BMJ Open

A balanced approach to identifying, prioritising and evaluating all potential consequences of quality improvement: modified Delphi study

Madalina Toma,[1] Tobias Dreischulte,[2,3,4] Nicola M Gray,[1] Bruce Guthrie[5]

To cite: Toma M, Dreischulte T, Gray NM, *et al*. A balanced approach to identifying, prioritising and evaluating all potential consequences of quality improvement: modified Delphi study. *BMJ Open* 2019;**9**:e023890. doi:10.1136/bmjopen-2018-023890

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-023890).

Received 30 April 2018
Revised 16 January 2019
Accepted 6 February 2019

[1]Scottish Improvement Science Collaborating Centre (SISCC), School of Nursing and Health Sciences, University of Dundee, Dundee, UK
[2]Division of Population Health and Genomics, University of Dundee, Dundee, UK
[3]NHS Tayside, Prescribing Support Unit, Dundee, UK
[4]Institute of General Practice and Family Medicine, University Hospital of Ludwig-Maximilians-University, Munich, Germany
[5]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK

Correspondence to
Dr Madalina Toma;
m.t.toma@dundee.ac.uk

ABSTRACT

Objectives Healthcare is a complex system, so quality improvement will commonly lead to unintended consequences which are rarely evaluated. In previous qualitative work, we proposed a framework for considering the range of these potential consequences, in terms of their desirability and the extent to which they were predictable or expected during planning. This paper elaborates on the previous findings, using consensus methods to examine what consequences should be identified, why and how to prioritise, evaluate and interpret all identified consequences, and what stakeholders should be involved throughout this process.

Design Two-round modified Delphi consensus study.

Setting and participants Both rounds were completed by 60 panellists from an academic, clinical or management background and experience in designing, implementing or evaluating quality improvement programmes.

Results Panellists agreed that trade-offs (expected undesirable consequences) and unpleasant surprises (unexpected undesirable consequences) should be actively considered. Measurement of harmful consequences for patients, and those with high workload or financial impact was prioritised, and their evaluation could also involve the use of qualitative methods. Clinical teams were agreed as important to involve at all stages, from identifying potential consequences, prioritising which of those to systematically evaluate, undertaking appropriate evaluation and interpreting the findings. Patients were necessary in identifying consequences, managers in identifying and prioritising, and improvement advisors in interpreting the data.

Conclusion There was consensus that a balanced approach to considering all the consequences of improvement can be achieved by carefully considering predictable trade-offs from the outset and deliberately pausing after implementation to identify any unexpected surprises and make an informed decision as to whether quantitative or qualitative evaluation is needed and feasible. Stakeholders' roles in in the process of identifying, prioritising, evaluating and interpreting potential consequences should be explicitly addressed within planning and revisited during and after implementation.

Strengths and limitations of this study

- ► To the best of our knowledge, this will be the first study to generate Delphi-based expert consensus on the identification, prioritisation, evaluation and interpretation of a wide range of quality improvement consequences, an area that has been largely overlooked in the existing literature.
- ► This study provides insights into how a balanced approach to determining all consequences of quality improvement projects can be achieved, the specific factors that need to be considered and the stage at which relevant stakeholders can be actively involved.
- ► The Delphi panel was purposively selected with the majority of participants identifying themselves as academics, quality improvement advisors and providers of healthcare services across the UK, with experience of designing, implementing or evaluating quality improvement interventions.
- ► Although the selection of experts was appropriate for the purpose of this study, the answers provided may not be appropriate for all practice settings, and therefore might limit the generalisability of our findings beyond the UK healthcare system.

BACKGROUND

The complexity of the healthcare system, along with the multiple pressures it faces, means that efforts to improve quality and safety often achieve only limited benefits and can have unintended consequences,[1] (Manojlovich, 2016 #19; Merton, 1936, The Unanticipated Consequences of Purposive Social Action) which may impact positively or negatively on care processes and outcomes. However, several systematic reviews have shown that most papers evaluating quality improvement programmes mainly report impact on targeted goals, with minimal reporting of other unintended consequences.[2 3] For example, only 1 of 121 interventions aiming to reduce falls and catheter-related infections,[4] none of 34 studies

of improvement interventions to improve surgical care,[5] only 6 of 94 (6.4%) studies examining the application of Plan Do Study Act improvement methods[6] and only 1 of 100 perioperative care improvement interventions reported any impact on unintended consequences.[7]

Furthermore, improvement projects rarely evaluate consequences identified after full implementation.[3] A recent Cochrane review of interventions to improve antibiotic prescribing practices for hospital inpatients showed that only 11 (10%) of 110 studies reporting interrupted time series data of improvement interventions (which typically evaluated healthcare rather than research interventions) reported any data about unintended negative consequences.[8] Overall, while there are a number of recommendations for systematically designing quality improvement interventions,[9–11] there is a lack of evidence that improvement programmes routinely evaluate the presence of unintended consequences either preimplementation or postimplementation,[12–15] with little specific guidance on how to best account for improvement consequences other than goals and what potential stakeholders should be consulted in planning, conducting and interpreting evaluations.

We have previously conducted a qualitative analysis of data from 15 semistructured interviews and 2 focus groups with 24 experts to explore the current understanding of unintended consequences of quality improvement.[2] Based on the findings of this analysis, we proposed a structured framework for considering the range of potential consequences of improvement interventions, in terms of their desirability and the extent to which they were expected during the initial planning. As described in figure 1, the framework proposes that a balanced approach should consider goals (expected desirable consequences) and predictable trade-offs (expected undesirable consequences) early in the design of a quality improvement programme and pause to identify and take stock of pleasant (unexpected desirable consequences) and unpleasant surprises (unexpected undesirable consequences) after a period of implementation.

Framed by our previous qualitative work,[2 3] this paper aims to:

1. Validate the previously developed framework and, through expert consensus, establish what potential quality improvement consequences should be identified.

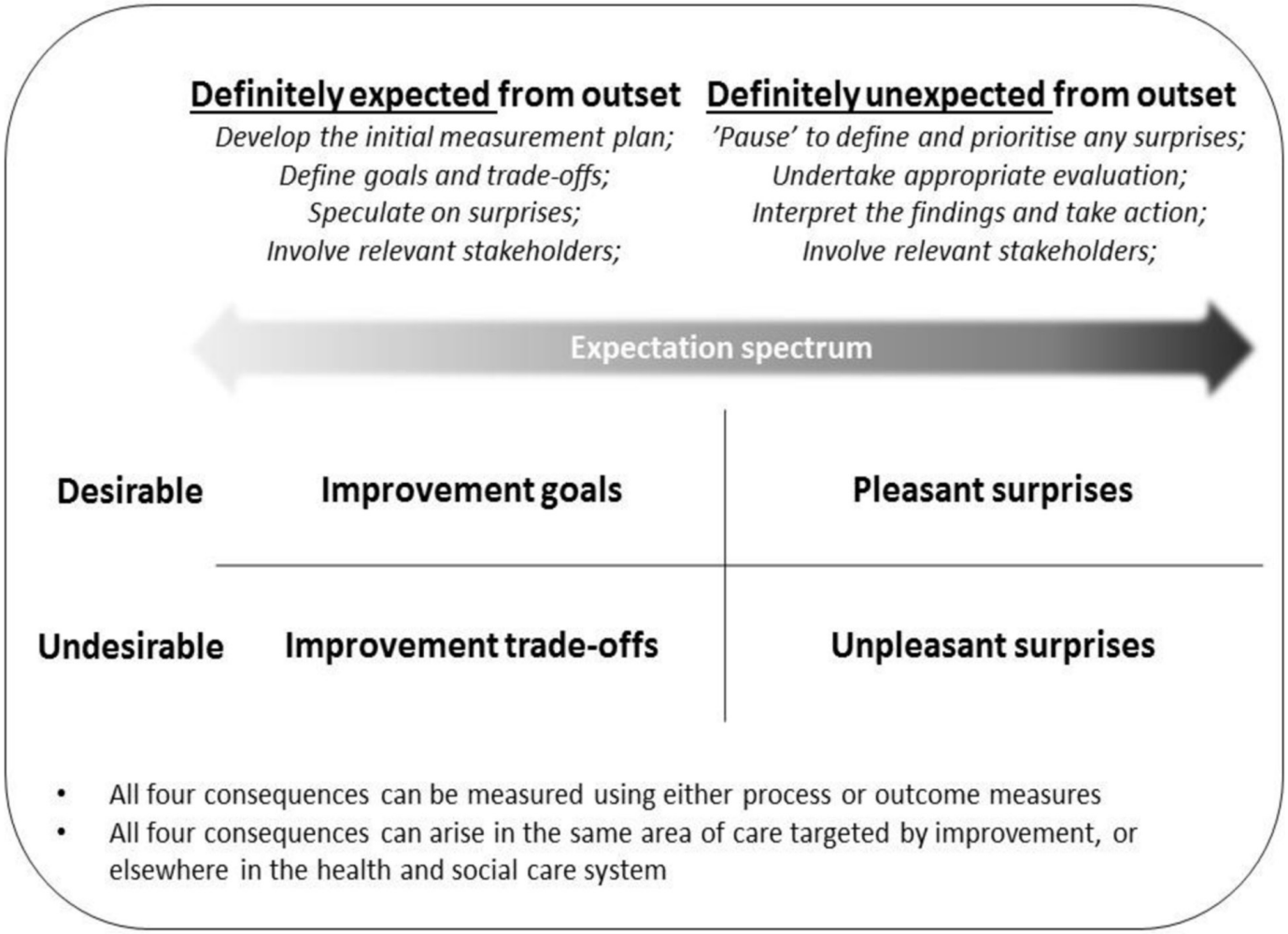

Figure 1 A framework describing different types of consequences of quality improvement projects (derived from previous qualitative work and wider literature, and validated through a two-round consensus study).

2. Extend the framework wider applicability within the quality improvement measurement context, by exploring and developing consensus in relation to why and how to prioritise, evaluate and interpret all identified consequences and what stakeholders should be involved throughout this process.

METHODS

Study design

The well-established consensus process incorporated a two-round modified Delphi method,[16] which consisted of rating and ranking of the importance of various propositions whose focus and scope were determined through the framework developed in our previous qualitative study.[2] The modified Delphi process was chosen as it is recommended for use in the field of quality improvement and patient safety as a reliable means of determining consensus for a defined multifactorial and complex problem. It is also useful in minimising the impact of group interaction and influence, as well as using valuable expert knowledge where understanding is only partial or incomplete.[17]

The development of the Delphi survey

Four key survey sections were defined as follows:

Section 1: Identifying potential consequences of improvement: Delphi panellists were asked what types of improvement consequences should be identified and who should be involved in identifying them.

Section 2: Prioritising which identified consequences to systematically evaluate: Panellists were asked under what circumstances should evaluation be conducted to assess and/or explore any identified potential consequences, and who should be involved in this decision.

Section 3: Undertaking appropriate evaluation for any identified consequences. Panellists were asked to rate statements about how to appropriately evaluate consequences of improvement, and who should be involved in this.

Section 4: Interpreting the emerging data: Panellists were asked to rate statements about who should be involved in understanding and interpreting findings to inform potential action.

Delphi survey piloting

Draft statements were initially pretested by a group of 10 clinical academics who commented on clarity and appropriateness, followed by two rounds of piloting by 14 additional participants with similar academic, clinical or management background as the targeted sample, using the same open access web platform (Bristol Online Surveys) as the main study. Pretesting and piloting led to some additional statements being proposed and added, and to refine the survey in terms of wording and sequencing.

Panel selection and recruitment

The main study panel members were purposively selected to be individuals with experience of designing, implementing or evaluating quality improvement interventions with an academic, clinical or management background. We generated a list of experts by including all the stakeholders approached for the original qualitative study,[2] plus additional improvement advisors, clinical academics, providers of health and social care services, policy-makers and patient representatives identified from online searches of articles with the highest number of citations in two leading quality improvement journals, authors of policy documents setting the general direction of quality improvement, keynote speakers at relevant conferences in the field, members of established quality improvement partnerships, service users attending community groups to advise local health boards and the research team's own networks. Additional Delphi invitees were identified through a snowballing technique,[18] whereby contacted panellists proposed suitable others with similar experience and knowledge.

Data collection

Consistent with consensus development procedures,[17] we used sequential, iterative stages as follows.

Delphi round 1

Participants were invited to take part by a personalised email which included a weblink to the round 1 online questionnaire, a letter of invitation to participate, the study information sheet and a briefing paper explaining the development process and theoretical underpinnings (online supplementary appendix 1). Participants were asked to score statements using a five-point Likert scale with a neutral option, ranging from 'not at all important' to 'extremely important'. The type of Likert scales normally varies from study to study; however, the five-point Likert scale has been the most consistently used as an acceptable compromise between the conflicting goals of offering enough choice to measure strength of opinion and designing items that are readily comprehensible to respondents.[19] Furthermore, five-point Likert scales represent a valid and reliable mean of measuring different levels of item agreement or assigned importance across similar modified Delphi studies.[20]

Space for free-text comments about existing statements was also given in the questionnaire, including justification for response and/or any other important areas which were not addressed. The survey also included open-ended questions for structured elicitation and demographic information relating to participants' role and experience.

Delphi round 2

Participants from round 1 were sent a revised briefing paper (online supplementary appendix 2) along with feedback on their scores on each statement compared with the distribution of all scores (online supplementary appendix 3). Given that there were no significant differences in scoring associated with any of the participant characteristics examined in round 1 (eg, geographical location, roles or experience), feedback was presented in

a combined form. This is consistent with previous literature which shows that if agreement is already satisfactory between stakeholder groups, then the type of feedback given may not make any difference in terms of the number of outcomes retained or reducing the variability of opinion.[21]

Using the same voting method as round 1, participants were subsequently asked to complete a revised questionnaire (online supplementary appendix 4) which included notes to indicate what and where changes had been made in response to all round 1 comments. Example of changes included removing examples which were ambiguous and clarifying and amending the wording of some statements (online supplementary appendix 5). Participants were given 1 month to complete each round of the survey and a reminder letter was sent via email to everyone who had not replied within 14 days.

Data analysis and definition of consensus

There is no accepted, set standard for the target percentage of agreement, with thresholds and definitions of consensus ranging between 51% and 80%.[22 23] We chose to take a conservative approach to defining consensus, deeming it to be present if ≥80% of participants rated each individual statement as very important or extremely important on the five-point Likert scale. All data entered via the web platform were downloaded and analysed using SPSS V.21.0 to calculate frequencies and mean ratings. The synthesis and thematic analysis of free-text responses was undertaken in NVivo V.11 following each round. After coding the initial sets of responses, MT and BG met to compare the labels attached and agreed on a set of codes that MT applied to all subsequent data. The wider team also convened regularly to discuss the summaries of any emerging findings. The focus of the analysis presented in the paper is the round 2 quantitative responses and supporting free-text findings.

RESULTS

Panel characteristics

We attempted to email 180 individuals, with 170 emails delivered successfully. Seventy-two (42.4%) individuals completed the round 1, and of these, 60 (83.3%) completed round 2. Table 1 shows that 50 (83.3%) of the 60 participants completing both rounds worked in the UK, 8 (13.3%) worked in another European country and 2 (3.3%) worked in the USA. Participants had a variety of (often multiple) roles with 28 (46.6%) having an academic background and 26 (43.3%) currently working as improvement advisors. Despite our best efforts to optimise recruitment and retention, only two panellists were providers of social care services (3.3%), two were patients or carers (3.3%) and one identified as a service user representative (1.6%).

Furthermore, more than half of the sample (53.4%) had 6 or more years' experience of working in healthcare quality improvement while 31 (51.6%) reported to have undertaken formal training in this area, most commonly (16.6%) to become Institute for Healthcare Improvement accredited improvement advisors. On the contrary, 29 (48.3%) revealed little or no experience of systematically measuring unintended consequences.

SECTION 1: IDENTIFYING POTENTIAL CONSEQUENCES OF IMPROVEMENT

Table 2 shows that all participants rated measurement of predefined improvement goals as very important or extremely important. There was also consensus that measures are important for identifying trade-offs expected before the implementation (95% rated this as important or very important) or unpleasant surprises emerging after a period of implementation (90%). Although some participants valued pleasant surprises as being important, ratings of this statement (70%) did not achieve the prespecified consensus level, with some participants describing pleasant surprises as being less critical than other consequences in reaching a more balanced approach.

> Improvement needs to be judged on its merits alone and an unpleasant surprise detracts from those merits, but I am not sure whether a more pleasant surprise necessarily augments them. Simply measuring everything in sight just in case it had a positive influence is neither desirable nor feasible and probably not the best use of resources. (Improvement advisor)

There was consensus that clinicians and non-clinicians who directly engage with patients in the targeted area (100%), patients (83%) and managerial staff (80%) involved in organising the targeted care should be involved in identifying all potential consequences of improvement activity (table 3).

Section 2: prioritising which identified consequences to systematically evaluate

There was consensus that potential consequences should be measured if there was a likelihood for high (100%) and moderate harms (95%) to patients, followed by high (98%) or moderate (90%) negative workload implications for the service doing the improvement as well as high (95%) or moderate (85%) negative workload implications for other health and social care services. Consensus was also achieved for reasons including high (95%) and moderate negative financial implications (85%) for services within the area targeted for improvement and high (88%) financial implications for services outside healthcare.

There was consensus about the importance of potential high benefits to patients (90%), to the service doing the improvement (95%) and to other health or social care services (81%), but not for matching moderate benefits (61%–75%), reinforcing the view that the occurrence of trade-offs and unpleasant surprises is probably the current focus when making informed decision as to whether

Table 1 Demographics of the 60 participants completing both rounds of ratings

Panel composition	N (%)
Geographical location	
UK (Scotland 39, England 8, Wales 2, Northern Ireland 1)	50 (83.3)
Europe	8 (13.3)
USA	2 (3.3)
Role within healthcare quality improvement*	
Academic research and/or teaching	28 (46.6)
Quality improvement advisor	26 (43.3)
Provider of healthcare services	11 (18.3)
Managerial staff	6 (10.0)
Policy-maker and regulator	4 (6.6)
Provider of social care services	2 (3.3)
Patient or carer	2 (3.3)
Service user representative	1 (1.6)
Experience in working in quality improvement and patient safety	
No experience	5 (8.3)
Less than 2 years	7 (11.7)
2–5 years	16 (26.7)
6–10 years	16 (26.7)
11–15 years	7 (11.7)
More than 15 years	9 (15.0)
Training in improvement science or quality improvement	
No	29 (48.4)
Yes	31 (51.6)
If yes, then type of training attended†	
Institute of Healthcare Improvement-Improvement Advisor Development Program	10 (16.6)
Lean or Lean Six Sigma Training Programme	5 (8.3)
The Scottish Patient Safety Programme Fellowship	5 (8.3)
Improvement Science Training for European Healthcare Workers	5 (8.3)
Academic qualifications in quality improvement or implementation science	5 (8.3)
Other training or fellowships (various)	11 (18.3)
Experience of using balancing measures in quality improvement or patient safety	
No experience	20 (33.3)
Less than 2 years	9 (15.0)
2–5 years	14 (23.3)
6–10 years	10 (16.7)
11–15 years	2 (3.3)
More than 15 years	5 (8.3)

*Number adds up to more than 60 because participants had the option to select multiple relevant roles.
†Number adds up to more than 31 because participants had the option to select multiple relevant training.

systematic evaluation is needed. No 'low severity' consequences reached the agreed consensus level although, as explained by one participant, the perception of severity can sometimes be a subjective assessment:

> What might appear trivial to an outsider like a one min increase in the time taken for clinical staff to do something might be perceived by the clinical staff as considerably longer and possibly with knock on consequences for scheduling of other tasks. Similarly, what might appear to an outsider to be minor inconvenience for patients might be the last straw. (Policymaker and regulator)

Table 2 All statements about the relative importance of all four types of improvement consequences, in descending order of average strength of agreement

	% rating very or extremely important	Mean rating*
Improvement goals	100†	4.85†
Improvement trade-offs	95†	4.58†
Unpleasant surprises	90†	4.30†
Pleasant surprises	70	3.68

*1=not at all important, 2=slightly important, 3=somewhat important, 4=very important, 5=extremely important.
†Indicates consensus among panellists meaning that ≥80% of participants rated a statement as very important and extremely important.

Beyond effects on the quality of the service or the patient care, increasing staff engagement with the improvement activity (86%) and reducing staff resistance to change (85%) were additional reasons to evaluate outcomes, because it demonstrated that improvers were taking staff concerns seriously (table 4).

There was only consensus that clinical teams (96%) and managerial staff (83%) involved in delivering and organising the targeted care should be involved in prioritising whether the identified consequences are important enough to be evaluated systematically (table 3).

Section 3: undertaking appropriate evaluation for any identified consequences

There was consensus that, irrespective of whether data are collected bespoke or for another purpose, both quantitative (90%) and qualitative data (86%) could be used to evaluate trade-offs, pleasant and/or unpleasant surprises with the same rigour as evaluating the predefined improvement goals (table 5).

As one participant described, qualitative data have much to offer both for the identification of trade-offs before implementation, and supporting postimplementation reflection on surprises, especially when retrospective measurement is not feasible.

> Numerical measures will be important for pre-identified consequences while qualitative data will be particularly important for identifying consequences that fall at the right-hand end of the expected-unexpected continuum. It is important to be curious and find a reliable mechanism for galvanising insights and stories that ultimately bring those conventional metrics alive (Improvement science academic)

There was only consensus that clinical teams delivering the targeted care (91%) should be involved in the implementation of appropriate evaluation for any identified consequences (table 3).

Section 4: interpreting the emerging data

There was consensus that both clinical teams delivering the targeted care (86%) and improvement advisors (91%) should be involved in interpreting the data about unintended consequences (table 3). Panellists explained in their free-text comments how making more use of external expertise in interpreting the data could help make findings more meaningful and readily useful.

> There is a need for the methodological expertise and critical distance of improvement advisors who can see data with fresh eyes. It may be more robust to have them interpret the data initially and then discuss the findings with other groups (Provider of healthcare services)

DISCUSSION

Overview of the main findings

Overall, there was consensus in the Delphi panel about the importance of the majority of the propositions. All participants rated measurement of predefined improvement goals as important, and there was agreement that trade-offs and unpleasant surprises should be actively considered, but no consensus about *pleasant surprises*. Participants prioritised the evaluation of seriously harmful consequences for patients, and those with high workload or financial impact in both the local implementation context, and in other health and social care services. There was also consensus that evaluation of a wider range of consequences could have additional value in terms of increasing staff engagement with the improvement activity and reducing resistance to change irrespective of whether measurement led to any change in implementation. Participants agreed that both quantitative data and qualitative data were helpful to evaluate trade-offs and surprises, with free-text comments highlighting that qualitative data are often useful either to contextualise quantitative data or to understand impact when formal measurement is not feasible. Agreement about the importance of involving various internal and external stakeholders varied depending on the stage of the improvement work. Clinical teams delivering the targeted care were agreed to be necessary to involve in all stages from identifying potential consequences, prioritising which consequences to evaluate, undertaking appropriate evaluation and interpreting the data. Patients were necessary in identifying potential consequences of improvement activity, managers in identifying consequences and prioritising which of those have to be systematically evaluated, and improvement advisors in interpreting the emerging data.

Strengths and limitations of the study

Strengths of the study are that it built on our previous qualitative work on this topic,[2 3] that we recruited and retained a substantial expert panel with an 83.3% response rate between rounds thereby reducing attrition bias, and

Table 3 All statements about the relative importance of WHO should be involved in identifying potential consequences, prioritising which consequences to systematically evaluate, undertaking appropriate evaluation and interpreting the data

	Identifying potential consequences		Prioritising which consequences to systematically evaluate		Undertaking appropriate evaluation		Interpreting the data	
	% rating very or extremely important	Mean rating*	% rating very or extremely important	Mean rating*	% rating very or extremely important	Mean rating*	% rating very or extremely important	Mean rating*
Clinical teams delivering the targeted care (clinicians and non-clinicians who directly engage with patients in the targeted area)	100†	4.80†	96†	4.75†	91†	4.60†	86†	4.48†
Managerial staff involved in organising the targeted care	80†	4.22†	83†	4.25†	75	4.12	78	4.18
Patients or carers	83†	4.21†	70	3.98	58	3.61	56	3.70
Clinical teams outside the targeted area of improvement who directly engage with patients	66	3.73	55	3.55	48	3.35	40	3.28
Improvement advisors (people with healthcare improvement expertise external to the local clinical and managerial teams	61	3.86	71	3.95	73	4.00	91†	4.30†
Third sector (eg, voluntary and community organisations, charities or social enterprises)	50	3.42	18	3.40	36	3.01	38	3.15
Academics (people with relevant expertise with a university or similar academic base and perspective)	46	3.36	53	3.48	38	3.15	73	3.93
Policy-makers and regulators	46	3.33	48	3.33	40	3.01	56	3.45

*1=not at all important, 2=slightly important, 3=somewhat important, 4=very important, 5=extremely important.
†Indicates consensus among panellists meaning that ≥80% of participants rated a statement as very important and extremely important.

Table 4 Statements about the relative importance of different reasons to evaluate any trade-offs, pleasant surprises and/or unpleasant surprises, in descending order of average strength of agreement

	% rating very or extremely important*	Mean rating†
Potential high harm to patients (any serious harm such as death, or common significant harm such as recoverable injury)	100*	4.96*
High negative workload implications for the service doing the improvement (cannot accommodate without compromising other work)	98*	4.80*
High negative workload implications for other health or social care services (cannot accommodate without compromising other work)	95*	4.71*
High negative financial implications for healthcare services	95*	4.70*
Potential high benefits for the service doing the improvement (significant improvement in staff morale or high financial savings)	95*	4.63*
Potential moderate harm to patients	95*	4.58*
Potential high benefits to patients (major health improvements which are not related to the initial improvement goal)	90*	4.53*
Moderate negative workload implications for the service doing the improvement	90*	4.35*
High negative financial implications for services outside healthcare	88*	4.31*
Increasing staff engagement with the improvement activity	86*	4.46*
Reducing staff resistance to the improvement activity	85*	4.31*
Moderate negative workload implications for other health or social care services	85*	4.25*
Moderate negative financial implications for healthcare services	85*	4.28*
Potential high benefits for other health or social care services	81*	4.25*
Moderate negative financial implications for services outside healthcare	78	3.90
Potential moderate benefits for the service doing the improvement	75	4.05
Potential moderate benefits to patients	70	4.01
Potential moderate benefits for other health or social care services	61	3.73
Increasing staff ownership of data and measures	78	4.25

No 'low severity' consequences reached consensus, so these are not shown in the table.
*Indicates consensus among panellists meaning that ≥80% of participants rated a statement as very important and extremely important.
†1=not at all important, 2=slightly important, 3=somewhat important, 4=very important, 5=extremely important.

Table 5 All statements about the relative importance of undertaking appropriate evaluation, in descending order of average strength of agreement

	% rating very or extremely important	Mean rating*
Use quantitative data to measure if trade-offs, pleasant and/or unpleasant surprises have occurred	90†	4.48
Use data (eg, qualitative, quantitative, already available or bespoke) to make evaluative judgements/ measure trade-offs, pleasant and/or unpleasant surprises with the same rigour as measuring improvement goals	88†	4.45
Use qualitative data to make evaluative judgements about the presence and extent of trade-offs, pleasant and/or unpleasant surprises	86†	4.41
Use bespoke data collection by clinical teams to measure trade-offs, pleasant and/or unpleasant surprises	46	4.13
Use data that is already collected for another purpose to measure trade-offs, pleasant and/or unpleasant surprises	41	4.05

*1=not at all important, 2=slightly important, 3=somewhat important, 4=very important, 5=extremely important.
†Indicates consensus among panellists meaning that ≥80% of participants rated a statement as very important and extremely important.

that it used a predefined criterion to define agreement on the importance of a proposition. A potential weakness is that the round 1 questionnaire was structured by our qualitative work, meaning that there was less contribution from the Delphi panel in defining the scope of the propositions, although participants did have the opportunity to add, alter or comment on each section. Additionally, there are no generally accepted rules for how the presence of consensus should be defined, with several factors, such as the aim of research, number of respondents and sequence of rounds, influencing the cut-off chosen.[22 23] Given the exploratory nature of the study, we deliberately chose to take a conservative approach to defining consensus, requiring ≥80% of participants to agree that proposition was very important or extremely important. An implication is that lack of consensus does not necessarily mean lack of importance, and such propositions may be relevant under some but not all circumstances. We, therefore, report all results in detail, and others may choose to consider different cut-offs suitable for their purposes.

Lastly, four-fifths of participants were UK based and two-thirds Scotland based. The UK has a well-developed quality improvement infrastructure which does vary somewhat in the different UK countries, so panel composition may limit the generalisability of the findings. However, participants came from a variety of quality improvement, health service and academic backgrounds, and we believe that the problems, findings and recommendations described in this paper are likely have general application.

Patient and public involvement

The priorities, experience and preferences of the people who used the local services were represented through their participation in the individual and group interviews that informed this study,[2] and their willingness to take part in piloting draft versions of the instrument and completing both rounds of the Delphi. However, we recognise that the final expert panel predominately consisted of academics, quality improvement advisors and providers of healthcare services. There is a need to engage a larger number of participants from outside the immediate word of frontline led to quality improvement, particularly service users, public, third sector partners and social care providers. However, this does not mean that everyone will choose to be involved to the same extent, or indeed will be responsible for planning, monitoring or evaluating care. We instead suggest moving beyond this narrow and exclusive approach,[24] and engage in a critical appraisal of the focus, methods and benefits of involvement, regardless of whether participants are using or providing services.

Implications for quality improvement programmes

The importance of balanced measurement systems is well established, with Drucker making the case for this 50 years ago, and encouraging improvers and managers to think broadly about what success constitutes for their organisation and hence what should be actually evaluated.[25] Furthermore, many of the practical guides to healthcare quality improvement emphasise the importance of developing a balanced set of measures during the planning of an improvement programme,[10 26–28] but the focus of such guides is generally on the measurement of goals,[29–31] with a smaller number of measures for expected undesirable consequences *(trade-offs)* which are easily predictable from the outset. The evidence from multiple systematic reviews of quality improvement evaluations is that few report any measures of unintended consequences,[4–7] consistent with almost half of our participants having little or no experience of using them, despite considerable quality improvement experience overall.

The findings of this consensus study reiterate and confirm the results of our previous work,[2 3] and suggest that those involved in improvement programmes should first articulate clear assumptions and formulate explicit predictions for both goals and trade-offs before implementation, and seek to identify relevant process and outcome measures for both. Second, an 'improvement pause' should be planned after implementation to deliberately step back from goal delivery to take stock and reflect on unexpected consequences of improvement activity. Unpleasant surprises in particular need to be carefully evaluated to see if any harm being caused is enough to stop or adapt the intervention to reduce the likelihood of any unpleasant surprises both within and outwith the area targeted for improvement.

Improvers and managers could anticipate these vulnerabilities by making careful and continuously planned efforts to explore all possible process and/or outcome failures both before and after implementation and as ongoing surveillance mechanisms. However, all improvement programmes are resource constrained and there will always be more risks than can feasibly be measured. Moving beyond simply identifying potential consequences, improvers need to reflect on whether measurement is truly meaningful,[32] make choices as to what identified consequences should be systematically evaluated and rationally account for the relative balance of risk and benefits. Based on the findings of this study, we suggest that this decision should be made by assessing if the depth, seriousness and severity of any trade-offs and unpleasant surprises in relation to patient care and other widespread workload and financial implications are likely to be so significant that they warrant particular attention to ensure these undesirable consequences are identified correctly and evaluated thoroughly, and if necessary action is taken to mitigate them.

Lack of data to measure unexpected consequences remains a significant problem. This is particularly common in healthcare systems where electronic health records are in limited use or where the usable routine data available have limited scope for use in evaluation. This is particularly common in healthcare systems where electronic health records are in limited use or where

the usable data have limited scope. Furthermore, even where relevant quantitative data are available retrospectively it will rarely provide a full explanation for what happened within or outwith the healthcare system. Consistent with other literature,[33 34] our participants agreed that qualitative data have an important role in evaluating surprises directly, as well as contextualising quantitative data where these are available. In the (often misquoted) words of Deming, 'It is wrong to suppose that if you can't measure it, you can't manage it—a costly myth'.[35]

Finally, improvers should think more broadly about the stakeholders they involve as different levels of engagement can be appropriate for different stages in the evaluation process. There was a clear agreement in our study that the clinical teams delivering care should be involved throughout the whole process, but that other stakeholders' importance varied with the stage of improvement. For instance, patients might have a unique perspective on care which is often invisible to most professionals but can usefully inform the identification of wider unintended consequences.[32 36 37]

Managerial staff directly involved in organising the targeted care who also understands the implications of changes on other parts of the system can play a significant role in identifying both intended and unintended consequences and deciding whether measurement is needed by aligning the focus on short-term external demands with internal priorities and long-term focus on quality improvement. However, without interpretation, measurement has little meaning and can be misleading, particularly when unpleasant surprises tend to be under-reported. Improvement advisors can, therefore, actively contribute to summarising and distilling the data, bringing a body of expertise in explicit change theories which are different from, but complementary to, the expertise of managers and clinicians.

The active involvement of other stakeholders (eg, academics, clinical teams outside the targeted area, third sector representatives and policy-makers) was perceived as relatively important but failed to reach the consensus standard, potentially being judged as 'nice to have, but not always crucial'.[38] This was particularly surprising in relation to academics involvement, whose perceived importance was only marginal despite the majority of our expert panel having academic links. This lack of consensus might reflect the recognised pros and cons of using a more rigorous, generalisable, but time-consuming research approach as opposed to small-scale, rapid and locally responsive quality improvement methods.[30] What is important to reiterate, is that in practice, improvers will have to make decisions appropriate to their own context, but we recommend that they actively consider these findings when making situational judgements, and notably that the development of the research skills of local teams might help ensure academic input is viewed more favourably.[39]

CONCLUSION

Based on evidence and consensus opinions of diverse stakeholder community, we conclude that a balanced approach should consider goals and predictable trade-offs early in the design of a quality improvement programme, and subsequently pause to take stock of unpleasant surprises after a period of implementation. Evaluation should be done iteratively during the improvement journey and simultaneously with implementation, using both qualitative and quantitative methods. Vigilance for unexpected consequences should be an ongoing, active pursuit for all relevant stakeholders, whose roles in the process of identifying, prioritising, evaluating and interpreting potential consequences should be explicitly addressed within planning and, if required, revisited during and after implementation.

Acknowledgements This work was undertaken by and on behalf of The Scottish Improvement Science Collaborating Centre (SISCC).

Contributors MT and BG were responsible for planning the study and led the data collection and analysis. TD and NMG contributed to data analysis. MT drafted and led the writing of the manuscript. BG, TD and NMG participated in critically appraising and revising the intellectual content of the manuscript. All authors read and approved the final manuscript.

Funding The Scottish Improvement Science Collaborating Centre (SISCC) is funded by the Scottish Funding Council (SFC), Chief Scientist's Office, NHS Education for Scotland and The Health Foundation with in kind contributions from participating partner universities and health boards. The grant reference number is 242343290 was received from SFC on behalf of all funders.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The ethical approval for the study was granted by the University of Dundee Research Ethics Committee (UREC 15069).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Relevant data supporting the results reported in this paper have been included in the submission as online supplementary material. Other data will be available from the corresponding author on reasonable request.

REFERENCES

1. Dixon-Woods M, McNicol S, Martin G. Ten challenges in improving quality in healthcare: lessons from the Health Foundation's programme evaluations and relevant literature. *BMJ Qual Saf* 2012;21:876–84.
2. Toma M, Dreischulte T, Gray NM, *et al*. Balancing measures or a balanced accounting of improvement impact: a qualitative analysis of individual and focus group interviews with improvement experts in Scotland. *BMJ Qual Saf* 2018;27:547–56.
3. Toma M, Davey PG, Marwick CA, *et al*. A framework for ensuring a balanced accounting of the impact of antimicrobial stewardship interventions. *J Antimicrob Chemother* 2017;72:3223–31.
4. Manojlovich M, Lee S, Lauseng D. A systematic review of the unintended consequences of clinical interventions to reduce adverse outcomes. *J Patient Saf* 2016;12:173–9.
5. Nicolay CR, Purkayastha S, Greenhalgh A, *et al*. Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. *Br J Surg* 2012;99:324–35.
6. Curnock E, Ferguson J, McKay J, *et al*. Healthcare Improvement and Rapid PDSA Cycles of Change: A Realist Synthesis of the Literature.

2012 http://www.nes.scot.nhs.uk/media/1389875/pdsa_realist_synthesis.pdf (Accessed 07 Oct 2017).
7. Jones EL, Lees N, Martin G, *et al*. How well is quality improvement described in the perioperative care literature? a systematic review. *Jt Comm J Qual Patient Saf* 2016;42:196–AP10.
8. Davey P, Marwick CA, Scott CL, *et al*. Interventions to improve antibiotic prescribing practices for hospital inpatients. *Cochrane Database Syst Rev* 2017;2:CD003543.
9. van Bokhoven MA, Kok G, van der Weijden T. Designing a quality improvement intervention: a systematic approach. *Qual Saf Health Care* 2003;12:215–20.
10. Langley GJ, Nolan KM, Nolan TW, *et al*. The improvement guide: a practical approach to enhancing organizational performance. *San Francisco,CA; Jossey-Bass* 1996.
11. Reed JE, McNicholas C, Woodcock T, *et al*. Designing quality improvement initiatives: the action effect method, a structured approach to identifying and articulating programme theory. *BMJ Qual Saf* 2014;23:1040–8.
12. Pronovost P, Wachter R. Proposed standards for quality improvement research and publication: one step forward and two steps back. *Quality and Safety in Health Care* 2006;15:152–3.
13. Brainard J, Hunter PR. Do complexity-informed health interventions work? A scoping review. *Implementation Science* 2015;11:127.
14. Jones E. SIQINS: strengthening reporting of quality improvement interventions and methods in surgery. https://lra.le.ac.uk/bitstream/2381/39747/1/2017-JONES-EL-PHD.pdf (Accessed 11 Sep 2017).
15. McDonald K, Schultz EM, Chang C. Evaluating the state of quality-improvement science through evidence synthesis: insights from the closing the quality gap series. *Perm J* 2013;17:52–61.
16. Dalkey NC. The Delphi Method: An experimental study of group opinion. *Rand Corp Public RM; Santa Monica* 1969.
17. Murphy MK, Sanderson CFB, Black NA, *et al*. Consensus development methods, and their use in clinical guideline development. *Health Technol Assessment* 1998;2:1–88.
18. Streeton R, Cooke M, Campbell J. Researching the researchers: using a snowballing technique. *Nurse Res* 2004;12:35–46.
19. Revilla MA, Saris WE, Krosnick JA. Choosing the number of categories in agree–disagree scales. *Sociol Methods Res* 2014;43:73–97.
20. Wang VC, Reio Jr TG. *Handbook of research on innovative techniques, trends and analysis for optimized research methods*. New York: IGI Global, 2017.
21. MacLennan S, Kirkham J, Lam TBL, *et al*. A randomized trial comparing three Delphi feedback strategies found no evidence of a difference in a setting with high initial agreement. *J Clin Epidemiol* 2018;93:1–8.
22. Lynn MR. Determination and quantification of content validity. *Nurs Res* 1986;35:382–6.
23. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs* 2000;32:1008–15.
24. Ocloo J, Matthews R. From tokenism to empowerment: progressing patient and public involvement in healthcare improvement. *BMJ Qual Saf* 2016;25:626–32.
25. Drucker PF. *The Practice of Management*. New York: Harper & Row Publishers, 1954.
26. Scottish Government. Quality improvement and data measurement - what non-executive directors need to know. 2016 http://www.gov.scot/Publications/2016/01/3484 (Accessed 01 Nov 2017).
27. Clarke J, Davidge M, James L. How-to guide for measurement for improvement. *Patient Safety First* 2009 https://eoeleadership.hee.nhs.uk/sites/default/files/Patient%20Safety%20First%20How%20To%20Guide%20measurement%20for%20improvement.pdf (Accessed 23 Oct 2017).
28. Institute of Healthcare Improvement. Science of improvement: establishing measures. http://www.ihi.org/resources/pages/howtoimprove/scienceofimprovementestablishingmeasures.aspx (Accessed 11 Oct 2017).
29. Dixon-Woods M, Amalberti R, Goodman S, *et al*. Problems and promises of innovation: why healthcare needs to rethink its love/hate relationship with the new. *BMJ Qual Saf* 2011;20:i47–i51.
30. Dixon-Woods M, Martin G. Does quality improvement improve quality? *Future Hospital Journal* 2016;3:191–4.
31. Vincent C, Burnett S, Carthey J. Safety measurement and monitoring in healthcare: a framework to guide clinical teams and healthcare organisations in maintaining safety. *BMJ Qual Saf* 2014;23:670–7.
32. Vincent C, Carthey J, Macrae C, *et al*. Safety analysis over time: seven major changes to adverse event investigation. *Implementation Science* 2017;12:151.
33. Martin GP, McKee L, Dixon-Woods M. Beyond metrics? Utilizing 'soft intelligence' for healthcare quality and safety. *Soc Sci Med* 2015;142(Suppl C):19–26.
34. Ovretveit J, Gustafson D. Evaluation of quality improvement programmes. *Qual Saf Health Care* 2002;11:270–5.
35. Deming WE. *The new economics for industry, government, education cambridge, massachusetts, london*. England: The MIT Press, 1994.
36. Armstrong N, Herbert G, Aveling E-L, *et al*. Optimizing patient involvement in quality improvement. *Health Expect* 2013;16:e36–e47.
37. Pomey MP, Clavel N, Aho-Glele U, *et al*. How patients view their contribution as partners in the enhancement of patient safety in clinical care. *Patient Experience* 2018;5:1:35–49.
38. Leviton LC, Melichar L. Balancing stakeholder needs in the evaluation of healthcare quality improvement. *BMJ Qual Saf* 2016;25:803–7.
39. Vindrola-Padros C, Pape T, Utley M, *et al*. The role of embedded research in quality improvement: a narrative review. *BMJ Qual Saf* 2017;26:70–80.