, cleaning and analysis), which will likely identify further areas for improvement. In addition, this study was limited to the review of the actual surveys, and did not consider any (formal or informal) protocols for sampling, data collection, cleaning or analysis.

This study is limited in that it only considered the independent reviews of the surveys. Three of the five reviewers had prior engagement coaching a small number of teams within the collaborative. This was a strength in terms of the expert knowledge and contextually understanding of the relevance and utility of the surveys, but is also a limitation as a potential source of bias. The inclusion of five interprofessional reviewers who scored independently was used to reduce the risk of bias and increase the trustworthiness of the data. The QI teams that developed the surveys were not interviewed or observed to understand their perceptions of the survey instruments, or to establish the true utility of the surveys in a practice setting. Further research should be conducted to explore how locally developed surveys are perceived by QI teams, how they are used in practice, and what, if any, impact they have on the QI initiative. For example, it may well be that a survey with poor face validity nonetheless provokes useful insights and leads to meaningful change. Understanding the holistic value of surveys in a practical QI setting is critical to inform how much time and resources should be invested. Successful improvements have been observed in response to patient surveys where there is strong QI infrastructure and culture to support acting on survey results,[23 24] suggesting it is also important to understand the context in which surveys are being used.[25]

Implications for research and practice

This research was of direct value to the CF LTT Learning and Leadership Collaborative, highlighting which surveys were of high quality and which required improvements, which can help focus efforts on gathering information of interest. In addition, the identification of high-quality

surveys as a result of the review has the potential to save time of improvement teams, avoiding the need for teams to create, format and test their own surveys, when instead they can build on the experience of others in the collaborative. The reviewed surveys also serve to increase the quality of information gathered to inform improvement plans.

This research is also of value to other collaboratives and QI teams, providing advice to think rigorously about value and science of survey development. We recommend that leaders of larger QI collaboratives invest in expertise to support cross-site survey design and timely review to increase rigour of measurement and value to the clinical teams and evaluators while meeting the pace at which QI projects operate.

The methodology used in this paper can be applied to the review of surveys in other QI collaboratives to help improve the quality of survey instruments. The quality assessment criteria and common errors identified through consensus discussion provide explicit guidance to others developing or reviewing surveys in QI. This list is not intended to be comprehensive as it is solely reflective of the issues identified within this study, but can act as a valuable starting point to guide review and expansion. The comparative time invested by the reviewers was considered a valuable investment in complement to the extensive effort already invested by all of the local QI teams.

This research also has implications for other researchers in considering what it means to apply QI approaches with fidelity, how learning can be supported in complex systems, and with implications for the organisational resource infrastructure required to effectively support improvement in practice.

CONCLUSION

The development of surveys requires careful consideration, time and expertise. Before developing a survey, consideration should be given as to whether a survey is the most appropriate form of capturing information, and whether a survey best meets the need of the QI team and the targeted population. Once QI teams decide that a survey is the best way to gain knowledge of the patient and/or staff, multiple issues require considerations to ensure the rigorous design of the survey. There is a need to educate and support QI teams to adhere to good practice and avoid common errors, thereby increasing the value of surveys for evaluation and QI. The methodology, quality assessment criteria and common errors described in this paper can provide a useful resource for others, and highlights the value of having an oversight group to ensure the quality of surveys and to facilitate sharing of learning between improvement teams.

Patient and public involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Acknowledgements We would like to acknowledge the Cystic Fibrosis Foundation leadership and all members of the CF LTT LLC and RDN communities who made this research possible.

Contributors MMG conceived the study idea and is the guarantor of the research and publication. JER designed the study with input from JJ. JER led the conduct of the study with contributions made by JJ, RZ, RM and FA to data collection and analysis. All authors were involved with interpretation of the findings and writing of the manuscript. All authors have read and approved the final manuscript.

Funding This research was supported by award number GODFRE20QI2 from the Cystic Fibrosis Foundation. The authors gratefully acknowledge the financial support provided by the Cystic Fibrosis Foundation. MMG receives CF improvement collaborative grant funding from the CF Foundation. RZ, FA, RM, JJ are CF Quality team coach consultants for CF improvement collaboratives and JR was contracted as an independent improvement scientist consultant to conduct this research.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval IRB approval was obtained from University of New Hampshire IRB: UNH IRB-FY2021-60.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Julie E Reed http://orcid.org/0000-0002-9974-2017

REFERENCES

1 Boaden R, Harvey G, Moxham C, *et al*. *Quality improvement: theory and practice in healthcare*. NHS Institute for Innovation and Improvement, 2008.
2 Langley G, Moen R, Nolan K, *et al*. *The improvement guide: a practical approach to enhancing organizational performance, 2nd edn*. Jossey-Bass, 2009.
3 von Thiele Schwarz U, Nielsen K, Edwards K, *et al*. How to design, implement and evaluate organizational interventions for maximum impact: the Sigtuna Principles. *Eur J Work Organ Psychol* 2021;30:415–27.
4 Woodcock T, Liberati EG, Dixon-Woods M. A mixed-methods study of challenges experienced by clinical teams in measuring improvement. *BMJ Qual Saf* 2021;30:106–15.
5 Taylor MJ, McNicholas C, Nicolay C, *et al*. Systematic review of the application of the plan–do–study–act method to improve quality in healthcare. *BMJ Qual Saf* 2014;23:290–8.
6 Waring JJ, Bishop S. Lean healthcare: rhetoric, ritual and resistance. *Soc Sci Med* 2010;71:1332–40.
7 Berenholtz SM, Needham DM, Lubomski LH, *et al*. Improving the quality of quality improvement projects. *Jt Comm J Qual Patient Saf* 2010;36:468–73.
8 Pronovost PJ, Berenholtz SM, Goeschel CA. Improving the Quality of Measurement and Evaluation in Quality Improvement Efforts. *Am J Med Qual* 2008;23:143–6.
9 Mountford J, Shojania KG. Refocusing quality measurement to best support quality improvement: local ownership of quality measurement by clinicians: Table 1. *BMJ Qual Saf* 2012;21:519–23.

10 Harvey G, Wensing M. Methods for evaluation of small scale quality improvement projects. *Qual Saf Health Care* 2003;12:210–4.
11 Al-Abri R, Al-Balushi A. Patient satisfaction survey as a tool towards quality improvement. *Oman Med J* 2014;29:3–7.
12 Manacorda T, Erens B, Black N, *et al*. The Friends and Family Test in general practice in England: a qualitative study of the views of staff and patients. *Br J Gen Pract* 2017;67:e370–6.
13 Ahmed F, Burt J, Roland M. Measuring Patient Experience: Concepts and Methods. *Patient* 2014;7:235–41.
14 Ginsburg LR, Tregunno D, Norton PG, *et al*. "Not another safety culture survey": using the Canadian patient safety climate survey (Can-PSCS) to measure provider perceptions of PSC across health settings. *BMJ Qual Saf* 2014;23:162–70.
15 Grogan S, Conner M, Norman P, *et al*. Validation of a questionnaire measuring patient satisfaction with general practitioner services. *Qual Health Care* 2000;9:210–5.
16 Godfrey M. *Improvement Capability at the Front Lines of Healthcare: Helping through Leading and Coaching*. Jonkoping University, 2013.
17 Kelly K. *Measurement Made Accessible: A research approach using qualitative, quantitative & quality improvement methods*. 1999.
18 Smith PJ, Dunitz JM, Lucy A, *et al*. Incorporating patient and caregiver feedback into lung transplant referral guidelines for individuals with cystic fibrosis—Preliminary findings from a novel paradigm. *Clinical Transplantation* 2020;34. 10.1111/ctr.14038 Available: https://onlinelibrary.wiley.com/toc/13990012/34/10
19 Kilo CM. A framework for collaborative improvement: lessons from the Institute for Healthcare Improvement's Breakthrough Series. *Qual Manag Health Care* 1998;6:1–13.
20 Godfrey MM, Oliver BJ. Accelerating the rate of improvement in cystic fibrosis care: contributions and insights of the learning and leadership collaborative. *BMJ Qual Saf* 2014;23 Suppl 1:i23–32.
21 Nelson E, Batalden P, Godfrey M. *Quality by design: a clinical microsystems approach*. John Wiley & Sons, 2007.
22 Sangoseni O, Hellman M, Hill C. Development and Validation of a Questionnaire to Assess the Effect of Online Learning on Behaviors, Attitudes, and Clinical Practices of Physical Therapists in the United States Regarding Evidenced-based Clinical Practice. *IJAHSP* 2013.
23 Davies E, Shaller D, Edgman-Levitan S, *et al*. Evaluating the use of a modified CAHPS® survey to support improvements in patient-centred care: lessons from a quality improvement collaborative. *Health Expectations* 2008;11:160–76. 10.1111/j.1369-7625.2007.00483.x Available: https://onlinelibrary.wiley.com/toc/13697625/11/2
24 Davies EA, Meterko MM, Charns MP, *et al*. Factors affecting the use of patient survey data for quality improvement in the Veterans Health Administration. *BMC Health Serv Res* 2011;11:334.
25 Reed JE, Kaplan HC, Ismail SA. A new typology for understanding context: qualitative exploration of the model for understanding success in quality (MUSIQ). *BMC Health Serv Res* 2018;18:584.
26 Olson K. An Examination of Questionnaire Evaluation by Expert Reviewers. *Field Methods* 2010;22:295–318.
27 Presser S, Blair J. Survey Pretesting: Do Different Methods Produce Different Results? *Soc Methodol* 1994;24:73.
28 DeMaio TJ, Landreth A. Examining expert reviews as a pretest method. In: Prüfer P, Fowler J, Jackson F, eds. *ZUMA-Nachrichten Spezial Band 9, questionnaire evaluation standards*. Mannheim, Germany: ZUMA, 2003: 60–73.
29 Woodcock T, Adeleke Y, Goeschel C, *et al*. A modified Delphi study to identify the features of high quality measurement plans for healthcare improvement projects. *BMC Med Res Methodol* 2020;20:8.
30 Taylor HA, Pronovost PJ, Sugarman J. Ethics, oversight and quality improvement initiatives. *Qual Saf Health Care* 2010;19:271–4.
31 Tsianakas V, Maben J, Wiseman T, *et al*. Using patients' experiences to identify priorities for quality improvement in breast cancer care: patient narratives, surveys or both? *BMC Health Serv Res* 2012;12:271.
32 Pope C, van Royen P, Baker R. Qualitative methods in research on healthcare quality. *Qual Saf Health Care* 2002;11:148–52.
33 Meyer GS, Nelson EC, Pryor DB, *et al*. More quality measures versus measuring what matters: a call for balance and parsimony. *BMJ Qual Saf* 2012;21:964–8.
34 Kelley K, Clark B, Brown V, *et al*. Good practice in the conduct and reporting of survey research. *Int J Qual Health Care* 2003;15:261–6.