BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

A modified Delphi study to identify the features of high quality measurement plans for healthcare improvement projects

Thomas Woodcock[1*], Yewande Adeleke[1], Christine Goeschel[2], Peter Pronovost[3] and Mary Dixon-Woods[4]

Abstract

Background: The design and execution of measurement in quality improvement (QI) initiatives is often poor. Better guidance on "what good looks like" might help to mitigate some of the problems. We report a consensus-building process that sought to identify which features are important to include in QI measurement plans.

Methods: We conducted a three-stage consensus-building approach: (1) identifying the list of features of measurement plans that were potential candidates for inclusion based on literature review and the study team's experience; (2) a two-round modified Delphi exercise with a panel of experts to establish consensus on the importance of these features; and (3) a small in-person consensus group meeting to finalise the list of features.

Results: A list of 104 candidate questions was generated. A panel of 19 experts in the Delphi reviewed these questions and produced consensus on retaining 46 questions in the first round and on a further 22 in the second round. Thematic analysis of open text responses from the panellists suggested a number of areas of debate that were explicitly considered by the consensus group. The exercise yielded 74 questions (71% of 104) on which there was consensus in five categories of measurement relating to: design, data collection and management, analysis, action, and embedding.

Conclusions: This study offers a consensus-based view on the features of a good measurement plan for a QI project in healthcare. The results may be of use to QI teams, funders and evaluators, but are likely to require further development and testing to ensure feasibility and usefulness.

Keywords: Measurement, Quality improvement, Quality measurement, Delphi technique

Background

Prospective measurement of quality of care over time, known as *measurement for improvement*, is a defining feature of many quality improvement (QI) approaches, [1, 2] important for monitoring systems, assessing progress, and generating feedback [3]. Given its influence on the decisions and behaviours of staff, improvement teams, hospital leaders, and policy-makers, measurement quality, validity, and accuracy, analysis, and presentation are all critical. However, despite some published guidance [4, 5], the standard of measurement in QI initiatives is highly variable [6–8].

Current practice in measurement for improvement compares unfavourably with clinical trials, where high-quality measurement is recognised as a priority and accordingly is expertly led, is well-resourced, and has clear protocols for data collection. By contrast, QI teams often (albeit not always) may seem to lack capability and capacity to plan and conduct appropriate measurement: [6, 9] they often have to resort to locally designed and poorly validated measures, with data collection and analysis undertaken amid the messy realities of clinical practice [7]. These difficulties, along with the perception by some that QI measurement does not need to be rigorous, have been implicated in lack of progress and low investment in improving measurement standards [4, 8].

* Correspondence: thomas.woodcock99@imperial.ac.uk
[1]NIHR ARC Northwest London, Imperial College, Reynolds Building, St. Dunstan's Road, London W6 8RP, UK
Full list of author information is available at the end of the article

We propose that it is possible to mitigate some of the problems associated with measurement for improvement through meticulous and well-informed planning. The production of a written measurement plan is analogous to a research protocol: it allows QI teams to develop and communicate intentions and reach a shared understanding of how the impact of a QI initiative will be measured. This is critical, because the choices made at the planning stage (conscious or otherwise) have long-lasting implications. For example, if a team does not plan to establish baseline levels of variables, it risks not being able to evaluate the outcome of the initiative objectively.

What a QI measurement plan should include, however, has not been systematically established: little research has focused on the standards that should apply to planning for measurement for improvement. The available resources are predominantly textbooks and guides developed by organisations that support or fund QI teams. Examples include Health Quality Ontario's Measurement Plan Tool, a checklist focusing on the data collection process [10]; NHS Scotland's QI Hub Measure Plan and Data Collection Forms [11], developed based on a framework that is informed by Robert Lloyd's Quality Measurement Road Map [12]; and the NHS Elect's Measurement Checklist, which is based on its seven steps to measurement for improvement [13]. Some focus solely on a specific subset of issues, such as design of measurement. None has been developed through a formal consensus process.

In this article, we aim to address the void in guidance on planning measurement for improvement by reporting a consensus-building process to identify which features are important to include in a QI measurement plan. We do not seek to establish standards for measurement, but instead to identify the features that might benefit from standards being set, with the aim of supporting planning and review.

Methods

Guided by a Steering Group (MDW, PP, CG), we conducted a three-stage consensus-building approach: (1) identifying the list of features of measurement plans that were potential candidates for inclusion on the basis of importance (September 2015–February 2016); (2) conduct of a modified Delphi exercise (March–May 2016); and (3) an in-person consensus group meeting to finalise the list of features (8 December 2016).

Stage 1: Identifying the candidate features to include in the measurement plan

We generated a list of possible candidate features that could be entered into the modified Delphi study. We drew on two sources to do this.

First, we reviewed the existing literature on good (and bad) practice in measurement for improvement, including both peer-reviewed and grey literature. We started with articles recommended by members of the steering group, and then checked the reference list of these articles to identify other relevant articles. We read the articles and evaluation reports, identifying features that were mentioned either as good practice, or as mistakes or pitfalls. We added articles cited by the initial list, until this ceased to yield further candidate features. The final list of articles reviewed comprised 22 journal articles [4, 7, 8, 14–32] and 17 reports and textbooks [10–12, 33–46].

Second, we drew on the experience of the core study team (TW, YA) who have supported over 50 QI initiatives over 7 years as part of the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care (CLAHRC), Northwest London programme. We drew in particular on knowledge and experience of the challenges that teams encounter in seeking to do measurement, and of the CLAHRC's measurement planning process, which was developed to support QI teams [47].

Consistent with the vision outlined by Berenholtz et al. [8] of producing practical tools that would be helpful to frontline staff, we decided to frame each identified candidate feature as a question that would help practitioners to review the strengths and weaknesses of plans. This process resulted in a list of candidate questions for inclusion in the modified Delphi study. For each question, we used the supporting literature to draft an explanation of what the measurement plan might include (see Table 3).

Stage 2: Consensus-building study to select and refine the questions

A modified Delphi technique [48, 49] was used to build consensus on which questions from stage 1 were important to include as features of a QI measurement plan, using two rounds of rating and review by an expert panel over an eight-week period.

The inclusion criteria for the panel of experts invited to take part in the study were: experience of measurement planning for healthcare QI initiative(s) or specialist expertise and authority or influence in the science of improvement. Authors of this article and colleagues in the NIHR CLAHRC for Northwest London were excluded from this stage. Potential panellists were suggested by the steering group and core study team and were invited by email to take part in the study and asked to consent to participation. We asked these potential panellists to suggest others who they considered suitable to participate, who we then invited in addition, provided they met the inclusion criteria. We used Qualtrics survey software

(version 1.5) to create and administer the questionnaires. We aimed to achieve a panel of 11–30 members, sample sizes in this range are typical for the Delphi method and have been shown to be effective and reliable [48, 50].

Delphi round 1

Round 1 of our modified Delphi used a structured questionnaire comprising the candidate questions identified in stage 1. Panellists were shown each of the candidate questions alongside the explanatory text. They were then asked to vote to keep, remove, or modify the question, or to state that they had no opinion. We used categorical response options to ensure that panellists were clear about the consequences of their votes, to make interpretation clear, and to ensure that the results were actionable in terms of establishing a final list of questions at the end of the study.

For each candidate question, panellists were given the option to provide free-text comments to support their decision, or to suggest changes to the question. We also asked panellists to give their opinion on the overall structure and completeness of the list of questions. We conducted a simple thematic analysis of free-text responses to these open-ended questions by manually searching, reviewing, defining, and naming themes [51].

Consensus was set a priori at 75% agreement with any one of the available actions (keep, remove, or modify the question), consistent with previous Delphi studies reported in the literature [52]. Responses to the round 1 questionnaire were analysed by the research team during a two-week period. We excluded responses of 'no opinion' from percentage agreement calculations. Any question reaching consensus to keep or remove was not fed back into the round 2 questionnaire. In cases where the panel did not reach at least 75% agreement to keep or remove a question, we examined the comments and proposed either to remove the question or to amend it to improve the framing. These questions then formed the round 2 questionnaire.

Delphi round 2

We asked panellists to review the aggregated agreement percentages for each question as part of the round 2 questionnaire, alongside their previous individual ratings and a summary of the panel's comments from round 1. We then asked them to reconsider their rating using the same categories as in round 1. At the end of round 2 of the Delphi study, the analysis and feedback process was repeated.

Thus the output of the questionnaire stage was three sets of questions: a set to retain by consensus, a set to remove by consensus, and a set with no consensus to retain or to remove.

Stage 3: In-person consensus meeting to finalise the questions

To decide what to do with the set of questions with no consensus to keep or to remove, we held an in-person meeting of the core study team and the steering group, and we also invited panellists who had completed both rounds of the Delphi study. The aim of this meeting was to finalise the structure and content of the question list. The objectives were to:

- review and resolve proposals for questions that did not reach consensus through the questionnaire rounds, drawing on panellists' responses from stage 2
- address themes emerging from the Delphi panel's free-text responses on the overall structure and completeness of the list of questions.

Based on the results of the Delphi study (both qualitative and quantitative), the study team proposed to retain or remove each of the remaining questions. The consensus group discussed these proposals in light of the panel voting and comments from stage 2, and accepted or rejected them, allowing the study team to finalise the question list. Any questions that reached consensus in stage 2 were not discussed in stage 3 as panel consensus was considered final.

Results

Stage 1: Identifying the candidate features to include in the measurement plan

We identified 104 candidate questions as potential features of measurement plans that would be relevant in reviewing strengths and weaknesses (Additional file 1: Appendix 1). We identified five high-level categories of questions: design of measurement, data collection and management, analysis, action, and embedding. These categories were further divided into ten subcategories.

Stage 2: Consensus-building study to select and refine the questions

We invited 76 experts who met the selection criteria. Figure 1 shows the flow of participants and questions through stages 2 and 3 of the study. Of the 23 panellists who consented to take part, 19 completed both rounds of the Delphi study (Table 1).

At the end of round 1, the panel had reached consensus on keeping 46 (44%) questions. These questions were included in the final content and were not entered into round 2. The remaining 58 questions were amended based on the panel's comments and suggestions and re-entered into round 2.

At the end of round 2, the panel had reached consensus on 24 (41%), agreeing to keep 22 and remove two

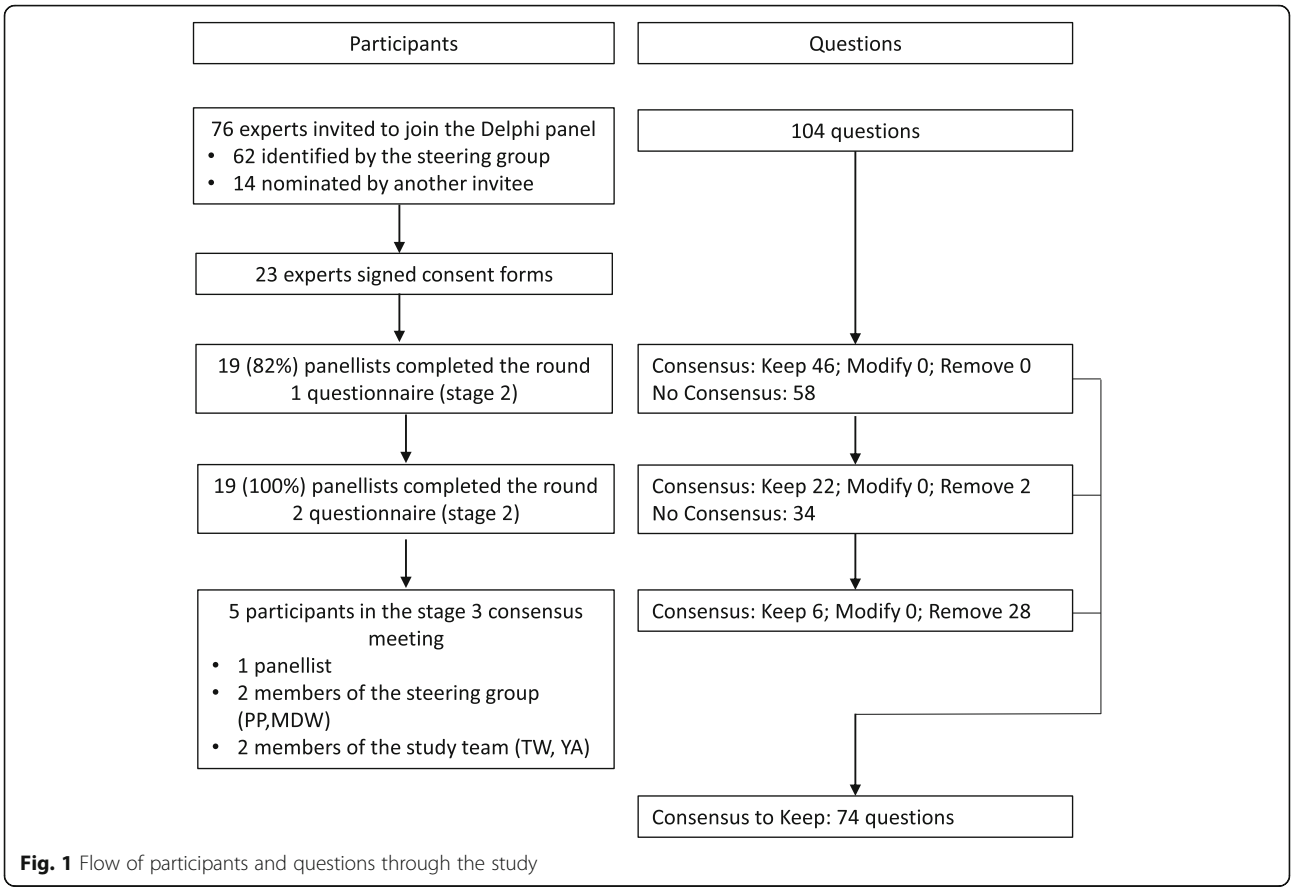

Fig. 1 Flow of participants and questions through the study

Table 1 Demographic characteristics of panellists in the modified Delphi study

	Panellists (*n* = 19)
Gender	
Male	12 (63%)
Female	7 (36%)
Country of employment	
United Kingdom	10 (53%)
United States of America	6 (32%)
Australia	3 (16%)
Role based on self-reported job titles	
Manager	8 (42%)
Academic researcher	6 (32%)
QI expert	3 (16%)
Nurse	1 (5.3%)
Doctor	1 (5.3%)
Experience in healthcare QI measurement planning	
Mean number of years	12 years

Data are number (%) unless otherwise stated. Note that some panellists held multiple roles in addition to that relating to their job title

questions. Thus at the end of stage 2, 70 (67%) of the 104 candidate questions met the predefined 75% agreement level: 68 to keep and two to remove. The number of questions in each subcategory reaching the specified consensus level is shown in Table 2. The panel did not reach consensus on 34 questions.

Stage 3: In-person consensus meeting to finalise the questions

For the 34 questions with no consensus at the end of stage 2, the core study team proposed to remove 27 questions and keep seven questions based on the Delphi panel's comments. During the stage 3 in-person consensus meeting, the group (YA, TW, PP, MDW, and one Delphi panellist) agreed with removal of all 27 questions proposed as meriting removal, and with six of the seven proposals to keep questions – the seventh was deemed to be sufficiently covered by another question. Therefore in this stage six questions were kept and 28 removed.

In response to comments from the Delphi panel, the consensus group also made minor revisions to the phrasing of the six questions that were kept. This resulted in the final list of 74 questions (71% of the original 104) shown in Additional file 1: Appendix 2, along with an explanation against each question of what it

Table 2 Number (%) of questions in each subcategory through each stage of the study

Stage 1		Stage 2				Stage 3			
Category	Subcategory	Round 1		Round 2		Total questions	Questions removed	Questions kept	Final number of questions in each subcategory
		Total questions	Questions with consensus	Total questions	Questions with consensus				
Design of measurement	Aim	7	3 (43%)	7	5 (71%)	2	1	1	10
	Measure set	13	4 (31%)	9	4 (44%)	5	4	1	8
	Operational definition	29	16 (55%)	11	2 (18%)	9	8	1	19
Data collection and management	Data collection process	13	8 (62%)	5	0 (0%)	5	4	1	9
	Training in and embedding of consistent data collection	5	2 (40%)	3	0 (0%)	3	3	0	2
	Database design	4	1 (25%)	3	2 (67%)	1	1	0	3
	Outliers and missing data	3	0 (0%)	3	2 (67%)	1	1	0	2
Analysis	Planning the analysis	17	6 (35%)	10	3 (30%)	7	5	2	11
Action	Planning for action	4	2 (50%)	2	2 (100%)	0	0	0	4
Embedding	Planning for sustainability	9	4 (44%)	5	4 (80%) (2 to remove; 2 to keep)	1	1	0	6
Totals		104	46 (44%) (all to keep)	58	24 (41%) (22 to keep; 2 to remove)	34	28	6	74

Note that, following feedback from the panel, some questions changed subcategory after each round of the modified Delphi

means and why it is important. Two example questions are shown in Table 3.

The consensus group also discussed the themes emerging from the Delphi panel's free-text responses on the overall structure and completeness of the list of questions, which broadly concerned whether the questions should concern methodological validity or simply transparency; the number of questions; inclusion of general project-level questions; and whether the standards that apply to research measurement should also apply to QI.

Methodological validity and transparency of the measurement plan

The Delphi panellists' free-text responses highlighted some tensions about whether the measurement questions should seek solely to address the transparency of a plan (whether the methodological approach is clearly articulated) or should also seek to assess the quality of the approach to measurement. The consensus group concluded that at this stage in the development of the field, it would be difficult to add to the study's original goal of identifying the important features of a measurement plan to seek additionally to specify the means by which quality of the methods would be assessed, especially given the complexity and context-specificity of this task. For example, decisions about which analytical methods are most appropriate for dealing with missing data depend on factors such as the extent and nature of the missing data, and the relation to other variables in the analysis. Furthermore, the decision as to which approach is optimal may involve knowledge of advanced statistical concepts that may not be readily accessible to clinical teams.

Table 3 Example questions with explanations

Question 51: Is there a plan in place for the prospective (1) identification and (2) minimisation of missing data? [No; Yes]
Category 2: Data collection and management **Subcategory 7:** Outliers and missing data
Explanation: Data reviews, visual cues, and reminders can be used to identify missing data. It is important to carefully distinguish between data items that are not applicable versus missing data, and branching logic is useful for this. Methods to minimise missing data include database controls; review of the data collection tool by designated staff or an independent reviewer at the time of data entry; immediate reporting of problems to the data collection staff and project leaders – for example, if missing data are over a certain threshold, quality assurance review is needed etc. Missing data threaten the progress of QI initiatives and the validity of evaluation findings derived from them.
Question 61: Is the intended frequency of feedback of the analysis to the team stated? [No; Yes]
Category 3: Analysis **Subcategory 8:** Planning the analysis
Explanation: Continuous communication of ongoing evaluation results to stakeholders is important for a capable improvement initiative. Updating and reporting the measures can be done on a daily, weekly, monthly, quarterly, or yearly basis. The frequency should be sufficient to see a pattern and for quicker action on the system of interest. Monthly feedback of analysis is often best suited for appropriate outcome measures, whereas weekly (or more frequent) analysis is often preferable for key process measures. However, there can also be less frequently analysed measures of interest. For instance, smoking quit rates at 1 year, or effects of lifestyle changes, may take longer to manifest. For QI work, it is important that at least for a small number of measures (the improvement measures), the frequency of feedback is high – ideally at least weekly. Explicitly stating the frequency of feedback at the planning stage can help to surface and deal with potential barriers to effective feedback at an early stage.

Number of questions

The panel raised the issue of the number of questions, noting that while the questions were comprehensive in scope, there were a lot of them taken as a whole. Given that there were already 68 questions according to the pre-defined threshold for consensus, the consensus group was careful to keep only questions deemed essential from the remaining 34 questions under consideration at the consensus meeting.

Inclusion of general project-level questions

Some panellists commented that some questions could be seen as not specific to measurement, but instead as pertaining to general project issues – e.g. project management or governance. The group agreed to remove any questions not specific to measurement, except those that were essential for subsequent questions to make sense. Each of these more general questions had received some votes to keep, revealing the blurred boundary between measurement and other activities undertaken by QI teams.

Research and QI – methodological and practical considerations

Some panellists commented that certain questions seemed more appropriate for 'research' rather than QI initiatives. The experts in the consensus group commented that while there are valid differences in methodology appropriate for answering different types of questions, the purpose of this study was to help bring to QI the rigor that research work benefits from. For example, in making inferences from a random sample to a fixed population versus understanding whether a change has occurred in a process over time, one would use different statistical methods. In both cases, having a clear definition of the measures used and an understanding of the quality of data against the definitions is important for the integrity of the conclusions drawn from subsequent analysis.

Discussion

This article presents a consensus-building study aimed at identifying the important features of measurement

plans for healthcare QI projects. The result is a list of 74 questions that may support QI teams in identifying the features relevant to planning and reviewing transparency and completeness of measurement planning, along with explanations for each feature. It is one of the first formal studies of this area, synthesising the cumulative learning from literature on measurement and experience in the field with the expertise of 19 leaders in the field of measurement for improvement. The findings may be of value to QI teams (for example in identifying where expert statistical or methodological advice may be necessary) and to funders, designers, and evaluators of QI programmes, though they may require further development and evaluation.

This study has a number of strengths and limitations. The initial literature search, which formed one source of candidate features for entry into stage 2 of the study, was not a systematic review. It is therefore possible that some potential candidate features present in the literature were missed, and that some candidate features used in the study are not evidenced in the literature. Use of the modified Delphi technique in Stage 2 offered a number of advantages, preserving the anonymity of panellists and allowing unrestricted expression of opinions, and thus helping to reduce the influence of dominant personalities and the effect of panellists' status on results [49]. It allowed the panel to choose freely whether to keep, modify, or remove questions. It also permitted coverage of the full range of QI initiatives, rather than, for example, focusing on those based in hospitals or particular health conditions. However, the long list of questions that emerged – over 70 – offers insight into the complexity of measurement as an endeavour in QI work. Such complexity is a theme emerging in the study of QI more generally [53], but the number of questions may be a risk to feasibility for general use and, if unattainable, may risk alienating or demoralising teams. Face-to-face discussions in Stage 3 enabled decisions to be made where no consensus could be reached using the remote survey in stage 2, but may have been vulnerable to typical group norms and effects. The study has helped to identify requisite features of a good plan and whether they are articulated transparently, but has not addressed the issue of quality of methods. This may be a focus of future work.

Further research is also needed to understand the relative importance of the questions identified through this study, to allow prioritisation of resources in planning improvement, and to convert the long list of questions we have identified into a practically useful guide for QI teams. Such work might focus on systematic approaches to help teams develop measurement plans that are scientifically valid, practically feasible, and promote successful improvement. This may require, for example, presenting the study findings in a user-friendly format suitable for QI teams, perhaps through development of an interactive guidance and support tool to facilitate adoption of the findings of this study. Such a tool could also provide useful data on which areas of measurement planning are particularly challenging for QI teams, and therefore where systematic support might best be aimed. A prototype tool has been developed [54], and the authors plan to report its development in a subsequent article.

Conclusions

Existing checklists and templates to support measurement planning are not comprehensive and none has been developed through a formal expert consensus technique. QI teams may use the results of our study proactively to highlight areas where they need to seek additional expertise, or to develop their plans further. Further work may be needed to refine and test the tool to ensure feasibility and usefulness, and to ensure that QI teams are appropriately supported in developing and reviewing measurement plans.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-019-0886-6.

Additional file 1: Appendix 1. Table of candidate questions for measurement plan review. **Appendix 2.** Consensus results: table of measurement plan review questions.

Abbreviations

CLAHRC: Collaboration for Leadership in Applied Health Research and Care; QI: Quality Improvement

Acknowledgements

The authors would like to thank all 19 panellists who took part in the Delphi study: Dr. Maren Batalden, Dr. Jonathan Benn, Professor Sean Berenholtz, Mr. Kurt Bramfitt, Dr. Robert Brook, Dr. Andrew Carson-Stevens, Ms. Alison Cole, Dr. Jocelyn Cornwell, Mr. Mike Davidge, Dr. Daisy Goodman, Mr. Mike Holmes, Dr. Jill Marsteller, Mr. Lloyd Provost, Professor Sheree Smith, Dr. James Williamson, and four others who wished to remain anonymous. The authors would also like to thank the following people for their valuable contributions: Dr. Julie Reed, Dr. Alan Poots, Mr. Derryn Lovett, Dr. Mable Nakubulwa, Mr. Neil Stillman, Mr. David Sunkersing, and Katrina Brown for their comments on the draft manuscript and input in discussions.

Authors' contributions

TW obtained the funding. TW and YA formed the study team and designed and conducted all stages of the research, designed and administered the questionnaires, collated and analysed the data, and drafted the manuscript. TW supervised the work. CG, PP, and MDW formed the steering group. All authors contributed to the final manuscript. TW is the corresponding author.

Funding

This article presents independent research funded by the Health Foundation (MPAT award, TW Improvement Science Fellow award), hosted and supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) and now recommissioned as NIHR Applied Research Collaboration (ARC), Northwest London. Mary Dixon-Woods is supported by the Health Foundation's grant to the University of Cambridge for The Healthcare Improvement Studies (THIS) Institute. THIS Institute is supported by the Health Foundation – an independent charity committed to bringing about better health and health

care for people in the UK. MDW is a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0617-10026). The views expressed in this article are those of the authors and not necessarily those of the Health Foundation, NHS, NIHR, or the Department of Health and Social Care. The funders had no role in the study design, data collection, data analysis, interpretation of results, or writing the manuscript.

Availability of data and materials

The anonymised datasets analysed in this study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This study has been approved by the NHS Health Research Authority (IRAS 188851). Potential panellists were invited to participate in the Delphi study and received written information about the study. Panellists received the questionnaire after completing the consent form on Qualtrics. 23 panellists completed the consent form. Participation was completely voluntary. The information collected was confidential and the identity of the panellists was concealed in the results. Panellists remained anonymous to each other during the study.

Consent for publication

Not applicable. The manuscript does not include details relating to an individual person.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]NIHR ARC Northwest London, Imperial College, Reynolds Building, St. Dunstan's Road, London W6 8RP, UK. [2]Medstar Health, Institute for Quality and Safety, 10980 Grantchester Way, Columbia, MD, USA. [3]University Hospitals Management Services Center, 3605, Warrensville Center Road, Shaker Heights, OH 44122, USA. [4]THIS Institute (The Healthcare Improvement Studies Institute), University of Cambridge, Cambridge, UK.

Received: 24 April 2019 Accepted: 13 December 2019
Published online: 14 January 2020

References

1. Boaden R, Harvey G, Moxham C, Proudlove N. Quality improvement: theory and practice in healthcare. NHS Institute for Innovation and Improvement. 2008. https://www.england.nhs.uk/improvement-hub/publication/quality-improvement-theory-practice-in-healthcare/. Accessed 8 Feb 2016
2. Brandrud AS, Schreiner A, Hjortdahl P, Helljesen GS, Nyen B, Nelson EC. Three success factors for continual improvement in healthcare: an analysis of the reports of improvement team members. BMJ Qual Saf. 2011;20:251–9. https://doi.org/10.1136/bmjqs.2009.038604.
3. Parand A, Benn J, Burnett S, Pinto A, Vincent C. Strategies for sustaining a quality improvement collaborative and its patient safety gains. Int J Qual Heal Care. 2012;24:380–90. https://doi.org/10.1093/intqhc/mzs030.
4. Needham DM, Sinopoli DJ, Dinglas VD, Berenholtz SM, Korupolu R, Watson SR, et al. Improving data quality control in quality improvement projects. Int J Qual Health Care. 2009;21:145–50. https://doi.org/10.1093/intqhc/mzp005.
5. Dixon N. Proposed standards for the design and conduct of a national clinical audit or quality improvement study. Int J Qual Heal Care. 2013. https://doi.org/10.1093/intqhc/mzt037.
6. Dixon-Woods M, Martin G, Tarrant C, Bion J, Goeschel C, Pronovost P, et al. Safer clinical systems: evaluation findings. 2014. http://www.health.org.uk/publication/safer-clinical-systems-evaluation-findings. Accessed 13 Dec 2018
7. Dixon-Woods M, Leslie M, Bion J, Tarrant C. What counts? An ethnographic study of infection data reported to a patient safety program. Milbank Q. 2012;90:548–91. https://doi.org/10.1111/j.1468-0009.2012.00674.x.
8. Berenholtz SM, Needham DM, Lubomski LH, Goeschel CA, Pronovost PJ. Improving the quality of quality improvement projects. Jt Comm J Qual Patient Saf. 2010;36:468–73 http://europepmc.org/abstract/MED/21548508. Accessed 28 Nov 2014.
9. Woodcock T, Liberati EG, Dixon-Woods M. A mixed-methods study of challenges experienced by clinical teams in measuring improvement. BMJ Qual Saf. 2019:bmjqs-2018-009048. https://doi.org/10.1136/bmjqs-2018-009048.
10. Health Quality Ontario. Measurement plan instruction sheet. http://www.hqontario.ca/Portals/0/Documents/qi/qi-measurement-plan-instruction-sheet-ac-en.pdf. Accessed 13 Dec 2018
11. NHS Education for Scotland. Framework for developing a measurement plan. 2012. http://www.qihub.scot.nhs.uk/media/340181/2012-06-15_measurement_improvement_journey_process.pdf. Accessed 13 Dec 2018
12. Lloyd RC. Quality health care: a guide to developing and using indicators: Jones and Bartlett Publishers, Inc; 2004.
13. Davidge M, Holmes M, Shaw A, Should S, Tite M. Guide to measurement for improvement. 2015. www.nhselect.nhs.uk/file_download.aspx?id=16359. Accessed 13 Dec 2018
14. Reed JE, McNicholas C, Woodcock T, Issen L, Bell D. Designing quality improvement initiatives: the action effect method, a structured approach to identifying and articulating programme theory. BMJ Qual Saf. 2014;23:1040–8. https://doi.org/10.1136/bmjqs-2014-003103.
15. Feldstein AC, Glasgow RE. A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. Jt Comm J Qual Patient Saf. 2008;34:228–43. https://doi.org/10.1016/S1553-7250(08)34030-6.
16. Diette GB, Rubin HR, Pronovost P. The advantages and disadvantages of process-based measures of health care quality. Int J Qual Heal Care. 2001;13:469–74. https://doi.org/10.1093/intqhc/13.6.469.
17. McGlynn EA, Asch SM. Developing a clinical performance measure. Am J Prev Med. 1998;14:14–21. https://doi.org/10.1016/S0749-3797(97)00032-9.
18. Pronovost PJ, Nolan T, Zeger S, Miller M, Rubin H. How can clinicians measure safety and quality in acute care? Lancet. 2004;363:1061–7. https://doi.org/10.1016/S0140-6736(04)15843-1.
19. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf. 2015;25:986–92. https://doi.org/10.1136/bmjqs-2015-004411.
20. Pronovost PJ, Miller MR, Wachter RM. The GAAP in quality measurement and reporting. JAMA. 2007;298:1800–2. https://doi.org/10.1001/jama.298.15.1800.
21. Palmer RH, Miller MR. Methodologic challenges in developing and implementing measures of quality for child health care. Ambul Pediatr. 2001;1:39–52. https://doi.org/10.1367/1539-4409(2001)001<0039:MCIDAI>2.0.CO;2.
22. Terris DD, Litaker DG. Data quality Bias: an Underrecognized source of misclassification in pay-for-performance reporting? Qual Manag Healthc. 2008;17. https://doi.org/10.1097/01.QMH.0000308634.59108.60.
23. Bowman CC, Sobo EJ, Asch SM, Gifford AL. Measuring persistence of implementation: QUERI series. Implement Sci. 2008;3:21. https://doi.org/10.1186/1748-5908-3-21.
24. Speroff T, O'Connor G. Study designs for PDSA quality improvement research. Qual Manag Health Care. 2004;13:17–32.
25. Nothacker M, Stokes T, Shaw B, Lindsay P, Sipilä R, Follmann M, et al. Reporting standards for guideline-based performance measures. Implement Sci. 2016;11:6. https://doi.org/10.1186/s13012-015-0369-z.
26. Perla RJ, Provost LP, Murray SK. Sampling considerations for health care improvement. Qual Manag Health Care. 22:36–47. https://doi.org/10.1097/QMH.0000000000000042.
27. Ovretveit J. Evaluation of quality improvement programmes. Qual Saf Heal Care. 2002;11:270–5. https://doi.org/10.1136/qhc.11.3.270.
28. Brook RH, McGlynn EA, Cleary PD. Measuring quality of care. N Engl J Med. 1996;335:966–70. https://doi.org/10.1056/NEJM199609263351311.
29. Provost LP. Analytical studies: a framework for quality improvement design and analysis. BMJ Qual Saf. 2011;20(Suppl 1):i92–6. https://doi.org/10.1136/bmjqs.2011.051557.
30. Pope C, van Royen P, Baker R. Qualitative methods in research on healthcare quality. BMJ Qual Saf. 2002;11:148–52. https://doi.org/10.1136/qhc.11.2.148.
31. Dixon-Woods M, Leslie M, Tarrant C, Bion J. Explaining matching Michigan: an ethnographic study of a patient safety program. Implement Sci. 2013;8:70. https://doi.org/10.1186/1748-5908-8-70.
32. Mattke S. When should measures be updated? Development of a conceptual framework for maintenance of quality-of-care measures. BMJ Qual Saf. 2008;17:182–6. https://doi.org/10.1136/qshc.2006.021170.
33. Langley GJ, Moen R, Nolan KM, Nolan TW, Norman CL, Provost LP. The improvement guide: a practical approach to enhancing organizational performance. 2nd ed: Jossey-Bass; 2009.

34. Provost LP, Murray SK. The health care data guide: learning from data for improvement. 1st ed: Wiley; 2011.
35. How to Improve. Institute for Healthcare Improvement. http://www.ihi.org/resources/Pages/HowtoImprove/default.aspx. Accessed 29 Sep 2015
36. Batley S, Bevan H, Cottrell K, Christian D, Davidge M, Easton J, et al. Improvement leaders' guide: measurement for improvement. 2007. https://www.england.nhs.uk/improvement-hub/publication/improvement-leaders-guide-measurement-for-improvement-process-and-systems-thinking/. Accessed 29 Sep 2015
37. Romano P, Hussey P, Ritley D. Selecting quality and resource use measures: a decision guide for community quality collaboratives. Agency for Healthcare Research and Quality. 2010. http://www.ahrq.gov/professionals/quality-patient-safety/quality-resources/tools/perfmeasguide/index.html. Accessed 30 Sep 2015
38. Pencheon D. The good indicators guide: understanding how to use and choose indicators. 2007. https://www.england.nhs.uk/improvement-hub/publication/the-good-indicators-guide-understanding-how-to-use-and-choose-indicators/. Accessed 29 Sep 2015
39. Altman Dautoff D, Van Borkulo N, Daniel D. Safety net medical home initiative. 2013. http://www.safetynetmedicalhome.org/sites/default/files/Implementation-Guide-QI-Strategy-1.pdf. Accessed 29 Sep 2015
40. Raleigh V, Foot C. Getting the measure of quality: opportunities and challenges. The King's Fund: London, UK; 2010. https://www.kingsfund.org.uk/publications/getting-measure-quality
41. Hughes RG. Tools and strategies for quality improvement and patient safety. In: Patient safety and quality: an evidence-based handbook for nurses. Rockville, MD: Agency for Healthcare Research and Quality (US); 2008. https://www.ncbi.nlm.nih.gov/books/NBK2682/.
42. Carey RG, Lloyd RC. Measuring quality improvement in healthcare: a guide to statistical process control applications. American Society for Quality: Milwaukee, WI; 1995.
43. US Department of Health and Human Services: Health Resources and Services Administration. Quality improvement. 2011. https://www.hrsa.gov/sites/default/files/quality/toolbox/508pdfs/qualityimprovement.pdf. Accessed 29 Sep 2015
44. Schoenbaum SC, Sundwall DN. Using clinical practice guidelines to evaluate quality of care. US Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1995.
45. Wheeler DJ. Making sense of data: SPC for the service sector: SPC Press; 2003.
46. Maher L, Gustafson D, Evans A. Sustainability model and guide: NHS Institute for Innovation and Improvement; 2010.
47. Curcin V, Woodcock T, Poots AJ, Majeed A, Bell D. Model-driven approach to data collection and reporting for quality improvement. J Biomed Inform. 2014;52:151–62. https://doi.org/10.1016/j.jbi.2014.04.014.
48. Dalkey NC. The Delphi method: an experimental study of group opinion. Santa Monica, CA: RAND Corporation; 1969. https://www.rand.org/pubs/research_memoranda/RM5888.html. Accessed 19 Oct 2018
49. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ. 1995;311:376–80. https://doi.org/10.1136/BMJ.311.7001.376.
50. Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. BMC Med Res Methodol. 2005;5:37. https://doi.org/10.1186/1471-2288-5-37.
51. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3:77–101. https://doi.org/10.1191/1478088706qp063oa.
52. Keeney S, Hasson F, McKenna H. Consulting the oracle: ten lessons form using the Delphi technique in nursing research. J Adv Nurs. 2006;53:205–12. https://doi.org/10.1111/j.1365-2648.2006.03716.x.
53. Reed JE, Howe C, Doyle C, Bell D. Simple rules for evidence translation in complex systems: a qualitative study. BMC Med. 2018;16:92. https://doi.org/10.1186/s12916-018-1076-9.
54. Imperial College London. Impala QI. 2019. https://impalaqi.com. Accessed 4 Oct 2019

Publisher's Note