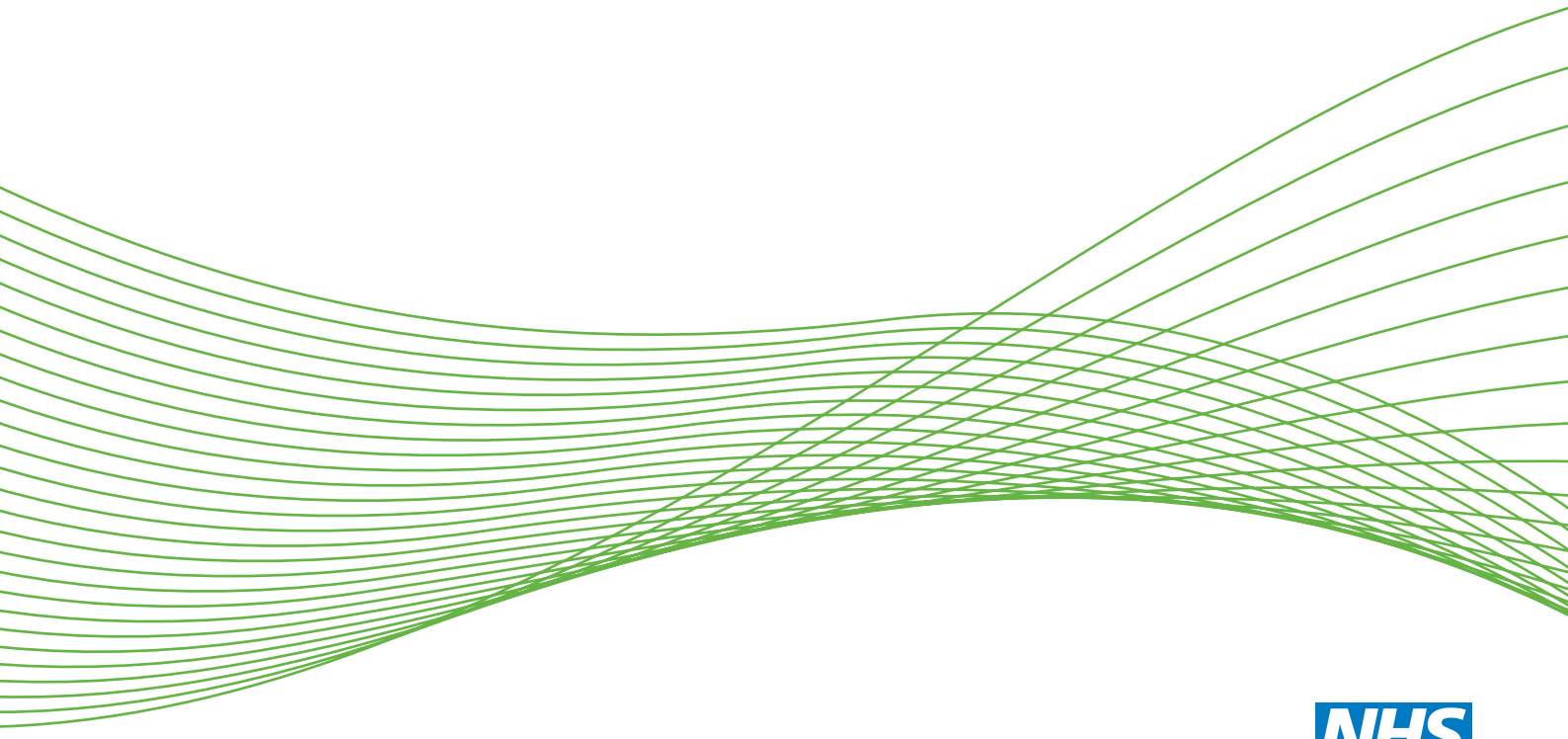
HEALTH SERVICES AND DELIVERY RESEARCH

VOLUME 3 ISSUE 23 MAY 2015
ISSN 2050-4349

A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West

Ruth McDonald, Ruth Boaden, Martin Roland, Søren Rud Kristensen, Rachel Meacock, Yiu-Shing Lau, Tom Mason, Alex J Turner and Matt Sutton



NHS
National Institute for Health Research

DOI 10.3310/hsdr03230

A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West

Ruth McDonald,[1]* Ruth Boaden,[2] Martin Roland,[3] Søren Rud Kristensen,[4] Rachel Meacock,[4] Yiu-Shing Lau,[4] Tom Mason,[4] Alex J Turner[4] and Matt Sutton[4]

[1]Manchester Business School and Centre for Primary Care, University of Manchester, Manchester, UK
[2]Manchester Business School, University of Manchester, Manchester, UK
[3]Institute of Public Health, University of Cambridge, Cambridge, UK
[4]Manchester Centre for Health Economics, University of Manchester, Manchester, UK

*Corresponding author

Declared competing interests of authors: Dr Martin Roland has received personal fees from Advice to Brazilian Ministry of Health on pay for performance and personal fees from Advice to Singaporean Ministry of Health on pay for performance, outside the submitted work.

Disclaimer: The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the National Institute for Health Research, the NHS or the Department of Health.

Published May 2015
DOI: 10.3310/hsdr03230

This report should be referenced as follows:

McDonald R, Boaden R, Roland M, Kristensen SR, Meacock R, Lau Y-S, *et al*. A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West. *Health Serv Deliv Res* 2015;**3**(23).

Health Services and Delivery Research

ISSN 2050-4349 (Print)

ISSN 2050-4357 (Online)

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HS&DR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hsdr. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Services and Delivery Research* journal

Reports are published in *Health Services and Delivery Research* (HS&DR) if (1) they have resulted from work for the HS&DR programme or programmes which preceded the HS&DR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

HS&DR programme

The Health Services and Delivery Research (HS&DR) programme, part of the National Institute for Health Research (NIHR), was established to fund a broad range of research. It combines the strengths and contributions of two previous NIHR research programmes: the Health Services Research (HSR) programme and the Service Delivery and Organisation (SDO) programme, which were merged in January 2012.

The HS&DR programme aims to produce rigorous and relevant evidence on the quality, access and organisation of health services including costs and outcomes, as well as research on implementation. The programme will enhance the strategic focus on research that matters to the NHS and is keen to support ambitious evaluative research to improve health services.

For more information about the HS&DR programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hsdr

This report

The research reported in this issue of the journal was funded by the HS&DR programme or one of its preceding programmes as project number 08/1809/250. The contractual start date was in March 2009. The final report began editorial review in May 2014 and was accepted for publication in November 2014. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HS&DR editors and production house have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the final report document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HS&DR programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HS&DR programme or the Department of Health.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Abstract

A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West

Ruth McDonald,[1]* Ruth Boaden,[2] Martin Roland,[3] Søren Rud Kristensen,[4] Rachel Meacock,[4] Yiu-Shing Lau,[4] Tom Mason,[4] Alex J Turner[4] and Matt Sutton[4]

[1]Manchester Business School and Centre for Primary Care, University of Manchester, Manchester, UK
[2]Manchester Business School, University of Manchester, Manchester, UK
[3]Institute of Public Health, University of Cambridge, Cambridge, UK
[4]Manchester Centre for Health Economics, University of Manchester, Manchester, UK

*Corresponding author ruth.mcdonald@mbs.ac.uk

Background: Advancing Quality (AQ) is a voluntary programme providing financial incentives for improvement in the quality of care provided to NHS patients in the north-west of England.

Objectives: (1) To identify the impact of AQ on key stakeholders and clinical practice; (2) to assess its cost-effectiveness; (3) to identify key factors that assist or impede its successful implementation; and (4) to provide lessons for the wider implementation of pay-for-performance schemes across the NHS.

Design: We tested whether or not the financial incentives of AQ had an impact on mortality using two methods: a between-region difference-in-differences analysis comparing the North West region and the rest of England for the incentivised and non-incentivised conditions and a triple-difference analysis comparing performance on the incentivised conditions, as well as the non-incentivised conditions, in the North West region and the rest of England. A cost-effectiveness analysis of AQ based on the first 18 months of the programme was also undertaken. We used interviews and observation to explore how and why changes occurred.

Results: Risk-adjusted mortality rates for all three of the conditions we studied (pneumonia, heart failure and myocardial infarction) decreased in both the North West region and the rest of England during the first 18 months of the scheme. The reduction in mortality for incentivised conditions was greater in the North West region than in the rest of England. Compared with non-incentivised conditions within the North West region, there was a significant reduction in overall mortality for incentivised conditions, comprising a statistically significant reduction in pneumonia and non-significant reductions in the other two conditions. Comparing mortality for the incentivised conditions with mortality for these conditions in other regions, there was a significant reduction in overall mortality in the North West region, again made up of individually significant reductions in pneumonia and non-significant reductions in the other two conditions. The reduction in mortality over the 18-month period studied for non-incentivised conditions was not significantly different between the North West region and the rest of England. The between-region difference-in-differences analysis after 42 months showed that risk-adjusted mortality for the incentivised conditions fell in the rest of England and the North West region. This reduction in the rest of England was significantly larger than in the North West region and was concentrated in pneumonia. However, the reductions in mortality were larger for the non-incentivised conditions in the North West region than in the rest of England between these periods.

For incentivised conditions, the triple-difference analysis shows a larger reduction in mortality for the rest of England than in the North West region between the short- and long-term periods.

Conclusions: Based on the first 18 months, the AQ programme was a relatively effective and cost-effective intervention. However, findings at 42 months are open to interpretation. One interpretation is that the short-term improvements were not sustained and that the observed improvements in mortality in the non-incentivised conditions within hospitals participating in AQ were unrelated to the programme. An alternative interpretation is that these improvements are related to the positive spillover effect of AQ. Further research should be undertaken to determine the explanation for the findings.

Funding: The National Institute for Health Research Health Services and Delivery Research programme.

Contents

List of tables xi

List of figures xiii

List of boxes xv

List of abbreviations xvii

Plain English summary xix

Scientific summary xxi

Chapter 1 Introduction 1

Chapter 2 Literature review and conceptual framework 3
Introduction 3
Financial incentives for quality in health care 3
Provider capacity and pay for performance 4
Provider support for pay for performance 4
Pay for performance and targeting at individual health-care professionals versus an organisation/team level 4
Pay-for-performance and single indicators versus whole-pathway incentives 5
Voluntary versus mandatory participation 5
Structuring bonuses 6
Size of the incentives 7
Provider costs of pay for performance 7
Applying optimal bonus values to pay-for-performance schemes 7
Pay for performance and penalties 8
Contents of incentivised measure sets 8
Provider case mix issues 9
Unintended consequences and pay for performance 9
Positive spillover effects 10
Supporting levers to accompany the financial incentives 10
Funding pay-for-performance initiatives 11
Designing achievement targets 11
Phasing in pay-for-performance schemes 11
Quantifying achievement in pay for performance 12
Adjusting the pay-for-performance scheme over time 13
Conceptualising change: overcoming challenges 14
Conceptualising change: understanding processes and mechanisms 16
Using the literature to inform our analysis 17

Chapter 3 Advancing Quality: background 19
Defining quality 19
Payment rules 20
Participants and set-up 20
Data definition, submission and monitoring 21

Developments during the evaluation period 22
Commissioning for Quality and Innovation Payment Framework 22
Advancing Quality clinical areas and metrics 22
Patient-experience measures 23

Chapter 4 Quantitative methods and findings **25**
Impact on mortality in the short term 25
Impact on mortality in the short term: methods 25
Impact on mortality in the short term: results 26
Distributional consequences of Advancing Quality in the first 18 months 29
Cost-effectiveness analysis in first 18 months 30
Impact of the change in incentive structure following the introduction of Commissioning for Quality and Innovation Payment Framework 31
Relationship between performance on process measures and observed outcomes 33
Impact on mortality in the longer term 34
Impact on mortality in the longer term: methods 34
Impact on mortality in the longer term: results 35

Chapter 5 Qualitative methods and findings **47**
Data collection 47
Developing a programme theory 48
Adapting a tried-and-tested initiative 49
Piloting 49
Investing in dedicated support and infrastructure 49
Using data to get attention 50
Using voluntarism and peer pressure to drive participation 50
Comparing apples with apples 50
Providing strategic leadership by sustaining senior-level commitment 50
Institutionalising behaviour change 50
Using feedback to inform learning 50
Using competition to drive performance 51
Using collaboration to drive improvement 51
Using money to spur improvement 51
Facilitating standardisation 51
Putting the theory into practice 52
Adapting a tried-and-tested initiative 52
Piloting 53
Investing in dedicated support and infrastructure 53
Using data to get attention 57
Using voluntarism and peer pressure to drive participation 58
Comparing apples with apples 58
Providing strategic leadership by sustaining senior-level commitment 59
Institutionalising behaviour change 60
Using feedback to inform learning 63
Using competition to drive performance 66
Using collaboration to drive improvement 67
Using money to spur improvement 68
Facilitating standardisation 68

Chapter 6 Discussion and conclusions 71
Impact on mortality in the short term 71
Cost-effectiveness of Advancing Quality 71
Relationship between performance on process measures and short-term outcomes 71
Impact on mortality in the longer term 72
Concluding remarks 75

Acknowledgements 77

References 79

Appendix 1 Advancing Quality measures 87

Appendix 2 Changes in trust performance over the first 12 months of Advancing Quality 89

Appendix 3 Changes in the distribution of composite quality score over time 91

Appendix 4 Sensitivity analysis 95

List of tables

TABLE 1 Characteristics of patients before and after the programme in the North West region and the rest of England 27

TABLE 2 Characteristics of hospitals in the intervention and control regions 28

TABLE 3 Changes over time in risk-adjusted mortality for the incentivised and non-incentivised conditions in the North West region and the rest of England 29

TABLE 4 Average hospital achievement on the incentivised indicators in the north-west of England in the short- and long-term periods 35

TABLE 5 Characteristics of patients before and after introduction of P4P in short-term (18 months) and long-term (24 months) periods in the North West region (intervention region) and the rest of England (control region) 36

TABLE 6 Risk-adjusted mortality for the conditions included in the P4P and those not included in the programme, before and after the introduction of the programme in the north-west of England 37

TABLE 7 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme before and after the introduction of the programme in the north-west of England and the adaptation of the programme quality metrics in the late-adopter regions 40

TABLE 8 Risk-adjusted mortality for the non-incentivised conditions (not included in the programme), before and after the introduction of the P4P programme in the north-west of England 43

TABLE 9 Percentage of patients treated under each specialty 45

TABLE 10 Percentage of specialists treating at least one patient with an incentivised condition who also treated at least one patient with a non-incentivised condition 45

TABLE 11 Composite quality scores at the bonus cut-offs, by domain and quarter 91

TABLE 12 Average quarter 1 and quarter 4 performance on each indicator 92

TABLE 13 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England, with different weight and standard error specifications 96

TABLE 14 Risk-adjusted total mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England 98

TABLE 15 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England 100

TABLE 16 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England, excluding hospitals with incentives for incentivised or non-incentivised conditions 101

TABLE 17 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England using 90-day in-hospital mortality 102

List of figures

FIGURE 1 Average quarterly hospital achievement for each condition across indicators 13

List of boxes

BOX 1 Aspects of organisational fields that influence attempts at change 16

BOX 2 Community-acquired pneumonia 19

BOX 3 Examples of changes to the organisational field following AQ introduction 74

List of abbreviations

ACS	appropriate care score
AMI	acute myocardial infarction
AQ	Advancing Quality
AQuA	Advancing Quality Alliance
CABG	coronary artery bypass graft
CAP	community-acquired pneumonia
CBS	Commissioning Business Service
CCG	Clinical Commissioning Group
CI	confidence interval
CMS	Centers for Medicare and Medicaid Services
CPS	composite process score
CQS	composite quality score
CQUIN	Commissioning for Quality and Innovation Payment Framework
HES	Hospital Episode Statistics
HQID	Hospital Quality Incentive Demonstration
ICD-10	*International Classification of Diseases,* Tenth Edition
P4P	pay for performance
P4R	pay for reporting
PCT	primary care trust
PDSA	Plan Do Study Act initiative
PEM	patient-experience measure
PROM	patient-reported outcome measure
QALY	quality-adjusted life-year
QI	quality improvement
QMR	Quality Measures Reporter
QOF	Quality and Outcomes Framework
SHA	Strategic Health Authority
SUS	Secondary Uses Service

Plain English summary

Background

In 2008, a scheme was introduced offering the potential for health-care providers to earn financial rewards by improving quality for NHS patients. All 24 eligible hospitals in the North West region of England participated.

What we did

We talked to people involved in the scheme and observed them in meetings related to the scheme. We also measured the impact of the scheme by looking at whether or not it had improved the death rate of various conditions. These were adjusted for risk and known as 'risk-adjusted mortality'.

What we found

After the first 18 months of the scheme, we found that there was a reduction in risk-adjusted mortality for three clinical conditions included in the scheme. Although there was a reduction elsewhere, the reduction in the North West region was larger.

After the first 42 months of the scheme, we found that the fall in risk-adjusted mortality was greater in the rest of England than in the North West region. However, there was a fall in risk-adjusted mortality in the North West region for some clinical conditions not included in the incentive scheme, which was greater than that for the rest of England.

What can we conclude?

One interpretation is that the short-term improvements we found after 18 months were not sustained and that the observed reductions at 42 months in mortality in the conditions not included in the scheme were unrelated to the scheme. An alternative interpretation, however, is that the incentive scheme led to positive benefits in other clinical conditions in the same hospital.

Scientific summary

Background

A wide variety of pay-for-performance (P4P) schemes have been developed for health-care providers. Such schemes are being increasingly adopted internationally with the aim of improving care quality. However, increased adoption of P4P is occurring despite a scant evidence base.

Advancing Quality (AQ) is a voluntary programme which provides financial incentives to health-care providers for improvement in the quality of care provided to NHS patients. It has been implemented in the North West region of England since 2008. Initially, quality of care was measured by clinical process and outcome measures in five clinical areas – acute myocardial infarction, heart failure, coronary artery bypass graft, pneumonia and hip and knee replacement. Subsequently, the programme expanded to include additional clinical areas, but these do not form part of this evaluation.

The AQ programme evaluation was undertaken over 5 years from 1 April 2009.

Objectives

The study objectives were to:

(a) identify the impact of AQ on key stakeholders (provider organisations, commissioners and patients) and clinical practice
(b) assess the cost-effectiveness of AQ
(c) identify key factors that assist or impede the successful implementation of AQ
(d) provide lessons for the wider implementation of P4P schemes across the NHS as a whole.

Methods

The study used a combination of qualitative and quantitative methods. We assessed the impact of AQ in quantitative terms using national data on mortality, readmissions and length of stay from Hospital Episode Statistics. This component helped us understand what happened. We tested whether or not the incentives had an impact on mortality using two methods: a between-region difference-in-differences analysis comparing changes in mortality over time between the North West region of England and the rest of England for the incentivised conditions and a triple-difference analysis comparing the changes in mortality over time between the incentivised conditions in the North West region and the rest of England with the changes in mortality over time between the North West region and the rest of England for the non-incentivised conditions. In addition, a cost-effectiveness analysis of AQ based on the first 18 months of the programme was also undertaken.

This quantitative analysis was combined with qualitative data collection and analysis aimed to shed light on how and why these impacts occurred. During the first 3 years of our 5-year evaluation we conducted interviews ($n = 391$) with relevant NHS staff and observations ($n = 52$) of meetings and events. During the final 2 years, we interviewed at least one member of staff from each participating provider organisation and 11 commissioners.

Results

Our assessment of impact found that AQ was associated with significant reductions in patient mortality during the first 18 months of the programme (Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 2012;**367**:1821–8). Risk-adjusted mortality rates for all three of the conditions we studied (pneumonia, heart failure and myocardial infarction) decreased over the study period in both the North West region and the rest of England. The reduction in mortality for incentivised conditions was greater in the North West region than in the rest of England, reducing from 21.9% to 20.1% in the North West region and from 20.2% to 19.3% in the rest of England. Compared with non-incentivised conditions within the North West region (within-region difference-in-differences analysis), there was a significant reduction in overall mortality for incentivised conditions of 0.9 percentage points [95% confidence interval (CI) 0.1 to 1.7 percentage points], comprising a statistically significant reduction in pneumonia and non-significant reductions in the other two conditions. Comparing mortality for the incentivised conditions with mortality for the same conditions in other regions, there was again a significant reduction in overall mortality in the North West region of 0.9 percentage points (95% CI 0.4 to 1.4 percentage points), which was also made up of individually significant reductions in pneumonia and non-significant reductions in the other two conditions. Combining these two suggested an overall reduction in mortality of 1.3 percentage points in the North West region (95% CI 0.4 to 2.1 percentage points), with a similar pattern for the individual conditions. The reduction in mortality over the 18-month period studied for non-incentivised conditions was not significantly different between the North West region and the rest of England.

Based on the first 18 months, we found AQ to be a cost-effective use of resources. The total cost of the AQ programme was just over £13M over the initial 18-month period, with only £5M of this consisting of the financial incentives. The ongoing running costs of the scheme exceeded the bonus payments, making up the majority of the costs at just over £7M. We estimated a gain of 6700 quality-adjusted life-years (QALYs) as a result of the reduction in mortality for the programme as a whole. At a QALY value of £20,000, this equals an estimated health gain worth £134M. Our estimates suggest that AQ also resulted in a reduction of 22,700 bed-days in the first 18 months. This is equivalent to a £5M reduction in costs.

The average performance reported by the participating hospitals on all of the quality measures improved in the first 18 months and improved further in the following 24 months, particularly for heart failure and pneumonia. Some of the process quality measures were significantly associated with better health outcomes at a trust level but the magnitudes of the estimated coefficients were too large to represent clinically plausible direct consequences of these process measures. The findings suggest that these financial incentives only weakly led to improved patient outcomes through their direct effects on the process measures that were incentivised.

Advancing Quality appears to have also led to improved patient outcomes by inducing positive spillover effects in terms of wider improvements in care quality across unmeasured dimensions and improvements in care for all patients. Our qualitative data provide support for this explanation, highlighting developments at sites (e.g. recruitment of specialist nurses to join up gaps in care and maintain a sustained focus on patients as they moved through the hospital) to improve care quality for patients in AQ clinical areas. They also suggest that clinician compliance with data-recording requirements varied between clinicians and across sites. Performance on process measures reflects what is recorded as opposed to the care that was delivered, and failure to record care delivery in a systematic fashion was a persistent problem. This further complicates the issue of quantifying relationships between performance on process measures and the relevant outcomes.

When we looked over the longer-term period from 18 to 42 months, risk-adjusted mortality rates continued to decrease in both the North West region and the rest of England for both incentivised and non-incentivised conditions. The between-region difference-in-differences analysis showed that risk-adjusted mortality for the incentivised conditions fell by 2.3 percentage points in the rest of England

and by 1.8 percentage points in the North West region. This reduction in the rest of England was significantly larger (0.7 percentage points, 95% CI 0.3 to 1.2 percentage points) than in the North West region, and was concentrated in pneumonia (1.1 percentage points, 95% CI 0.4 to 1.8 percentage points). However, the reductions in mortality were also larger for the non-incentivised conditions in the North West region than in the rest of England between these periods (1.2 percentage points more, 95% CI 0.4 to 2.0 percentage points).

We considered various explanations for the smaller reduction in mortality for the incentivised conditions in the North West region in the long term (i.e. at 42 months) compared with the rest of England. The first is the possibility that the scheme became less effective with the change in incentive structure, as the AQ programme switched from a tournament scheme with bonuses to a scheme involving penalties for failure to reach quality benchmarks. The continued improvement in performance on incentivised process measures in the AQ hospitals suggests that the incentives may still have been effective, but we have no data from control hospitals for these measures. Moreover, as described previously, we did not find a significant relationship between performance on process measures and outcomes.

A second possible explanation is that there was a positive spillover effect from the adopting region (i.e. the North West region) to other regions. The early results of AQ had been widely disseminated in England and two other regions had adopted a form of AQ programme incentives. These regions showed a greater reduction in mortality in the long term compared with other control regions that did not incentivise the AQ indicators, although the reduction was statistically significant only for acute myocardial infarction.

We also found limited evidence for positive spillover effects within the AQ hospitals, as the patients with non-incentivised conditions that were treated by specialists who also treated patients with incentivised conditions experienced the largest reductions in mortality in the long term.

A number of factors appeared to contribute to the success (as measured by improving performance on process measures and mortality at 18 months) of the scheme. These include in-person collaborative learning events, dedicated infrastructure support, financial rewards to invest in additional staff and a combination of competition to spur improvement and collaboration to facilitate learning. Additionally, programme participants were able to contribute to shaping the programme as it evolved, enhancing legitimacy and buy-in.

At the same time, there were a number of barriers to implementation. In the context of heavy workloads and competing priorities, frontline staff did not always adhere to AQ requirements. Furthermore, data collection was burdensome in a context in which AQ was not part of existing electronic patient information systems. AQ did not become institutionalised and embedded into routine behaviours. Instead, there was a reliance on core AQ staff to cajole and persuade other staff members, which often resulted in going around obstacles rather than resolving enduring problems. Although there were some common themes in the approach taken (in particular, the employment of specialist nurses), more generally, hospitals implemented AQ using a range of activities tailored to and developed in their local context. This suggests that there was no one blueprint for implementing AQ in each site.

In terms of impact on commissioners, input from staff in commissioning organisations was relatively limited in the first year of AQ. Although some commissioner staff had begun to engage with AQ by year 2, the subsequent reorganisation of NHS commissioning functions during the study period meant that input from commissioners was limited or non-existent for most of the study period.

The AQ scheme design incorporated features of what the literature identifies as good practice. It did not involve penalties and it rewarded relative, as well as absolute, performance. The fact that participation was on a voluntary basis and was universal (i.e. all 24 eligible organisations took part) appeared to add to the legitimacy of the AQ programme. Additionally, the competitive nature of the scheme did not crowd out knowledge sharing and collaboration more generally. However, our findings, which highlight

implementation challenges and a failure to embed change in routine practice, suggest that, although scheme design is important, there are other aspects relating to implementation that require attention if financial incentive schemes are to fulfil and maintain their potential.

Conclusions

Based on the first 18 months, AQ was a relatively cost-effective intervention. The findings after 42 months are open to several interpretations. Our failure to find a relationship between process and outcome measures at 18 months suggests that there were positive effects beyond the changes in the specific AQ measures. An alternative interpretation, however, is that short-term improvements were not sustained and that the observed improvements in mortality in the non-incentivised conditions within hospitals participating in AQ were unrelated to the programme.

The first explanation is supported by changes to care delivery identified by our evaluation. It may be that there were further positive spillover effects in quality of care both from participating to non-participating hospitals and from incentivised to non-incentivised conditions in the participating hospitals. We found some modest evidence for both of these hypotheses. However, we did not explicitly focus on non-incentivised conditions. Furthermore, because we collected qualitative data from a large number of sites ($n = 24$), we were unable to conduct detailed, in-depth research to explore these issues comprehensively.

Further research to investigate the relationship between AQ and changes in incentivised and non-incentivised conditions would shed light on this area. Linked to this, research exploring changes in rest-of-England sites would also add to our knowledge.

The study highlights the importance of considering costs beyond the incentive payments of financial incentive programmes intended to improve care quality. It also suggests that competition did not inhibit collaboration, with providers keen to share learning within the AQ community of practice. Instead, cohesive network relationships appeared to support the social enforcement of anticompetitive norms. In-person collaborative learning events were an important part of building and sustaining such relationships.

We found no evidence of changes in care resulting from AQ being institutionalised. Instead, modifications to practice were generally not systematised and behaviour change was still largely reliant on prompting by particular individuals. The success of AQ seems to have been a result of persistent and focused individuals working to remind staff and to plug gaps in data collection and/or care pathways. Furthermore, far from being everybody's business and part of organisation-wide change, AQ was delivered in a context in which many staff were unaware of its existence. Further research should be undertaken to determine the explanation for the findings.

Funding

The National Institute for Health Research Health Services and Delivery Research programme.

Chapter 1 Introduction

A wide variety of pay-for-performance (P4P) schemes have been developed for health-care providers, and such schemes are being increasingly adopted internationally with the aim of improving care quality.[1,2] Increased adoption of P4P is occurring despite a scant evidence base. A review published in 2009 found that only three hospital P4P schemes had been evaluated, and that good evidence was available for only one scheme. This was a US-based initiative called the Hospital Quality Incentive Demonstration (HQID), which was adopted by the Centers for Medicare and Medicaid Services (CMS) in 2003 and supported by Premier Inc.,[3] a nationwide organisation of not-for-profit hospitals. These evaluations[4–6] and later papers[7,8] show, at best, modest impacts on hospital processes of care. Evidence of an effect on patient outcomes is even weaker; HQID was shown to have had no effect on patient mortality[9] and a 2011 Cochrane review found no evidence that financial incentives improve patient outcomes.[10] This report details the methods and findings of a 5-year National Institute for Health Research, Health Services and Delivery Research programme-funded project evaluating Advancing Quality (AQ), a hospital-based P4P initiative in the north-west of England.

Advancing Quality is a voluntary programme that provides financial incentives for improvement in the quality of care provided to patients. It has been implemented in the North West region of England since 2008. The programme is based closely on the P4P demonstration project implemented in the USA, HQID, which involved a partnership between CMS and Premier Inc. AQ was initially designed and supported by Premier Inc., and involved similar quality indicators and financial incentive structures. However, it differed from HQID in involving the universal participation of eligible providers and in being implemented in a different health system.

The AQ programme evaluation was undertaken over 5 years from 1 April 2009. Given the changing NHS context, a two-stage process was adopted which involved agreement of aims for the first 3 years of the evaluation, with aims for the final 2 years being agreed in year 3 of the evaluation.

The aims of the evaluation were to:

1. provide a wide-ranging and in-depth evaluation of AQ in the north-west of England
2. identify key lessons for the adoption of P4P systems in the UK NHS
3. add to the evidence base concerning P4P systems.

In order to achieve these aims, the four main objectives were to:

1. identify the impact of AQ on key stakeholders (provider organisations, commissioners and patients) and clinical practice
2. assess the cost-effectiveness of AQ
3. identify key factors that assist or impede the successful implementation of AQ
4. provide lessons for the wider implementation of P4P schemes across the NHS as a whole.

During the first 3 years of the evaluation, we identified that the first 18 months of the programme led to significant reductions in patient mortality and lengths of stay in hospitals.[11] After the initial 18 months, the programme underwent a number of changes (notably a change from financial bonuses to penalties, extension to new disease areas, and a change in supporting contractor). Consequently, the main objectives agreed for the second and final phase of the evaluation were as follows:

(a) to analyse how the impact on mortality was generated
(b) to examine whether or not the impacts and cost-effectiveness of AQ are sustained over time
(c) to assess whether or not the structure of the financial incentives impacts on performance
(d) to develop a framework for efficient design of financial incentives in the future.

Chapter 2 Literature review and conceptual framework

Introduction

The literature in the area of incentives is large and growing and much of it is concerned with financial incentives. In addition to studies demonstrating positive effects of changes in incentive structures (financial and otherwise),[12] there is a substantial literature derived from a wide range of sectors on the potential for such performance management systems to generate unintended and dysfunctional consequences.[13] Owing to the large volume of theoretical and empirical literature, which may have some relevance to this topic, and the need to limit the review to manageable proportions, it has been necessary to draw some boundaries with regard to the scope of the review.

We have also attempted to avoid duplication of other research. Davies *et al.*[14] recently undertook a review of the literature on incentives and governance, which provides an in-depth and highly accessible review of the literature in this area. Christianson *et al.*[15] also reviewed the literature to assess the impact of financial incentives on the quality of care delivered by health-care organisations and individuals, and rather than reiterate its contents in detail here we refer interested readers to this accessible report.

Our recent report *The Impact of Incentives on the Behaviour and Performance of Primary Care Professionals*[16] contains an extensive (c. 5500 words) discussion on incentives and motivation drawing on economic, psychological and sociological literatures. Interested readers should refer to this report for an in-depth discussion of the literature in this area.

In what follows, we present a selective review of some of the literature, chosen for its relevance to the subject of incentives in the context of AQ. The first part of the review discusses issues relating to the design of incentive schemes, highlighting good practice where possible. This is followed by selective discussion of literature relevant to the understanding of change processes in health-care organisations.

Financial incentives for quality in health care

For a number of years, many Western countries have operated fee-for-service payment systems, with payments based on some measure of the volume of care, such as the numbers of procedures undertaken or the total number of bed-days.[17] However, such schemes may not lead to optimal quality, as health-care providers are incentivised to maximise volumes through unwarranted procedures or superfluous lengths of stay.[18] Alternative payment systems, such as bundled payments and capitation, are an attempt to curb unnecessary reimbursement. However, these can have the opposite effect of discouraging behaviours by the organisation that, while clinically necessary or desirable, would not provide them with additional remuneration (e.g. screening and other preventative measures or some elective procedures).[19] P4P is intended to act as a middle ground between these two designs, incentivising to a greater extent the completion of behaviours related to improved quality of care, without discouraging the completion of other necessary yet non-incentivised procedures.

Implementation of P4P schemes is a growing trend in health-care settings.[20] In the NHS, several programmes have been introduced in the past 10 years, including the national Quality and Outcomes Framework (QOF)[21] and Commissioning for Quality and Innovation Payment Framework (CQUIN).[22] Frequently, P4P schemes are justified by a need for improved, standardised levels of care quality and as a method to increase transparency, particularly when coupled with a public-reporting element.[3] However,

reviews of the ability of P4P to fulfil such goals suggest that improvements are inconsistent and often only temporary.[2]

Provider capacity and pay for performance

The introduction of a quality measurement and reward initiative requires reliable technical support and a capable administrative staff. At its most basic, this requires a sufficient data system to gather and (if necessary) analyse data, and requires individuals responsible for advocating and implementing the scheme internally. A number of papers recognise electronic medical records as a key element in implementing both a successful P4P scheme and, in turn, an efficient health-care system.[23,24] In summarising the effectiveness of best practice tariffs, McDonald *et al.*[25] noted reduced performance on the quality measures that required a large amount of additional data collection and a restructuring of electronic data systems. Additionally, providers identified insufficient capacity in relation to the built environment (e.g. day-case surgery facilities) and workforce (e.g. orthogeriatrician capacity as a result of a failure to invest locally but also as part of a more general scarcity workforce issue). The built environment may help or hinder quality improvement (QI) initiatives. Additionally, where such initiatives require the shifting of responsibilities across professions, a rebalancing of the provider workforce may be needed.

Provider support for pay for performance

McDonald *et al.*[25] also identified a lack of support for certain aspects of a P4P programme in areas where clinicians disputed the content of best practice pathways. This suggests that obtaining provider support for a P4P scheme is vital for its success. Furthermore, an evaluation of the national CQUIN initiative in a later study by McDonald *et al.*[26] found that, although support was often present among managers involved in negotiating quality goals and related payment rules, such support was often absent with regard to frontline clinicians. There were various reasons for this, including a failure by managers to engage and inform frontline clinicians in the process.

Additionally, in some cases, commissioners sought to impose quality goals, which led to friction and less collaborative forms of working in addition to a tick box approach in some cases. Other studies highlight how a lack of engagement and ownership reduces the impact of such initiatives.[27] This suggests that scheme content needs to be informed by close discussion with clinicians and experts in order to ensure relevance, feasibility and commitment. Clear and robust communication and support mechanisms are needed between relevant parties during all phases of a scheme's design and development, in order to avoid implementing a scheme of which clinicians are unaware or unsupportive.[28]

The tendency of financial incentives to crowd out intrinsic motivation[29] raises concerns that P4P may compromise altruistic desires to maximise the welfare of patients. Such effects may be minimised by ensuring that tangible rewards (such as money) are complemented by symbolic rewards (such as praise or public recognition). Provider support is likely to be more readily forthcoming when P4P initiatives involve applying incentives to standardised, simple processes, rather than more complex processes requiring greater cognitive application; limit commissioner use of coercive methods, such as surveillance and threats, when promoting quality measures; and acknowledge – if not ensure – that the level of incentive reflects the cost of additional effort required by participants.[30] Crowding out of intrinsic motivation may also be reduced by targeting the scheme at teams and groups rather than individuals and by involving clinicians in the development of indicators and exclusion criteria.[31]

Pay for performance and targeting at individual health-care professionals versus an organisation/team level

In an aggregated (i.e. team as opposed to individual) measure, poor performance of one contributor may be hidden by high performance from other individuals.[32] Through performance monitoring of distinct individuals, there exists a more direct link between behaviour and output and a greater visibility of free-riding participants. Moreover, some procedures and care practices, such as the provision of discharge information or smoking cessation guidance, rely on the autonomy of single individuals, which may be undermined by team-level measures.[33]

Evaluating individual-level performance for providers that see a low relevant caseload in the scheme timeframe may be overly complex when individual contributions to care are unclear, and misleading when individual physicians see only a small relevant caseload.[34] It may be unfitting to reward just one individual involved in a patient's care where such care has spanned a number of health-care professionals. Rewards across clinical teams place greater emphasis on incentive schemes as a form of altruism and a way of improving welfare for patients, rather than as a source of additional income for individuals.[35]

A team-level incentive scheme may still be supplemented with individual-level performance reporting, minimising issues of accountability and free-riding in group schemes. However, setting a sufficiently challenging minimum target for performance within a team may be preferable because this can also substantially reduce free-riding, by allowing peer effects to increase motivation.

Pay-for-performance and single indicators versus whole-pathway incentives

It may be inappropriate to incentivise single-indication quality measures without consideration of the patient's other conditions, which may advance the risk of drug adverse events or treatment contraindications.[36] Furthermore, providing health care often involves the expertise of several individuals, across a substantial time period, and this is particularly the case for patients with chronic diseases or multiple comorbidities.

Many health conditions lack a clear care plan, particularly when the patient remains undiagnosed or with an unclear course of treatment.[37] Patient interactions with primary and secondary care services – hospitalisations, repeat prescriptions, rehabilitation, and so on – are often unpredictable in chronic health conditions such as diabetes or mental health disorders. This complicates the development of incentivised measures significantly, as they need to span the whole care pathway.

Pay-for-performance schemes based on care plans have been introduced, for example in Germany and Australia, and designed around a package of care spanning prevention and treatment, with several co-ordinated providers incorporated within one payment contract.[38] Rather than externally assessing the relative contributions of clinical teams or individuals, bonuses are often received at the system level, with teams and individuals rewarded at the provider's discretion. However, these care plan P4P schemes have yet to be evaluated for effectiveness.[38]

Voluntary versus mandatory participation

Voluntary P4P schemes allow providers with a sufficient level of infrastructure to join the scheme at their discretion, without the need to wait for lesser-resourced providers to develop their labour or IT capabilities. Voluntary P4P schemes may only attract those who are already performing highly, excluding those most in need of the programme. Furthermore, highly resourced providers may reap benefits from early participation, further widening the performance gap between providers.[39] Mandatory participation in P4P would therefore remove inequity between providers resulting from self-selection. However, coercing participation is likely to impact adversely on intrinsic motivation.[27]

Evidence of over-representation of high performers in voluntary P4P schemes is inconclusive. One study of a Hawaiian quality initiative suggests that low- and high-performing providers were uniformly represented in the scheme.[40] A study of HQID, the US-based predecessor of AQ, suggests that those who volunteered for the scheme exhibited significantly different levels of patient volume and mortality.[9] Overall, conclusive evidence on the extent to which voluntary or mandatory participation affects performance is relatively absent from the literature.

Structuring bonuses

There are four principal structures for rewards under P4P:

1. a predefined, single threshold of achievement (benchmarking/target – e.g. completion of a given measure among *x*% of eligible patients)
2. multiple thresholds of achievement at predefined intervals (e.g. a small incentive for completion of a given measure among *x*% of eligible patients, followed by a larger incentive for completion among *x* + 10% of patients)
3. continuous reward systems, with incentive provided on a per-completed measure/patient basis
4. relative performance (a tournament – e.g. incentives shared among the top *x* performing providers).

These are often used in combination to reward a mixture of absolute, relative, and improvement in, performance.

Single performance thresholds

Rewarding providers based on a single threshold of performance has the benefit of simplicity both at the point of scheme design and in communication with participants. However, a threshold that is set at a very high level could discourage lower-performing providers from engaging, particularly when there is no secondary reward for improvement. If thresholds are not increased relative to achievement levels, high performers may find themselves eligible for incentive with no additional effort required and providers may disengage from the scheme (as occurred in one US programme[41]).

Multiple performance thresholds

Multiple performance thresholds, however, create incentives for both high and low performers. Much like in a single-threshold structure, however, motivation to improve could decrease past the point of attaining threshold performance – an issue predicted among high performers in the QOF.[42] Increasing targets in a threshold scheme as providers improve should reduce the likelihood of this, but creates an additional administrative burden. Multiple thresholds[35] and continuous rewards work to reward all providers in proportion to their level of achievement, and, thus, they are more likely to be effective in rewarding absolute performance and improvement over time simultaneously.

Continuous reward systems

Much like a single-threshold system, continuous reward structures have the benefit of simplicity and also guarantee a reward for all participants (in the absence of a minimum performance threshold). Unlike threshold systems, however, rewards are directly proportionate to performance. In the absence of any additional bonus for improvement, the structure by its very nature guarantees that lower-performing providers will receive lower payments; several studies suggest that this can widen inequalities in service provision.[39,43] At the same time, providers that can quantify performance to date may feel sufficiently rewarded and discouraged from exerting any further effort for the remainder of the period, particularly on quality measures that require significant investment.

Furthermore, the idea of rewarding participants irrespective of achievement could be contentious among stakeholders who feel that a minimum standard should be met before any incentive is given, particularly when the indicator represents standard clinical guidelines of care. A minimum threshold of achievement could be set, below which providers would be ineligible for reward (or even penalised). Performance on quality measures often exhibits a plateau effect past a given level of achievement; assuming constant returns to effort, there will be a point at which the marginal cost of initiating the measure in another patient outweighs the marginal benefit to the provider.

Relative performance (tournaments)

Tournament schemes, based on a ranking of providers, incentivise providers to compete against one another for rewards. In the absence of perfect knowledge of other participants' behaviour, the level of performance required for reward is unknown, eliminating the risk that providers would slow performance past a given level of achievement and instead motivating participants towards a system of continuous improvement.[44]

In other scheme types, commissioners may risk a shortfall of funds if achievement is higher than expected. In a tournament scheme, on the other hand, the number of winners is fixed. Commissioners therefore possess a greater level of certainty of the final reward value.[33] However, the fact that providers do not know what level of performance is required could discourage large-scale investment in QI, on the basis that the expected return on investment is uncertain. Similarly, if some providers believe that they are low performers relative to their competitors, they could disengage from a programme that rewards solely on relative performance, under the impression that they are unlikely to receive payment.[45]

When the incentive is contingent upon outperforming other providers, evidence suggests that collaboration and knowledge sharing is less likely.[46] Tournament-style schemes can be inappropriate where there exists low variability in performance across participants.[47] In this case, the differences in performance that decide who receives a bonus and who does not can be very small and may undermine confidence and credibility in the scheme.

Size of the incentives

Historical evidence suggests that incentive values in P4P schemes are rarely correlated with subsequent improvements in care quality.[20] Kristensen *et al.*[22] note that many systematic reviews of the effectiveness of P4P often only touch on the notion that bonus values and subsequent performance may be correlated or that bonuses should be equated with the marginal costs and benefits of the scheme.

Provider costs of pay for performance

Increased effort results in a cost to providers. This can involve communication, administration and data-collection costs, as well as the cost of the behaviour change itself. Some quality indicators will be more straightforward and/or less costly to implement than others: persuading individuals to receive a vaccination, for example, is generally easier than convincing them to quit smoking; and, conversely, provision of either service would be cheaper than administering magnetic resonance imaging or computerised tomography for a patient.

In converting the cost savings of QIs into provider tariffs, rewards reflect the impact of high performance on cost savings in the system – a tariff design used in the first major Medicare demonstration[48] but yet to be seen in the UK. However, many improvements in quality are difficult to quantify in financial terms, such as programmes to raise patient satisfaction or to target long-term primary prevention. Furthermore, using expected savings to set incentive levels would represent a significant administrative burden and require frequent re-evaluation as clinical evidence evolves. Cost savings-based incentives that do not explicitly refer to the provider and commissioner costs of implementing and maintaining the programme when estimating returns result in flawed and/or partial evaluations of cost-effectiveness.

Many P4P schemes (such as AQ) have been introduced with an intentionally minimal direct focus on cost savings; the principal objective is often instead to improve and standardise care quality for patients, with any subsequent cost saving superfluous to the general cause. Setting bonuses proportionate to cost savings automatically places emphasis on the scheme as a method for saving money, rather than for any altruistic reasons, which may lead to disengagement from participants (see *Provider support for pay for performance*).

Applying optimal bonus values to pay-for-performance schemes

Optimal service provision, from the perspective of a provider, is to continue implementing quality measures until the cost of doing so for one extra patient is equal to these marginal benefits. Hypothetically, providers would still choose to participate in the absence of a financial bonus, as long as the return from the altruistic component and marginal benefits to patients outweigh the costs. In reality, however, this perspective would be financially unsustainable, from the perspective of the commissioner at least; all NHS funds have an opportunity cost in terms of the potential health gains foregone in other areas.

Evaluating the benefit associated with increased service quality among providers is difficult: altruism as a concept would be complex to quantify and measure, and individual physicians and providers would obtain differing levels of personal satisfaction from improving patient care via P4P measures.

In practice, the cost savings and altruistic sentiments resulting from P4P are often unknown or unquantifiable. In a recent evaluation of the CQUIN scheme, although some attempt was made to mirror incentives to effort required or the level of priority assigned to a measure, decisions on reward weightings were taken on an ad hoc, localised basis, rather than with any formalised evidence to hand.[26]

Pay for performance and penalties

Economic theory[49] suggests that 'people impute greater value to a given item when they give it up than when they acquire it'.[50] Although this implies that penalties may more effectively encourage behaviour changes among providers than would equivalent rewards, more recent evidence suggests an increased level of gaming of the system and a greater demotivating effect when providers are faced with potential losses.[30] Penalties may also reinforce a cycle of poor performance, often penalising providers most in need of investment that already lack the financial ability to innovate and implement change (see Werner *et al.*[39] for an example comparing safety-net and non-safety-net hospitals in the USA).

Contents of incentivised measure sets

A key consideration in scheme design is whether or not measures should focus on processes or outcomes. Patient outcomes are often driven by factors outside the control of providers; mortality and morbidity rates are likely to be higher in areas of high poverty, with greater disease prevalence and with lower patient education. Process or structure measures are less likely to be influenced by environmental factors outside the control of the provider and so may reduce provider concerns. Some specialties may be more ideally suited to specific measure types, owing to the way in which care is provided and care quality measured.[51]

Some quality measures, such as smoking cessation, have applicability across a number of conditions; often, however, quality measures suffer from a lack of applicability past one or two similar clinical areas. Indicators are therefore often targeted at specific diseases, particularly those with a significant patient population, within which such schemes may have a greater potential impact. P4P is suited to standardised, well-defined procedures; it is less well suited when measures of performance are difficult to define, obtain or evaluate.[47]

Process measures should be within the provider's control to implement. For example, patient compliance with lifestyle changes such as diet or exercise would be difficult for a physician to monitor and verify.[52] P4P measures should be visible in terms of accountability, applicability and effectiveness; there may exist significant gaps between processes that can be viably quantified and subsequent patient health outcomes.

Quality measures will be most easily monitored and reported at the point at which providers collect appropriate data as part of standard procedure. Providers need to ensure that they are collecting the relevant data for measurement; calculating risk adjustments for patients, for example, may require non-standard, supplementary demographic or clinical information.

Measures should be designed such that at the point of evaluation there exists a clear distinction between high- and low-performing participants;[34] a reduction in the spread of achievement across providers is a key identifier of improvement and a need to amend the scheme's design. In the absence of baseline variability, evidence of provider improvement is less clear.[53]

Rewarding care quality in a limited number of measures may lead to a more narrow focus on only incentivised measures or conditions, whereas larger bundles of measures may induce more general improvements in care quality. However, using a large number of measures increases the overall size and burden of the scheme and correlations between indicators are more likely to exist in a larger set, thereby increasing the risk of large gains and losses between participants.[54] Increasing the number of measures

also dilutes the importance attached to each, which may result in a lack of effort and therefore improvement, if the influence of a measure on overall performance is perceived to be insignificant.

Locally developed measures have the potential benefit of greater relevance to local health needs. Providers can develop P4P schemes aligned to their own long-term QI efforts, with local schemes used to pilot indicators for later use at national level. In promoting CQUIN to providers, the NHS advocated the new framework as 'an opportunity for commissioners and providers to focus on delivering higher levels of quality of care for their populations, rather than responding to centrally directed targets'.[55]

National measures, however, allow participants to benchmark progress and achievement against other providers and are more likely to represent a cohesive, comparable set of quality standards. Evidence from the CQUIN evaluation suggests that, although local contribution to quality measures was considered vital to the success of the scheme, frontline clinicians were rarely encouraged to involve themselves in the development of CQUIN goals and the technical design may have been better suited as a centrally directed initiative.[22] A mixture of national- and local-level targets may be best suited – although clinicians may quickly tire of the multiple requests for data potentially resulting from schemes in place at both local and national levels, with a risk of overlap or conflict in quality measures as their volume increases.

Provider case mix issues

The significant resource requirements of unusual patient case mixes are likely to affect achievement potential for P4P participants, with slower improvement rates and lower achievement levels. For example, specialist providers may treat and manage patients with more severe diagnoses and may therefore have greater difficulty meeting certain process and outcomes indicators. Participants in areas with greater poverty will perform less well on outcomes indicators that measure risky behaviours such as smoking and drinking levels. Low levels of education are likely to affect patient adherence and co-operation, thereby slowing improvements in process measures relying on actions from the patient, such as attending support groups or adhering to medication.[56]

Setting risk adjustments across providers creates a method for standardising performance across different population groups. However, risk-adjustment methods represent an additional administrative cost and, for the most part, remain underdeveloped and unsupported by clinicians.[57] Categorising participants into comparator groups based on patient demographics would limit benchmarking to a smaller number of similar providers and is therefore limited to situations where a sufficient number of comparator providers exist.

Pay-for-performance schemes often incorporate a system of exclusion reporting for patients considered to be ineligible for receipt of one or more quality indicators (such as a contraindication for a particular pharmacological product or opting for palliative care instead of treatment). Such criteria are seen as important in ensuring that patients are not given unsuitable care and allowing health-care professionals to exercise their clinical judgement without fear of being penalised. Eligibility criteria must, however, be clearly defined in order to avoid providers viewing exclusion reporting as an opportunity to exclude seriously ill or non-compliant patients who would otherwise reduce achievement in quality measures.

Unintended consequences and pay for performance

Exception reporting and patient selection

More sick or less adherent patients may be excluded from treatment within a P4P scheme, on the basis that such patients risk contributing negatively to the provider's quality measures.[52] The potential for manipulation of exception reporting, in which providers exclude certain patients who, on face value, should have been included in the initiative, has attracted studies into whether or not providers manipulate this to their personal advantage.

Evidence from the QOF suggests that the majority of general practitioners (GPs) did not participate in unwarranted exception reporting, despite some wide variations in the level of reporting across practices.[58] Gravelle *et al.*,[59] however, identified that a small proportion of practices gamed exception reporting to maximise their income. The presence of perverse incentives induced by P4P is strongly suggested in surveys of US physicians,[60] particularly when risk adjustments are non-existent or considered inadequate and inaccurate.

Focusing on process or structural measures, rather than final outcomes, should discourage providers from selecting patients based on their predicted response to treatment. Commissioners could monitor for changes in provider case mix following the introduction of P4P or use risk-adjusted bonuses based on relative patient complexity, in order to discourage the exclusion of more severe patients. Both strategies, however, represent a time and cost burden.

Effort diversion and multitasking

Much like the issues encountered in fee-for-service reimbursement systems, participants may be encouraged to overuse incentivised measures for the sake of additional reward unless the scheme is properly monitored and audited.[61] P4P schemes may lead to the prioritisation of incentivised conditions or measures, at the possible expense of non-incentivised areas. Overall quality of care may decrease, owing to an inefficient redistribution of investment between indications resulting from P4P incentives.[62] However, evidence of such an effect is mixed. One study claims systematic effort diversion present in QOF,[42] whereas another, conversely, notes a relative absence of effort diversion, instead identifying substantial positive spillover effects (see *Positive spillover effects*) within the same policy.[63] Ensuring that incentive monies are not sourced from clinical areas in need of the finance should help to minimise effort diversion; measures that encapsulate the whole provider system, such as patient experience or ward hygiene ratings, would also be less discriminatory in nature.

Positive spillover effects

Some QI measures have the potential to bring broader quality changes to non-incentivised conditions or aspects of care. A US P4P scheme incentivising completion of measures for diabetes patients under a managed care programme encouraged the rollout of the measures to patients across all health plans, despite the absence of an incentive provision for such patients.[52] Such positive spillover effects are likely to be amplified when participants are encouraged to communicate and collaborate with non-participating departments and providers, and when quality measures are applicable across a number of indications (such as discharge documentation) rather than strictly limited to incentivised conditions.

Supporting levers to accompany the financial incentives

Performance reporting

A number of P4P schemes include a reporting element, enabling internal staff and external stakeholders to view provider performance. An effectively designed public reporting initiative can have significant effects on commissioner and provider behaviours, even in the absence of a monetary incentive,[64] and can also act as a useful preparatory tool for providers that are new to a P4P scheme.[65] The sole act of reporting quality measures can encourage the development of standardised reporting systems.[65] De-anonymised reporting measures at individual level may also reduce problems with accountability and free-riding in group incentive schemes. Even if bonuses are distributed to individual physicians or providers, performance reporting can be produced at 'multiple levels of the care delivery system – physician, physician group, hospital, community – to identify gaps in performance and foster accountability at each level'.[66]

However, much like a reward scheme, performance reporting carries a risk of unintended consequences, such as gaming the system and patient selection. Indeed, physicians often react more negatively to a P4P scheme that includes external reports, with greater concern placed on the quality of the data and measurement.[67] As quality measures are often a small subset of a provider's services, publication of individual performance may be considered to provide an incomplete picture of care provision.[68]

Public reports may also be difficult to interpret for their intended audience. Composite scores and statistical and clinical derivations of provider achievement may hinder the use of public reports as decision-making tools for patients.[68] Reports that use a system of ranking participants may be misleadingly detrimental to low performers in schemes where provider scores exhibit limited variability. Regarding the use of reporting procedures as a complement to financial incentives, provision of bonuses allows participants to further invest in QI, and may therefore accelerate changes in care above that achievable with performance reporting alone.[69]

Feedback provision for participants

Negative feedback can reduce individuals' perceptions of their competence, leaving them feeling demotivated.[70] Positive feedback can make people feel happier and more competent. However, although praise may increase motivation, the relationship between feedback and performance is complex, with feedback that supplies the correct solution more effective at improving performance than praise.[71] A moderate supply of timely feedback,[72] complemented with serviceable suggestions for improvement, may be an optimal method for reporting results of P4P performance.[65,71] Such feedback may be best originating from the clinicians themselves, rather than commissioners, in order to ensure actionable methods for improvement.[69]

Funding pay-for-performance initiatives

The source of funding for P4P schemes has implications on how providers view such programmes. Withhold schemes encourage the view that high-quality care should be standard practice, rather than an action warranting additional rewards. However, they may also encourage the view that P4P acts as a way for commissioners to hold back much-needed funds, a sentiment acknowledged in an evaluation of CQUINs.[73] There is also evidence that suggests that clinicians view withholds as unfair, coercive and contrary to the spirit of collaboration that should characterise P4P initiatives.[27] The use of a system whereby any unearned bonuses or capitation are paid out as bonuses to high performers on top of their original incentive value – termed a challenge pool in US literature[74] – might encourage motivation when participants receive a reward greater than that of the original withheld funds, although this has yet to be used systematically. Similarly, in a combined reward–penalty scheme the income generated from fines to low performers would be used as bonus monies for high performers. Expected returns from penalties also need careful calculation in order to avoid owing bonuses greater than available funds (see Kahn *et al.*[75] for an example of this in HQID).

Designing achievement targets

Measure targets should be based on the capacity of a provider to improve in the subsequent period; although wealthier, highly resourced providers can be set a more challenging target than providers experiencing greater cost constraints, targets should maintain achievability among all participants. For tournament-style schemes, this could mean the grouping of providers based on resource availability, with rewards provided to winners in each group. For threshold-style schemes, participants may each work within their own set of performance thresholds or, again, may share targets with other similar providers. Alternatively, instead of benchmarking to competitor providers, targets may be set based on an individual participant's performance in the preceding period (with a predetermined level of performance in the baseline period). Nevertheless, targets for all providers should lead to sufficient QIs (and, in turn, health benefits) so as to justify the opportunity cost of the scheme.

However, participants may become aware that targets in a subsequent period are based on current performance and may, therefore, reduce effort levels in the preceding period in order to maintain a lower future target. Furthermore, if targets do not actively encourage low performers to catch up with higher achievers, performance disparities may become institutionalised across providers. Targets must therefore be based on both the capacity and the relative need of the provider to improve relative to its peers.

Phasing in pay-for-performance schemes

In the early stages of a P4P scheme, when provider engagement is still in formation, the intense focus on specific procedures of care may be a shock for participants not yet acquainted with the nature of P4P.

Larger schemes require the recruitment of, and support from, a greater number of health-care professionals, which may be difficult to achieve from the outset. The use of a pilot system has therefore been suggested as a valuable method for phasing in P4P initiatives.[65] This could involve testing the scheme in a limited geographical area or within a select number of indications or providers. In addition, baseline data during the piloting period can be used to set benchmarks and performance targets for when the scheme is fully implemented. Participation in the scheme may be voluntary for a set period, so that those providers requiring additional time to bring internal systems up to requirement need not be penalised in the interim.

Alternatively, participants could be incentivised on quality measure reporting levels alone (i.e. with no targets set for performance in the quality measures). Providers are then able to learn which areas of their infrastructure need investment prior to entering a full P4P scheme. Such a system of pay for reporting (P4R) was implemented in the USA prior to HQID, albeit not with the original intention of using it as a transition method into P4P. However, authors of a study into the Physician Quality Reporting Initiative concluded that P4R would act as a useful preparation technique for providers subsequently involved in P4P.[65]

The process by which providers join a P4P scheme will affect the ease with which performance can be evaluated. As discussed previously, allowing providers to volunteer to participate in a scheme is likely to attract participation from those who expect to obtain a net benefit from opting in. In turn, extrapolating performance potential and applying to the (non-participating) remainder of the provider network may bias a priori estimates upwards. Mandatory participation of all providers, however, removes the ability to compare performance with a control group of non-participants, so that the counterfactual performance across the same time period may be determined.[76]

Random selection into a compulsory scheme can instead ensure that the scheme maintains a mixture of providers operating at various levels of baseline performance. Subsequent achievement in P4P should therefore more accurately reflect the achievement potential of non-participating organisations. Alternatively, limiting performance to a small number of providers, selected on the basis of sharing some common characteristic, creates the ideal basis for a natural experiment. This is the case with our evaluation and enables us to compare performance in AQ providers in the north-west of England with those in the rest of the country. By observing performance across comparator organisations over the same time period, the relative effects of the P4P scheme may be isolated from national trends in performance. Ensuring that other contemporaneous shocks are accounted for when evaluating performance is vital, as is the ability to gather equivalent performance data from non-participating organisations from baseline onwards. In the case of AQ, although performance on process measures outside the NHS North West is not routinely available, we obtained outcomes (i.e. mortality rates) and use those for comparative purposes.

Quantifying achievement in pay for performance

Performance in P4P is often quantified by amalgamating achievement on individual indicators into a single composite score. Although methods for developing this score are numerous, two of the most frequent are the composite process score (CPS) and the appropriate care score (ACS).[77] The CPS represents the proportion of situations in which P4P measures have been appropriately administered. The ACS represents the proportion of patients receiving all care-quality measures for which they are eligible (*Figure 1* shows an example calculation of each in AQ). Performance as measured by an ACS is likely to be poor if one or more measure is particularly difficult to complete, suggesting the use of a CPS in the initial stages of a scheme when measures are often re-evaluated for feasibility. The ACS is particularly useful when performance in the CPS has little variation and when there exists scope for additional measures: 'it is more difficult to provide all the required measures of a large set than a small one'.[66] This makes the ACS more challenging as more measures are added. If measures complement one another, in the sense that the completion of a full set of measures leads to a superior outcome for the patient than completion of single measures alone, the ACS would be the appropriate score to use.

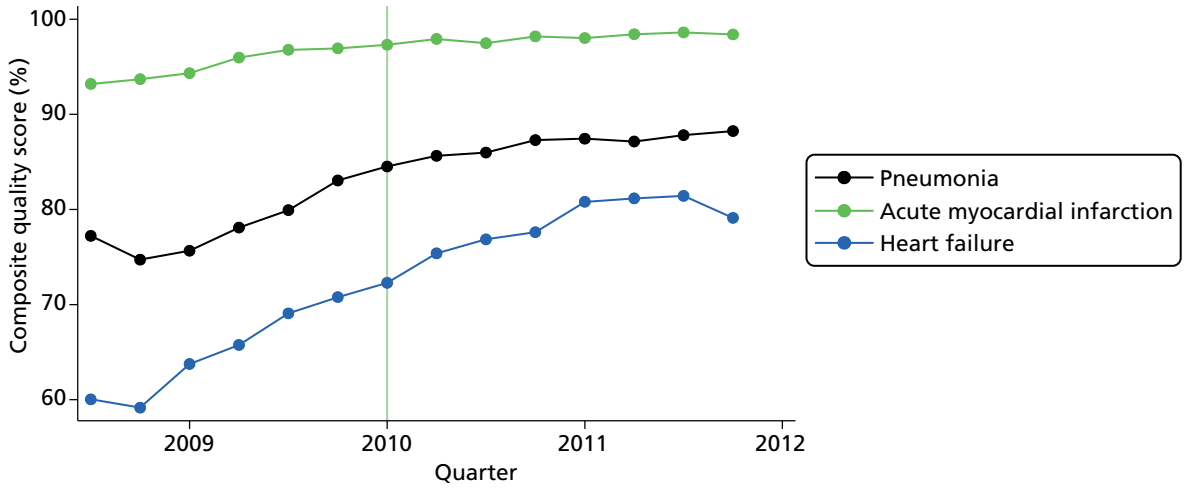


FIGURE 1 Average quarterly hospital achievement for each condition across indicators. Note that the vertical line indicates start of long-term period (month 19 of the programme).

In amalgamating indicator-specific scores into a composite value, it may be desirable to weight each indicator relative to its clinical benefit or financial cost of completion. As a result, performance in a heavily weighted indicator will influence substantially a provider's subsequent achievement. From the perspective of commissioners, however, individual weightings often represent additional analysis time and costs; from the perspective of clinicians and patients, they represent a complex obstacle to understanding performance.[78]

Adjusting the pay-for-performance scheme over time

Some targets may require amending or retiring in order to reflect new evidence, such as those relating to clinical effectiveness or safety (see Reeves *et al.*[79] for an example in the QOF). Some quality measures may become more difficult or impractical to attain, owing to changes in either the internal provider environment, such as the supply of hospital beds (see CMS[80] for an example in pneumonia), or the external environment, such as changes in socioeconomic conditions.[81] For most measures there exist natural ceiling or plateau effects and decreased returns to effort and investment, at which point financial incentives will no longer be sufficient to improve quality further. Finally, some incentivised behaviours may become so ingrained into standard practice that rewards are no longer necessary: a justification for re-evaluation again noted in a formal summary of QOF.[82] This may be seen as the ultimate success for a P4P programme.

If attainment in a particular measure is peaking at a suboptimal level, further investment or higher rewards may be required. As noted previously, certain internal and external barriers to achievement may mean that the marginal cost of initiating a quality measure on additional patients outweighs the marginal return in incentive. Inducing greater effort would therefore require a rebalancing of incentive value. Performance on a measure may be sufficiently high to warrant its discontinuation. Rather than an immediate removal of an indicator, however, it may be more appropriate to phase out the performance measure. If quality metrics are expected to change radically on a regular basis, support from providers may be more difficult to achieve, particularly for targets requiring significant investment, under the expectation that insufficient returns would be seen within such a short implementation period.[26]

This idea of gradually evolving P4P measures has been used in QOF. Instead of requiring participating GPs to measure the blood pressure of patients with diabetes, for example, participants were subsequently required to ensure that the patient's blood pressure was maintained below a predetermined threshold for the 15 months preceding evaluation. This has the benefit of maintaining focus on the same objective across both measures, while shifting the goalposts in the level of effort required from the provider.[79] Furthermore, a gradual shift from a process to outcomes measure should allow low-performing providers

to continue towards the original (easier) quality measure, at the same time as encouraging continuous improvement from high performers.

Retired indicators should continue to be monitored in order to check for depletions in performance following their removal, as occurred in one process measure for the QOF.[79] Such occurrences may be a result of a lack of support from participants prior to incentive provision or because they represented a significant burden on workloads.

As we explain in this report, the AQ programme embodied many aspects of what might be seen as desirable design features of P4P schemes. However, this is no guarantee of success, as the implementation of P4P initiatives is likely to require substantial changes to practice. Evidence suggests that efforts to introduce such change frequently fall short of intentions, resulting in variable outcomes.[83]

Conceptualising change: overcoming challenges

In addition to the burgeoning literature on change in health-care settings, there are also various frameworks conceptualising the change process. In our proposal for funding we chose to use a framework developed by Bate *et al.*[84] as part of their examination and explanation of quality journeys in a range of health-care organisations internationally. They discuss the findings in terms of six common challenges facing organisations undertaking QI initiatives in health-care settings. They conceptualise change as an ongoing journey and suggest that there is no one best way to achieve intended outcomes. The challenges they identify emphasise the multidimensional nature of change. What is important, they suggest, is the ability to address multiple challenges simultaneously and to adapt solutions to the local organisational context.

The challenges in their framework are:

> *structural (organising, planning and co-ordinating quality efforts)*
>
> *political (addressing and dealing with the politics of change surrounding any QI effort)*
>
> *cultural (giving quality a shared, collective meaning, value and significance within the organisation)*
>
> *educational (creating a learning process that supports improvement)*
>
> *emotional (engaging and mobilising people by linking QI efforts to inner sentiments and deeper commitments and beliefs)*
>
> *physical and technological (the design of technological infrastructure that supports and sustains quality efforts).*
>
> *Bate* et al.[84]

Addressing structural challenges is described as requiring strong and decisive executive leadership giving clear direction. A focus around QI in general and specific programmes in particular is likely to improve the impact of QI processes. In addition to structures to facilitate leadership and whole systems working, developing structures to overcome challenges related to data collection and monitoring systems is seen as hugely important. Furthermore, the provision of slack resources[85] to enable staff to stand back from everyday pressures is likely to facilitate the implementation of QI processes.

Boundary spanning refers to roles with a 'hybrid, dual bridging aspect, such as clinical leader/manager, which allow for lateral contact and communication between different groups and the linking of resources, people and ideas'[84] around QI efforts. In a professional bureaucracy[86] such as the NHS, institutionalised

occupational boundaries and related epistemic communities[87] and subcultures result in a landscape which is replete with 'structural holes'.[88] The holes between two groups do not mean that people on either side of the hole are unaware of one another, but that they are focused on their own activities and do not pay attention to what people outside their group are doing. 'Holes are buffers like an insulator in an electric circuit. People on either side of a structural hole circulate in different flows of information. Structural holes are thus an opportunity to broker the flow of information between people and control the projects that bring together people from opposite sides of the hole'.[89]

Political challenges relate to dealing with conflict and opposition, stakeholder buy-in and engagement and securing commitment to a common agenda for improvement. Political issues to consider include the extent of empowerment (of staff and patients to exercise control over their environment), clinical engagement, politically credible leadership and relationships between clinicians and managers in terms of their agreement to work together on improvement initiatives. Additionally, making and maintaining constructive relationships with relevant external partners is viewed as an important part of overcoming political challenges.

These are challenges that are connected with established cultures of working practice and the way in which work is performed and perceived by staff members. Bate *et al.*[84] identify various strands in relation to culture, which are important for QI. These include a group collaborative culture, which refers to a strong group culture that promotes teamwork and co-operation between staff and places a premium on values such as respect, integrity, trust, pride, honesty, inclusion and openness. Formal culture refers to an emphasis on formalised disciplines to ensure efficiency and effectiveness. A culture of mindfulness involves a working environment that promotes vigilance and reflective practice while discouraging habitual and mechanical approaches to work. A scientific culture concerns a commitment to evidence-based practice, and a culture of learning refers to a culture that values risk-taking and experimentation, constantly encouraging people to do more and do it differently, developing and sharing new knowledge, skills and expertise.

The educational challenges refer to a process of '[e]stablishing and nurturing a continuous learning process in relation to quality and service improvement issues, including both formal and informal mentoring, instruction, education and training, and the acquisition of relevant knowledge, skills and experience'.[84] In addition to education and training generally, a focus on evidence- and experience-based learning, as well as experimentation and piloting is emphasised. These involve learning and developing new understanding from analysis of routine evidence and data, together with a willingness to try out new ideas and assess their impact.

Emotional challenges relate to the task of '[e]nergizing, mobilizing, and inspiring staff and other stakeholders to want to join in the improvement effort by their own volition and sustain its momentum through individual and collective motivation, enthusiasm and movement'.[84] Important aspects in any efforts to overcome emotional challenges are the use of champions to engage peers, as well as quality activists driving improvement via informal networks. Quality should be seen as a mission or calling, rather than merely a job. It is crucial that people engage with their hearts, as well as with their heads.

Physical and technological challenges refer to the '[d]esign and use of a physical, informational and technological infrastructure that improves service quality and the experience of care'.[84] The emphasis here is on the built environment and the extent to which it supports and encourages (as opposed to inhibits) QI efforts. Additionally, supportive information technology, in terms of both its functionality and its location, is a key aspect enabling organisations to drive and maintain QI processes.

Conceptualising change: understanding processes and mechanisms

Studies of change in health care often provide explanations that describe the what.[90] Exploration of the how and why can range from black-box explanations to fairly detailed descriptions which distil a wealth of observational and interview data to identify common features. In relatively few cases (although this is becoming increasingly common[91]), researchers identify underlying mechanisms that lead to particular outcomes in specific contexts. In practice, however, identifying what constitutes a mechanism as opposed to aspects of the context can be difficult.[92] Additionally, identifying what counts as relevant in relation to contexts that span micro-, meso- and macro-level factors is no simple task. For this study, therefore, although we have attempted to go beyond describing surface characteristics or common ingredients, we have approached the understanding of the how and why of the programme from a different, but complementary, angle. This entails consideration of different kinds of explanation as we explain in the next section.

In order to understand how initiatives such as P4P programmes work, it is important to consider the broader field in which these are situated. 'An organisational field can be defined as a social area in which organisations interact and take one another into account in their actions. Organisational fields contain organisations that have enduring relationships to each other'.[93] In this case, the field is the social area where the 24 health-care providers operate and this area includes other organisations with which they interact.

Fields are characterised by formal rules, but understanding what happens within the field requires consideration of other factors.[94] It is important to take into account the part played by network structures and cognitive frames. For example, as part of the process of applying and implementing formal AQ rules, staff from participating organisations were involved in collaborative learning events similar to those commonly used within a QI collaborative approach. This process can be conceptualised as part of a process of creating new network structures given that new staff members (recruited specifically for the AQ programme) initiated these events. These events also involved bringing people together to develop collective understanding (cognitive frames). To understand how this was related to change, it is important to examine who participated and what participation involved (e.g. in-person learning was chosen as opposed to webinars). Additionally, this 'what' does not merely mean the ingredients of these events. The meaning should also include network structures and rules of the field more generally, because change programmes such as AQ are not free-floating but operate as part of a broader organisational field. For example, an examination of network structures reveals, among other things, the position of those involved, some of whom will be more powerful in the field than others. Furthermore, although many of these people may never meet each other, there is often a dependency relationship between them. For example, staff who code data from patient records rely on the information clinicians provide therein. Even assuming that we can capture some or all of this, we need to go beyond formal rules and network structures to consider other factors that may exert an influence (*Box 1*).

BOX 1 Aspects of organisational fields that influence attempts at change

Formal rules: these provide a regulatory framework to guide behaviour.

Network structures: people are positioned in different spaces in the field, with some more powerful than others. Many of these people may never meet each other, but there is often a dependency relationship between them.

Cognitive frames: individual and collective perceptions of field values and activities.

In recent years, partly as a reaction to a reliance on structural explanations which underplay the active part played by individuals and groups, various studies of change in organisational fields have focused on the ways in which actors within the field have interpreted events and structures and how this has influenced activities. For example, institutional theorists conceptualise fields in terms of prevailing social norms[95] and are concerned with understanding how things within the field achieve legitimacy. The emphasis here is on shared meaning (e.g. Reay and Hinings[96]) and understanding how changes in what counts as important and legitimate within the field occur. Attending to interpretation and the development of collective cognitive frames is important, but focusing solely on shared meanings is likely to produce a partial explanation at best.

Furthermore, there is a tendency within such accounts to blur or ignore completely the distinction between the cognitive frames of groups and individuals and the network structures in which they are positioned. Some view perception and meaning in terms of network relationships where 'interactions between people gradually acquire an objective quality, and eventually people take them for granted'.[97] Such an approach implicitly recognises cognition and perception but makes it indistinguishable from network structures, thereby conflating different types of factors. Other explanations attempt to endogenise cognition. Networks are interpreted as 'networks of meaning'[98] which express mental maps of the structure of social relations. Here the existence and structure of connections are ignored, as what is important is the dominant interpretations that characterise the network. In such cases, a focus on perception and interpretation underplays or fails entirely to acknowledge the ways in which formal rules, on the one hand, and network structures, on the other, influence activities within the field.[94]

To understand how changes do or do not happen, it is therefore important to acknowledge all three sets of factors (rules, network structures and cognitive frames) and the interplay between them in a way that does not privilege one over another.

Using the literature to inform our analysis

The foregoing outlines lessons pertaining to scheme design in relation to P4P initiatives. However, although it is necessary to pay careful attention to design in order to increase the likelihood of success, it is unlikely to be sufficient. The burgeoning literature reporting problems with implementation in relation to attempts to change practice in health-care settings draws our attention to the challenges involved. We combine the lessons on scheme design with Bate's[84] framework and, in particular, the nature of challenges involved when exploring the implementation of the AQ programme. We also seek to go beyond the details of the AQ programme to conceptualise change in the organisational field more broadly drawing on Beckert's[94] ideas of how change occurs in such fields. We do not suggest that this is the only way to approach the issues. Instead, we suggest that it is a useful approach which enables us to consider change as a dynamic process involving the interplay of various factors and challenges in a way that resonates with our findings.

Chapter 3 Advancing Quality: background

In this section we present a background description of AQ. This description has been kept to the minimum information required to enable readers to understand the findings. An accessible and detailed description of AQ is available on the AQ website (see www.aquanw.nhs.uk). AQ has evolved over time and new indicators and clinical areas have been added, but this report is mostly concerned with describing and evaluating AQ in relation to the five clinical areas included at October 2008.

Defining quality

Quality of care, for the purpose of the financial incentive programme, at the outset, was intended to be measured in three different ways as follows:

- Clinical process and outcome measures in five clinical areas – acute myocardial infarction (AMI), heart failure, coronary artery bypass graft (CABG), pneumonia, and hip and knee replacement. (See *Appendix 1* for a full list of clinical process and outcome measures used in AQ.)
- Patient-reported outcome measures (PROMs) – patients undergoing elective hip and knee surgery are asked a series of questions about their health status before their procedure and a series of questions 6 months after their procedure.
- Patient experience of care provided – various approaches to measuring patient experience were trialled during the programme. These included two questions based on questions contained in the Hospital Consumer Assessment of Healthcare Providers and Systems survey instrument used in the USA. Patients were asked to rate the hospital on a scale of 1 to 10 and to indicate the likelihood (on a scale from definitely no to definitely yes) that they would recommend the hospital to family and friends. A further six questions, 'Six of the best', covered experience of service delivery, staff and information.

The choice of clinical areas was based on high-volume conditions and availability of metrics to measure processes and outcomes of care. The clinical measures used at the start of AQ are contained in *Appendix 1*. As an illustration of the nature of AQ measure, the community-acquired pneumonia (CAP) measures are listed in *Box 2*.

BOX 2 Community-acquired pneumonia

1. Percentage of patients who received an oxygenation assessment within 24 hours prior to or after hospital arrival.
2. Initial antibiotic selection.
3. Blood culture collected prior to first antibiotic administration.
4. Antibiotic timing, percentage of pneumonia patients who received first dose of antibiotics within 6 hours after hospital arrival.
5. Smoking cessation advice/counselling.

Payment rules

The first year of AQ was run as a pure tournament, with hospitals that scored in the top quartile on the incentivised quality metrics receiving a 4% bonus payment and those in the second quartile a bonus of 2%. For the next 6 months, financial incentives were awarded based on three criteria. Providers for which performance in this period was above the median score from the first year were awarded an attainment bonus. Those earning this attainment bonus were then eligible for two further payments, which were awarded to the top quartile of improvers and those that achieved quality scores in the top two quartiles. There were no penalties or withholds for poor performers during these first 18 months.

After the first 18-month period, payments were made under a new P4P framework that applied across the whole of England. Under this CQUIN framework, a fixed proportion of the hospital's expected income was withheld and paid out only if hospitals achieved required performance thresholds. The majority of topics, quality measures and threshold values were negotiated and agreed between the hospital and the primary care trust (PCT), the NHS organisation responsible for planning and commissioning health-care services on behalf of the local population.[14] However, the regional authority could also specify some CQUIN requirements, and the North West region included the AQ indicators in the CQUIN requirements of all 24 hospitals. Required levels of achievement were based on the quality scores that had been achieved by each hospital in the first year of AQ.

The potential total amounts of money linked to performance were kept constant throughout the period. In total, £3.2M in bonuses was paid to hospitals in the North West region for the first year and £1.6M was paid for the next 6 months. The transfer to the CQUIN framework meant that the P4P scheme effectively changed from bonuses to penalties. The CQUIN agreements were designed so that hospitals would lose £3.2M in total each year if they all failed to meet all of the targets for the five AQ conditions. At the same time, although comparative league table performance data continued to be reported, AQ was no longer a tournament-style system in that payment was not intended to be made to only a subset of providers.

Relative performance and payment was initially based on the CPS, an aggregate score that reflects the number of opportunities to do the right thing and the proportion that were achieved. The ACS is a reflection of what happened to individual patients. It measures the proportion of patients that received all of the relevant interventions (i.e. perfect care with regard to AQ measures). Although payment was based on CPS performance initially, recent changes mean that the ACS is used as a basis for payment.

Participants and set-up

All 24 acute trusts and the North West Ambulance Trust participated in AQ from its inception to the end of the period covered by the evaluation, although only four trusts undertook CABG. (During this period Trafford Healthcare NHS Trust was taken over by Central Manchester University Hospitals NHS Foundation Trust; therefore, the number of participating organisations was reduced to 23.) AQ was led from the (then) NHS North West Strategic Health Authority (SHA).

The AQ programme went live for all participating trusts on 1 October 2008, although only the clinical process and outcome measures were being collected from this date. PROMs data collection commenced on 1 January 2009 and patient experience data collection was anticipated to commence winter 2009/10. Seven trusts participated in a first-wave pilot. Wave 1 sites were recruited in May 2007 and site selection was informed by geographical location (because the intention was to obtain geographical spread), willingness to participate, readiness of systems, good evidence of partnership between primary and secondary care and organisation type. First-wave organisations received twice as much in set-up costs as second-wave trusts (£60,000 vs. £30,000).

The clinical-measures component of AQ was based on the HQID in the USA, which was a collaboration between the CMS and Premier Inc. (hereafter Premier), a nationwide organisation of not-for-profit hospitals. Following a competitive process, in November 2007 Premier was selected by the SHA as the partner organisation for AQ. Premier's role involved the provision of advice and support from staff based in the north-west of England as well as telephone and e-mail support from US-based staff who worked on the UK time zone. This was supplemented with occasional visits from these US-based staff to provide face-to-face training and advice workshops in the north-west of England. In addition, Premier provided software tools and data-management and reporting facilities adapted from its experience of helping to support the delivery of the HQID in the USA. Following Premier's decision not to tender for the new contract in 2010, an alternative provider undertook this role for the remainder of the period covered by this evaluation. [Clarity Informatics took on this role from December 2010. The new service provider designed a new online tool for data entry and analysis, called Clarity Assure, which has now replaced Premier's Quality Measures Reporter (QMR) tool.]

Seven trusts participated in a first-wave AQ pilot. Wave 1 sites were recruited in May 2008 and site selection was informed by geographical location (because the intention was to obtain geographical spread), willingness to participate, readiness of systems, good evidence of partnership between primary and secondary care, and organisation type. First-wave organisations received twice as much in set-up costs as second-wave trusts (£60,000 vs. £30,000).

Advancing Quality went live for all participating trusts on 1 October 2008, although only the clinical process and outcome measures were being collected from this date. PROMs data collection commenced on 1 January 2009 and patient experience data collection went live on 1 January 2010.

Data definition, submission and monitoring

The clinical-measures component of AQ is based on the US HQID. AQ measures are supported by a detailed data dictionary, compliance with which is intended to ensure standardisation within and between providers.

The NHS data for patients discharged within a particular month were sent to Premier 58 days after the end of that month. Once the data had been checked for completeness of general data elements, the patients were grouped to the five clinical areas (if they qualified). The data were then made available by Premier to participating providers using a web tool for data collection. These data showed patients for each of the clinical areas. Trusts were then able to complete data entry for the fields relevant to AQ measures, run reports on their data using this web tool, verify that their understanding of eligible patients tallied with Premier's data, check for missing data and resolve mismatches when appropriate. No changes were permitted after the resolution deadline and the process from month-end to the production of final reports by Premier took 158 days. Individual trusts submitted clinical data sets to Secondary Uses Service (SUS), the NHS comprehensive data repository. These data included information on dates of admission and discharge, primary diagnosis, procedure codes, age, sex, and so on, from their patient administration systems. This is a routine data-collection process for all trusts nationally. Patient data were extracted from SUS. For AQ, three organisations, the Greater Manchester Commissioning Business Service (CBS), the Cheshire & Mersey Contract Information Shared Services Unit and the Cumbria & Lancashire Contracting Information Service, extracted data sets for patients covered by their area of the north-west of England. CBS agreed to collate the three extracts, format them to fit Premier specifications and transmit them to Premier on behalf of all North West region providers. This enabled data extraction exercises to be reduced from 24 to three. This took place 45 days after the month-end. Preparing the data sets, including removing patient identifiers, meant that the data were not transmitted to Premier until day 58. Premier identified the relevant AQ population for each provider and then returned this to that provider on day 68. All providers had 30 days to enter data relating to AQ processes. For example, did the pneumonia patient receive the first dose of antibiotics within 6 hours after hospital arrival? Providers submitted these data via the web tool (QMR) and

30 days later the deadline for issue resolution was reached and the data set was closed (day 128). AQ reports were produced 30 days later (158 days).

Developments during the evaluation period

Commissioning for Quality and Innovation Payment Framework

From 1 April 2010, AQ became part of the NHS North West CQUIN payment framework, which meant that changes were made to payment rules. Whereas AQ was a tournament-style system that (in year 1) rewarded top performers, CQUIN involved setting provider-specific stretch targets that reward an agreed level of improvement from the previous year's baseline. This meant that all providers that achieved the agreed performance were eligible for CQUIN payment. The local stretch targets for each provider for AQ year 3 were set by the SHA AQ team, which used year 1 results as baseline owing to the unavailability of year 2 data at the time CQUIN targets had to be disseminated to providers (March 2010). In some cases, low year 3 targets resulted in payments to providers that had performance below the average performance of year 2.

In October 2010, the SHA AQ team was incorporated into the Advancing Quality Alliance (AQuA). This is a membership organisation established in the north-west of England to support QI work in the region. AQuA is funded by member organisations, which were asked to pay £55,000 each for 1-year membership in the first year, but subsequently paid £25,000 each per year. All North West England providers joined AQuA in the year following its creation. AQuA is a membership health-improvement organisation, whose vision involves supporting members 'to transform the health and quality of healthcare for the people they serve'.[99] (For more information about AquA, see www.aquanw.nhs.uk.) The AQ programme (initially funded by the PCTs) provides a large share (62%) of AQuA's income (£5,190,222 in 2011), with 4% of this spent on AQuA staff and the remainder on external consultancy support. In practice, for AQ providers, AQuA organises and runs collaborative events (such as AQ leads meetings and collaborative meetings in each clinical area), as well as facilitating the sharing and dissemination of collected data. AQ is regularly described as the flagship programme of AQuA. The creation of Clinical Commissioning Groups (CCGs) means that AQuA now has to negotiate with CCGs on the subject of payment and focus of work.

Advancing Quality clinical areas and metrics

Various changes were made to measures during the evaluation period. For example, the beta-blocker on arrival for AMI patients measure was phased out from July 2009. The taking of blood cultures for pneumonia patients was dropped from October 2012. New measures, such as the requirement to document a CURB-65 score (a measure of pneumonia severity) from April 2011, were added as part of an ongoing process of evolution. The approach was to trial measures in shadow form to test feasibility, after which they were incorporated into the incentive scheme.

Furthermore, AQ expanded its scope to three additional clinical areas: stroke, mental health and dementia. Stroke came into effect in October 2010 and was incentivised under the CQUIN scheme. AQ stroke built on the Stroke 90 : 10 initiative. This was a regional initiative launched in October 2008 by the Health Foundation, the Stroke Association and the Royal College of Physicians with the target of achieving a Sentinel Audit Score in the 26 participating hospitals of 90% by 2010, from a baseline in 2004 of 56%. It ran for 2 years from October 2008 to October 2010 and managed to drive up the Sentinel Audit Score for all participating trusts to just under 90%. Dementia and first-episode psychosis were introduced with effect from January 2011. However, the focus of this evaluation, so far, has been on the initial AQ measures.

Patient-experience measures

Owing to the evolving nature of the patient-experience measures (PEMs) process and problems with measures and data-collection tools, PEMs have not formed part of this evaluation.

Initially, the administration and collection of PEMs data were very challenging for providers and resulted in particularly low return rates among all participating trusts. As a result of this, the SHA AQ team organised a PEMs collaborative event in the middle of July 2010, at which it was decided that the requirement for patients to complete the PEMs survey on the day of discharge was to be relaxed to any day between admission and discharge. In that meeting it was also decided that PEMs surveys should be narrowed down to one question, that is 'did you receive the care that mattered to you?', although this would not come into effect until April 2011.

Dr Foster's involvement in the administration and collection of PEMs data also came to an end by January 2011, with the process subsequently being overseen by the CBS, which agreed to design a new web-based data-capture tool. From April 2011 PEMs data collection was undertaken by providers using existing in-house tools and processes.

Chapter 4 Quantitative methods and findings

The study used a combination of qualitative and quantitative methods. The impact of AQ was measured in quantitative terms. This focused on impact on outcomes and helped us understand what happened. This quantitative analysis was combined with qualitative data collection and analysis (described in *Chapter 5*), aimed at shedding light on how and why these impacts occurred. In addition, a cost-effectiveness analysis and an analysis of the distributional impact of AQ were also undertaken.

Impact on mortality in the short term

Impact on mortality in the short term: methods

We obtained patient-level data from national Hospital Episode Statistics (HES) from the NHS Information Centre (now the Health and Social Care Information Centre) for all patients in England treated for three of the five conditions included in the scheme. We did not include hip and knee surgery because mortality following elective joint replacement is < 1%. We also did not consider CABG because only four hospitals out of the 24 in the North West region of England undertook this procedure.

Hospital Episode Statistics in England records deaths that occur within any hospital. We focused on all deaths that occurred within 30 days of admission. Published national statistics show that over 90% of deaths within 30 days for the incentivised conditions occur in a hospital. To check that there were no changes in discharge policies that might have led to more deaths outside hospitals, we also analysed changes in the proportions of patients discharged to care institutions rather than their own homes.

We obtained equivalent data for patients admitted with six primary diagnoses which were not incentivised. These conditions were chosen based on published statistics at national level[13] to meet the following criteria: (1) not clinically linked to any incentivised condition; (2) sufficient volume (over 9000 admissions in England per year); (3) 30-day mortality over 6%; and (4) more than 80% of deaths within 30 days occurring in a hospital.

Six diagnoses met these four criteria and were treated as reference conditions: acute renal failure [*International Classification of Diseases,* Tenth Edition (ICD-10)[100] codes beginning N17]; alcoholic liver disease (K70); intracranial injury (S06); paralytic ileus and intestinal obstruction without hernia (K56); pulmonary embolism (I26); and duodenal ulcer (K26). We excluded from the reference group all patients with a diagnosis included in the incentive scheme on any of their admissions over the 3-year study period. Our comparators therefore include two mutually exclusive sets of patients: one set with an admission covered by the scheme in hospitals not included in the scheme; and one set with an admission for a reference condition and no mention of any diagnosis covered by the scheme on any admission within the 3-year period.

Data were obtained for patients admitted over a 3-year period from 1 April 2007 to 31 March 2010. This 3-year period includes 18 months prior to the introduction of the scheme and the first 18 months of its operation. The data set includes patients treated at the 24 NHS hospitals in the North West region and the 130 NHS hospitals in all other regions of England which admitted more than 100 patients with each condition over the 3-year period. The final sample contains 410,384 patients with pneumonia, 201,003 patients with heart failure, 245,187 patients with AMI and 241,009 patients with non-incentivised conditions, treated at 154 different hospitals. Hospital characteristics were obtained from the websites of national regulators (e.g. Healthcare Commission and Monitor) and the NHS Information Centre.

We calculated expected risks of mortality using regressions at patient level which included sex and age (the primary ICD-10 diagnosis code); 31 Elixhauser comorbidities derived from secondary ICD-10 diagnosis codes; the type of admission (emergency or transfer from another hospital); and the location from which

the patient was admitted (own home or institution). The analysis of risk-adjusted mortality was undertaken on data aggregated by the quarter of the year and by admitting hospital.

We tested whether or not the incentives had an impact on mortality in three ways: (1) a between-region difference-in-differences analysis comparing the changes in mortality over time between the North West region and the rest of England for incentivised conditions; (2) a within-region difference-in-differences analysis comparing the changes in mortality over time between the incentivised and non-incentivised conditions in the north-west of England; and (3) by estimating a triple difference, comparing the changes over time in mortality between the incentivised conditions in the North West region and the rest of England and between the incentivised and non-incentivised conditions. The triple-difference analysis captured the effect of the programme on mortality for the incentivised conditions in the North West region controlling for the effects of changes over time in mortality for the incentivised conditions owing to factors other than the initiative itself, in addition to changes over time in overall mortality in the North West region and differences in mortality between the incentivised and non-incentivised conditions between the North West region and the rest of England.

We estimated the effects for all three incentivised conditions combined and then separately for each condition. Each analysis allowed flexibly for time trends using a binary variable for each of the 12 quarter-years, and for hospital differences using a binary variable for each hospital, and includes an interaction term between the intervention group and the post-implementation period.

Impact on mortality in the short term: results

The characteristics of patient populations in the North West region and the rest of England before and after the scheme's introduction are shown in *Table 1*.

For all conditions, patients in the North West region were slightly younger but had more comorbidities. Similar changes over time in patient volumes and patient characteristics are observed in both areas. The profile of hospitals in the North West region is similar to the rest of England (*Table 2*), with a slight tendency for fewer hospitals in the North West region to have received the lowest ratings by the national regulators for overall care quality and financial management in 2007.

Risk-adjusted mortality rates for all of the conditions we studied decreased over the study period in both the North West region and the rest of England. The reduction in mortality for incentivised conditions was greater in the North West region than in the rest of England, reducing from 21.9% to 20.1% in the North West region and from 20.2% to 19.3% in the rest of England (*Table 3*). Compared with non-incentivised conditions within the North West region (within-region difference-in-differences analysis, *Table 3*), there was a significant reduction in overall mortality for incentivised conditions of 0.9 percentage points (95% CI 0.1 to 1.7 percentage points), comprising a statistically significant reduction in pneumonia and non-significant reductions in the other two conditions. Comparing mortality for the incentivised conditions with mortality for the same conditions in other regions (between-region difference-in-differences analysis, *Table 3*), there was a significant reduction in overall mortality in the North West region of 0.9 percentage points (95% CI 0.4 to 1.4 percentage points), again made up of individually significant reductions in pneumonia and non-significant reductions for the other two conditions. Combining these two (triple-difference analysis, *Table 3*) suggested an overall reduction in mortality of 1.3 percentage points in the North West region (95% CI 0.4 to 2.1 percentage points), with a similar pattern for the individual conditions. The reduction in mortality over the period studied for non-incentivised conditions was not significantly different between the North West region and the rest of England.

Our finding that risk-adjusted mortality for the non-incentivised conditions reduced by similar amounts in the North West region and in the rest of England suggests that our findings are not explained by higher preintervention mortality or by a general improvement in quality or reduction in case-mix complexity in the study region. Nonetheless, we undertook a wide range of further analyses to test the robustness of our findings. There were no significant changes in the proportion of patients discharged to care homes and all

TABLE 1 Characteristics of patients before and after the programme in the North West region and the rest of England

Health condition	North West region			Rest of England			Difference in differences
	Before	After	Difference	Before	After	Difference	
AMI							
Number of patients	20,080	18,753	–6.6%	104,915	101,485	–3.3%	–3.3%
Mean age (years)	70.2	70.2	0.1	70.3	70.7	0.4	–0.4
Percentage aged > 75 years	43.2	43.3	0.1	44.1	44.9	0.9	–0.8
Transfer from another hospital (%)	6.9	5.9	–1.0	10.8	8.2	–2.6	1.6
Average Elixhauser[a] conditions	1.60	1.73	0.13	1.51	1.68	0.17	–0.04
Discharged to care home (%)	2.9	2.7	–0.2	1.7	1.7	0.0	–0.2
Unadjusted mortality in 30 days (%)	12.4	11.0	–1.4	11.0	10.7	–0.3	–1.1
Heart failure							
Number of patients	15,445	15,476	0.2%	83,555	86,572	3.6%	–3.4%
Mean age (years)	75.9	76.6	0.7	77.5	78.1	0.6	0.1
Percentage aged > 75 years	61.5	64.0	2.6	67.2	68.8	1.6	0.9
Transfer from another hospital (%)	1.3	1.1	–0.2	1.7	1.5	–0.2	0.0
Average Elixhauser[a] conditions	2.28	2.43	0.15	2.17	2.40	0.23	–0.08
Discharged to care home (%)	4.0	4.1	0.1	3.3	3.2	–0.2	0.2
Unadjusted mortality in 30 days (%)	17.9	16.6	–1.3	16.6	16.1	–0.6	–0.7
Pneumonia							
Number of patients	28,275	36,428	28.8%	150,531	195,204	29.7%	–0.8%
Mean age (years)	71.8	72.4	0.6	72.4	73.1	0.7	–0.1
Percentage aged > 75 years	54.0	55.6	1.6	56.5	58.0	1.5	0.1
Transfer from another hospital	0.8%	0.7%	–0.1%	1.2%	1.0%	–0.2%	0.1%
Average Elixhauser[a] conditions	1.84	1.99	0.15	1.69	1.91	0.21	–0.06
Discharged to care home (%)	6.5	6.6	0.2	4.9	4.9	0.0	0.1
Unadjusted mortality in 30 days (%)	28.0	25.9	–2.2	27.2	26.3	–0.9	–1.3
Non-incentivised conditions							
Number of patients	16,705	18,407	10.2%	98,348	107,581	9.4%	0.8%
Mean age (years)	61.8	62.6	0.7	63.4	64.2	0.8	–0.1
Percentage aged > 75 years	30.6	32.6	2.0	34.4	35.9	1.5	0.4
Transfer from another hospital (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average Elixhauser[a] conditions	1.48	1.63	0.15	1.31	1.49	0.18	–0.03
Discharged to care home (%)	3.8	3.8	–0.1	2.8	2.8	–0.1	0.0
Unadjusted mortality in 30 days (%)	13.3	13.0	–0.3	11.7	11.0	–0.7	0.3

a Comorbidities that are predictive of mortality derived from secondary ICD-10 diagnosis codes.[17]

TABLE 2 Characteristics of hospitals in the intervention and control regions

	North West region		Rest of England	
Hospital characteristic	***n***	**%**	***n***	**%**
Scale and scope				
Teaching/specialist	5	21	21	17
Large general	7	29	37	30
Medium general	8	33	40	32
Small general	4	17	26	21
Foundation trust status[a]				
Non-foundation trust	17	71	90	73
Foundation trust	7	29	34	27
Rating of overall care quality in 2007[b]				
Excellent	7	29	37	30
Good	13	54	59	48
Fair/weak	4	17	27	22
Rating of financial management in 2007[c]				
Excellent	11	46	45	37
Good	7	29	30	24
Fair/weak	6	25	48	39

a Foundation trusts are hospitals that have been approved by the national regulator to have additional managerial and financial freedoms. We classify hospitals by their status in 2007.
b Composite rating of performance in 2007 by the national regulator (the Healthcare Commission) against core standards, existing national targets and new national targets for quality.
c Composite rating of performance in 2007 by the national regulators (the Healthcare Commission and Monitor) on financial standing, management and control.

TABLE 3 Changes over time in risk-adjusted mortality for the incentivised and non-incentivised conditions in the North West region and the rest of England

	North West region			Rest of England			Within region	Between region	
Health conditions	Before (%)	After (%)	Difference	Before (%)	After (%)	Difference	Difference in differences (95% CI)[a]	Difference in differences (95% CI)[b]	Triple difference[c]
Non-incentivised	13.1	12.1	–1.0	12.0	10.7	–1.3	–	0.3 (–0.4 to 1.1)	–
Incentivised	21.9	20.1	–1.8	20.2	19.3	–0.9	–0.9 (–1.7 to –0.1)	–0.9 (–1.4 to –0.4)	–1.3 (–2.1 to –0.4)
AMI	12.1	10.7	–1.4	11.3	10.4	–1.0	–0.4 (–1.3 to 0.6)	–0.3 (–1.0 to 0.4)	–0.6 (–1.7 to 0.4)
Heart failure	18.8	17.5	–1.3	16.9	15.8	–1.1	–0.4 (–1.5 to 0.7)	–0.3 (–1.2 to 0.6)	–0.6 (–1.8 to 0.6)
Pneumonia	29.4	27.0	–2.4	27.1	26.3	–0.7	–1.5 (–2.5 to –0.5)	–1.6 (–2.4 to –0.8)	–1.9 (–3.0 to –0.9)

a The within-region difference in differences represent the change over time for the incentivised conditions minus the change over time for the non-incentivised conditions in the north-west of England.
b The between-region difference in differences for each condition represent the change over time in the North West region minus the change over time in the rest of England.
c The triple differences represent [(the change over time in the incentivised conditions in the North West region minus the change over time in the incentivised conditions in the rest of England) minus (the change over time in the non-incentivised conditions in the North West region minus the change over time in the non-incentivised conditions in the rest of England)].
The within-region and between-region difference in differences and the triple differences are regression estimates from weighted least squares models including indicator variables for quarter of admission and admitting hospital. Regression point estimates (95% CIs) of absolute rate reductions.

differences were smaller than 0.3 percentage points. We verified that the trends in mortality were similar in the two areas prior to the scheme's introduction. We also checked that our findings were unaffected when we controlled for changes in patient volumes and baseline mortality and when we compared the North West region with a subset of similar English regions.

Further examination of the additional mortality reductions in the North West region showed few differences by hospital type. Small hospitals, and hospitals rated as having excellent- or good-quality services by the Care Quality Commission prior to the scheme, showed the largest mortality reductions. North West region hospitals rated as having weak- or fair-quality services prior to the scheme did not reduce mortality more than similar hospitals in other regions.

Distributional consequences of Advancing Quality in the first 18 months

Financial incentives may increase the quality of care patients receive but may also lead to (inappropriate) exclusion of patients from the metrics used to measure performance. Little is known about the consequences of financial incentives for the distribution of quality. Providers may select 'easier' patients or be induced to treat 'harder' patients who they would have neglected without the financial incentive.

Using patient-level data ($n = 165{,}000$) for four of the five AQ conditions (excluding CABG because of small numbers) for the period October 2008 to June 2011, we examined (1) whether or not patients' probabilities of being excluded from, or achieving, performance metrics depended on their characteristics

and the characteristics of the hospitals that treat them; and (2) whether or not these probabilities have changed over time and if quality differences have narrowed or widened over time.

Patient characteristics included sex, age, ethnicity and area deprivation. Provider characteristics included foundation trust status and ratings of quality and financial management by the national regulators. We estimated multinomial logistic regression models across all indicators and health conditions for three periods: October 2008 to September 2009 (when the financial incentive was a pure tournament system); October 2009 to March 2010 (when a mixed financial incentive was offered); and April 2010 to June 2011 (when the CQUIN system of withholds for performance below a locally agreed threshold was in place).

Probabilities of exclusion from, or achieving, performance metrics fell by 5.7 percentage points over time and of achievement of performance metrics increased by 9.0 percentage points over time. We found a substantial age gradient in rates of both achievement and ineligibility, which persisted over time. Compared with patients aged under 45 years, patients aged 85 years and over were 13 percentage points more likely to be excluded from, and 11 percentage points less likely to achieve, the AQ quality indicators.

There was also evidence of significant differences in achievement by area deprivation, with the distribution of quality being pro-poor in the first period and pro-rich in the third period. Foundation trusts were more likely to exclude patients (by 0.7 percentage points), and trusts that were rated fair or poor by the national regulator for the quality of their services were less likely to exclude patients and more likely to achieve the quality indicators (by 2.7 percentage points). The improvements in quality recorded under the AQ programme therefore varied by patient and provider characteristics.

Cost-effectiveness analysis in first 18 months

Previous reviews have noted the surprising lack of evidence on the cost-effectiveness of P4P initiatives.[101] We updated the most recent systematic review and identified 14 studies that had reported on the costs of P4P schemes. Most did not include a comprehensive range of costs. Only two studies clearly incorporated the costs associated with the development and set-up of the P4P schemes, and only six studies included the ongoing running costs. The coverage of the consequences of these schemes was similarly narrow. Many studies did not consider consequences beyond the effects on the incentivised measures, few studies considered the impacts on health outcomes and only one study reported on the effects on non-incentivised patient groups and aspects of care.

We then proposed a more comprehensive framework, suitable for evaluation of health-care programmes such as P4P. We applied this framework to the first 18 months of AQ by obtaining data on the costs of the administration and management of the programme (as well as the incentive payments) and by applying the between-region difference-in-differences method (described in *Impact on mortality in the short term: methods*) to readmission rates and lengths of stay, as well as mortality. We assigned monetary values to these consequences using the National Institute for Health and Care Excellence cost-per-quality-adjusted life-year (QALY) threshold and prices from the national Payment by Results tariff.

The total cost of the AQ programme was just over £13M over the initial 18-month period, with only £5M of this consisting of the financial incentives. The ongoing running costs of the scheme actually exceed the bonus payments, making up the majority of the costs at just over £7M. This result reinforces the importance of considering the costs of P4P beyond the incentive payments themselves.

We estimated a gain of 6700 QALYs as a result of the reduction in mortality for the programme as a whole. At a QALY value of £20,000, this equals an estimated health gain worth £134M. Our estimates suggest that AQ also resulted in a reduction of 22,700 bed-days in the first 18 months. This is equivalent to a £5M reduction in costs. For readmissions, we estimate a statistically insignificant £0.5M increase in costs across all conditions.

Owing to the uncertainty around the methods used to estimate the potential QALYs gained as a result of the estimated reductions in mortality, we calculated the number of additional QALYs that it would be necessary to gain as a result of AQ in order for the programme to be deemed cost-effective at the standard UK cost-effectiveness threshold. We estimate that if just one QALY were gained for each death that was averted, AQ would be deemed cost-effective if a QALY is valued at £20,000.

Although it appears that AQ is likely to have represented a cost-effective use of resources during the first 18 months, an important consideration for policy-makers is its ability to continue generating improvements in quality and health outcomes and, therefore, justifying continued expenditure, in the longer run.

Impact of the change in incentive structure following the introduction of Commissioning for Quality and Innovation Payment Framework

We assessed the impact of AQ becoming part of a regional CQUIN scheme, which employed a fines structure with locally negotiated thresholds. Data for the two studied conditions, pneumonia and heart failure, were available at a quarterly level for the first 10 quarters of the AQ programme, spanning the period from 1 October 2008 to 31 March 2011. Twenty-four trusts collected information on the heart failure condition and 23 on pneumonia, as the Liverpool Heart and Chest Trust does not treat pneumonia patients. In the analysis we used pooled quarterly data on pneumonia and heart failure.

We first tested how performance differed under different incentive schemes. We defined the following regression specification:

$$y_{izmt} = \beta_0 + \beta_1 \text{ quarter}_t + \beta_2 D1 + \beta_3 D2_{imt} + \beta_4 D3 + X_{it}\gamma + Z_i\theta + c_{iz} + \varepsilon_{izmt}, \quad (1)$$

where y_{izmt} is the value of the performance measure for trust *i,* on quality indicator *z*, for condition *m*, in quarter *t*. *D*1 is a proxy for pure tournament scheme. It takes a value of 1 in year 1, one-third in year 2 and zero in year 3. *D*1 values reflect the contribution of the pure tournament scheme to the incentive design in that year: a pure tournament in year 1, a mixed scheme in year 2, in which the tournament forms one-third of the scheme, and a full fixed-threshold system in year 3. *D*3 is a proxy for improvement with a value of one-third in year 2 and zero in years 1 and 3 to capture the contribution of the improvement component to the design schemes in the different years. We defined:

$$D2_{imt} = \text{CQUINthreshold}_{imt} - \text{minCQUINthreshold}_m, \quad (2)$$

which measures the difference between the CQUIN threshold set for trust *i*, for condition *m* in year 3, and the minimum CQUIN threshold over all trusts for that condition. $D2_{imt}$ is set to zero for years 1 and 2. Hence, our base design is a regime where each trust faces a common threshold, which captures the minimum attainment award in year 2. X_{it} is a set of time-varying covariates which consists of the proportion of patients treated within 18 weeks and the number of beds available for day-only procedures. Z_i is a set of time-invariant covariates which consists of a measure of the size and type of trust, a measure of the quality of financial management, foundation trust status, and measures of commitment to core standards and national priorities published by the Care Quality Commission. c_{iz} is the unobserved trust–indicator effect for trust *i* and indicator *z*. ε_{izmt} are idiosyncratic errors.

β_2 measures whether or not the average outcome under a pure tournament-style incentive scheme is greater relative to a scheme in which trusts face a common minimum threshold. β_3 measures whether or not outcomes differ between an incentive regime consisting solely of a payment for performance improvement, as extrapolated from the small part it played in the mixed scheme in year 2, and that of the same common threshold scheme. β_4 measures the effect of a unit increase in the CQUIN threshold above

its minimum value. We also tested whether or not there was any difference in average outcomes under the pure-tournament and pure-improvement incentive schemes.

We defined y_{izmt} in four ways. First, we used the achievement score on each indicator, defined as follows:

$$P_{izmt} = \frac{T_{izmt}}{E_{izmt}} \times 100, \tag{3}$$

where T_{izmt} is the total amount of patients treated on indicator z in quarter t, and E_{izmt} is the number of patients eligible for treatment on indicator z in quarter t.

Second, we considered the quarterly improvement in achievement to test how the design structure affected the rate of performance improvement. This was defined as follows:

$$dP_{izmt} = P_{izm,t} - P_{izm,t-1}. \tag{4}$$

However, performance on these measures is likely to be affected by the 100% ceiling on the score. Therefore, we also defined y_{izmt} as both the level and change in the log-odds ratio:

$$\log odds_{izmt} = ln\left(\frac{P_{izmt}}{100 - P_{izmt}}\right), \tag{5}$$

which takes account of the ceiling at 100.

To test the predictions on how the incentive design impacted the distribution of performance, we estimated:

$$\begin{aligned} dP_{izmt} = {} & \beta_1(y1 \times yrstartq1)_{mt} + \beta_2(y1 \times yrsartq2)_{mt} + \beta_3(y1 \times yrstartq3)_{mt} + \beta_4(y1 \times yrstartq4)_{mt} \\ & + \beta_5(y2 \times yrstartq1)_{mt} + \beta_6(y2 \times yrstartq2)_{mt} + \beta_7(y2 \times yrstartq3)_{mt} + \beta_8(y2 \times yrstartq4)_{mt} \\ & + \beta_9(y3 \times yrstartq1)_{mt} + \beta_{10}(y3 \times yrstartq2)_{mt} + \beta_{11}(y3 \times yrstartq3)mt + \beta_{12}(y3 \times yrstartq4)mt \\ & + X_{it}\gamma + Z_i\theta + u_{izmt} \cdots (2), \end{aligned} \tag{6}$$

where *yrstartq**, (*) = 1, 2, 3, 4, is a dummy variable which equals 1 if a trust belongs to quartile (*) at the beginning of each AQ period, where quarter 1 is the baseline period for the first year of AQ, quarter 4 is the baseline for the second AQ period, and quarter 6 is the baseline for the third AQ period. Quartiles are created for each condition in each period using ranks of trusts based on the level of composite quality score (CQS) in the baseline quarter.

The coefficient estimate on each dummy captures the average quarterly change in CQS for trusts in a given quartile in a given period. To test whether or not each incentive scheme led to significantly different improvements for poor-performing trusts relative to good-performing trusts we tested the difference in mean quarterly improvement between trusts in quartile 4 relative to quartile 1. To test which year generated the greatest performance improvements for top-performing trusts we test the equality of β_1 and β_5, β_1 and β_9, and β_5 and β_9, and for initially poor-performing trusts, the equality of β_4 and β_8, β_4 and β_{12}, and β_8 and β_{12}.

In addition to this regression, we also estimate the model (1) using quantile regression. We assessed the impact of the incentive scheme design on poor performance (the 25th percentile) and on good performance (the 75th percentile).

We found that average performance under the 100% tournament system and 100% improvement system did not differ significantly from a regime based solely on a common minimum threshold. The results for performance at the 25th percentile were similar. We estimated that mean performance for the 75th

percentile under a pure improvement system was –4.8 percentage points lower than with a common fixed-threshold regime, although this disappeared after controlling for ceiling effects. We found that, irrespective of scheme design, P4P was associated with greater performance improvements for trusts performing poorly but we were unable to identify a design that led to higher improvement for the lowest performers.

Relationship between performance on process measures and observed outcomes

Following our findings on reduced mortality for pneumonia, we then examined whether or not this could be attributed to changes in the quality of care at the individual or organisational level. We obtained data from the AQ programme. These included hospital records from the SUS containing patient- and trust-level characteristics at spell level and data from the programme's QMR, which records delivery of the process measures of care for each patient.

The linked data contained 98,771 patient spells for patients admitted with pneumonia across 18 quarters, from October 2008 to April 2013. These encompassed almost the entire population of pneumonia patients qualifying for inclusion in the AQ programme from the 24 participating trusts. The population of AQ-qualifying pneumonia patients is 104,435 over the 18 quarters. However, there were 5664 patients for whom not all covariates were available and who were therefore dropped from the sample because of missing data. The patients who were removed were, on average, slightly older, by 2 years ($n = 73$), and, as a percentage, had a 1% higher in-hospital death rate of 25%.

We also obtained data from HES, which provided an in-hospital 30-day unadjusted and risk-adjusted mortality rate at trust level over 14 quarters of data, from October 2008 to April 2012. These data encompassed all patients with pneumonia from the 24 trusts.

At the patient level, we generated a dichotomous variable for within-spell in-hospital mortality as our outcome variable. This variable was generated from the discharge method in the hospital record. To extend our analysis we used two more outcome variables at the trust level: the unadjusted and risk-adjusted in-hospital 30-day mortality rates.

The five process measures of care for pneumonia under AQ were oxygenation assessment; initial antibiotic selection; blood cultures before antibiotics; initial antibiotics received within 6 hours of hospital arrival; and smoking cessation advice. For each process measure of care, we knew if the patient was given the measure, was not given the measure or was excluded from the measure. Exclusions are made when hospitals are able to remove patients from receiving process measures of care.

For our spell-level analyses, we generated two sets of dichotomous variables taking values of 1 if the process measures were given or if the patient was excluded from each measure. We performed both cross-sectional and panel data econometric techniques to extract both causal and correlational effects of process measures on health outcomes.

We found that 24% of patients who were admitted with pneumonia died in hospital. The average age of the admitted population was 73 years. The achievement rates were much lower than the scores used by the AQ programme, as we included patients who were excluded in the denominator of the achievement measure. Much higher proportions of patients were excluded from the process measures of care than received the care measures.

Oxygenation assessment had the highest level of achievement. Sixty per cent of patients were given this treatment. Initial antibiotic selection and antibiotics received in a timely fashion had achievement rates of around 30%. Blood cultures performed before initial antibiotic and smoking cessation advice were the

two quality measures with the highest exclusion rates of around 80%. Exclusion from smoking cessation advice was high, as non-smokers are excluded from this measure.

We found that 1 percentage point increases in the achievement rates on the blood cultures and smoking cessation advice measures were associated with reductions in the mortality rate of approximately 0.2 percentage points. These effects were statistically significant at 5%. One percentage point increases in the rates at which patients were excluded from the blood cultures and smoking cessation advice measures were also associated with a statistically significant reduction in the probability of patients being discharged dead. Timely delivery of antibiotics and antibiotic selection did not have statistically significant associations with mortality.

Some of the process quality measures were significantly associated with better health outcomes at trust level, but the magnitudes of the estimated coefficients were too large to represent clinically plausible direct consequences of these process measures. We concluded that the improvements in some process measures were reflecting wider improvements in the quality of care delivered by trusts.

Impact on mortality in the longer term

Towards the end of the project, we returned to our impact analysis on mortality with more recent data. We considered how AQ had affected mortality in the longer term, between months 19 and 42.

Impact on mortality in the longer term: methods

Hospital-level data on quarterly performance on the metrics incentivised by the programme were obtained from the AQ programme. Patient-level data on patient characteristics, coexisting conditions and mortality were obtained from national HES. As in our 18-month analysis, we used data for all patients in England treated for the three conditions covered by the programme for which patients are admitted in an emergency and from which there is a substantial mortality rate within 30 days, that is, AMI, heart failure and pneumonia.

We obtained equivalent data for patients admitted in an emergency with five of the six non-incentivised conditions that we had considered in our initial work. The sixth condition was excluded as it had been the subject of a national QI programme. We considered the possibility that the quality of care provided for patients admitted with the remaining five conditions may have benefited from the improvements in the care offered in emergency departments of the participating hospitals in the longer term.

Data were obtained for patients admitted between 1 April 2007 and 31 March 2012. We divided the data into three periods: (1) before – 18 months prior to the introduction of the scheme; (2) short term – the first 18 months of its operation; and (3) long term – months 19–42 of the programme. The final sample contained 390,652 patients admitted for AMI, 338,921 for heart failure, 761,954 for pneumonia and 333,991 patients admitted for non-incentivised conditions, treated at 161 hospitals across England. We analysed mortality occurring in any hospital within 30 days of admission.

As before, we calculated expected risks of mortality using logistic regressions at patient level which included sex and age; the primary ICD-10 diagnosis code; 31 Elixhauser comorbidities derived from secondary ICD-10 diagnosis codes; the type of admission (emergency or transfer from another hospital); and the location from which the patient was admitted (own home or institution).

The analysis of risk-adjusted mortality was undertaken on data aggregated by 3-month period and by admitting hospital. We tested whether or not the incentives had an impact on mortality using a between-region difference-in-differences analysis comparing changes in mortality over time in the hospitals in the North West region with those in the rest of England.

We tested whether or not the incentives had an impact on mortality in two ways: a between-region difference-in-differences analysis comparing changes in mortality over time between the North West region and the rest of England for the incentivised and non-incentivised conditions; and a triple-difference analysis comparing the changes in mortality over time between the incentivised conditions in the North West region and the rest of England with the changes in mortality over time between the North West region and the rest of England for the non-incentivised conditions.

We estimated the effects for all three incentivised conditions combined, all five non-incentivised conditions combined, and then separately for each condition. We weighted the condition-specific mortality rates using total volumes over the entire period to ensure that the combined mortality series did not reflect changes in the relative volumes of patients admitted for different conditions. All analyses allowed flexibly for time trends using an indicator for each of the 20 quarter-years, and for hospital differences using an indicator for each hospital in the separate condition analyses and an indicator for each hospital–condition combination in the combined analyses. We estimated separate impacts for the short- and long-term periods by including interaction terms between the intervention group and each of the two post-implementation periods.

Impact on mortality in the longer term: results

The average performance reported by the participating hospitals on all of the quality measures improved in the first 18 months and improved further in the following 24 months, particularly for heart failure and pneumonia (*Table 4*). Analysis of performance by quarter (see *Figure 1* and *Appendix 4*) shows that rates of improvement slowed over time and some quality measures, especially for AMI, plateaued at high levels of achievement towards the end of the period.

TABLE 4 Average hospital achievement on the incentivised indicators in the north-west of England in the short- and long-term periods

Condition	Indicator name	Short term	Long term	Difference
Pneumonia	Blood cultures performed in the emergency department prior to initial antibiotics received in hospital	63.1	81.9	18.7
	Adult smoking cessation advice/counselling	44.9	62.7	17.8
	Initial antibiotic received within 6 hours of hospital arrival	66.2	76.2	10.0
	Initial antibiotic selection in immunocompetent patients	82.7	91.2	8.5
	Oxygenation assessment	97.6	99.5	1.9
Heart failure	Discharge instructions	29.2	55.1	25.9
	Adult smoking cessation advice/counselling	55.6	75.7	20.1
	Angiotensin-converting enzyme inhibitor or angiotensin receptor blocker for left ventricular systolic dysfunction	90.1	94.9	4.8
	Left ventricular systolic function assessment	89.3	93.6	4.3
AMI	Adult smoking cessation advice/counselling	86.8	94.1	7.3
	Beta-blocker prescribed at discharge	93.8	97.6	3.7
	Aspirin at arrival	97.1	98.8	1.8
	Aspirin prescribed at discharge	98.3	99.3	1.1
	Fibrinolytic therapy received within 30 minutes of hospital arrival	84.4	85.3	0.9
	Angiotensin-converting enzyme inhibitor or angiotensin receptor blocker for left ventricular systolic dysfunction	97.5	98.0	0.5

The characteristics of the patient populations in the North West region and the rest of England were similar before the scheme's introduction, with a slight tendency for patients in the North West region to be younger and to have more coexisting conditions (*Table 5*). Similar changes over the short- and long-term periods in patient volumes and patient characteristics are observed in both areas.

Risk-adjusted mortality rates decreased over time in both the North West region and the rest of England for both incentivised and non-incentivised conditions (*Table 6*). The between-region difference-in-differences analyses show a significant impact of the programme on patient health in the short term (–0.9 percentage points, 95% CI –1.3 to –0.4 percentage points), comprising a statistically significant reduction in pneumonia (–1.5 percentage points, 95% CI –2.3 to –0.4 percentage points) and non-significant reductions in the other two conditions [AMI (–0.1 percentage points, 95% CI –0.9 to 0.6 percentage points) and heart failure (–0.2 percentage points, 95% CI –1.1 to 0.7 percentage points)].

TABLE 5 Characteristics of patients before and after introduction of P4P in short-term (18 months) and long-term (24 months) periods in the North West region (intervention region) and the rest of England (control region)

	North West region			Rest of England		
Patient data fields	Before introduction	Short term	Long term	Before introduction	Short term	Long term
AMI						
Admissions	19,992	18,804	23,282	104,460	101,765	122,349
Male patients (%)	61.7	61.9	60.3	63.3	63.2	62.7
Patients aged ≥ 75 years (%)	43.1	43.3	44.7	44	44.9	46.1
Coexisting conditions (average number)	1.6	1.7	2	1.5	1.7	1.9
Unadjusted mortality in 30 days (%)	11.6	10.5	9.9	10.4	10.1	9.5
Heart failure						
Admissions	15,295	15,493	20,127	82,847	86,786	118,373
Male patients (%)	53	53.1	52.8	51.3	51.6	52.1
Patients aged ≥ 75 years (%)	61.4	64	65.7	67.1	68.8	68.8
Coexisting conditions (average number)	2.3	2.4	2.7	2.2	2.4	2.7
Unadjusted mortality in 30 days (%)	16.4	15.3	14.2	15.3	14.8	13.7
Pneumonia						
Admissions	28,159	36,656	53,180	149,579	196,381	297,999
Male patients (%)	50.4	49.8	50.3	52.2	51	51.2
Patients aged ≥ 75 years (%)	53.9	55.6	55.1	56.4	58	58.1
Coexisting conditions (average number)	1.8	2	2.2	1.4	1.6	1.8
Unadjusted mortality in 30 days (%)	26.6	24.7	22.9	25.9	25.1	21.9
Non-incentivised conditions						
Admissions	13,449	14,837	21,975	76,649	84,578	122,503
Male patients (%)	57.4	57.1	55.9	57.2	57	56.7
Patients aged ≥ 75 years (%)	30.5	33	33	35.1	37.1	37.9
Coexisting conditions (average number)	1.6	1.7	1.9	1.4	1.6	1.8
Unadjusted mortality in 30 days (%)	13.9	14.1	11.8	12.2	11.8	10.9

TABLE 6 Risk-adjusted mortality for the conditions included in the P4P and those not included in the programme, before and after the introduction of the programme in the north-west of England

	North West region		Rest of England		Between-region difference in differences		Triple difference	
Mortality comparisons	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Non-incentivised conditions								
Mortality before introduction	14.0	–	12.3	–	–	–	–	–
Change from before to short term	–	–0.5	–	–1.2	0.7	–0.2 to 1.6	–	–
Mortality after introduction (short term)	13.5	–	11.2	–	–	–	–	–
Change from short term to long term	–	–2.9	–	–1.7	–1.2	–2.0 to –0.4	–	–
Mortality after introduction (long term)	10.5	–	9.5	–	–	–	–	–
Change from before to long term	–	–3.5	–	–2.8	–0.5	–1.4 to 0.3	–	–
Incentivised conditions combined								
Mortality before introduction	20.5	–	18.9	–	–	–	–	–
Change from before to short term	–	–1.6	–	–0.8	–0.9	–1.3 to –0.4	–1.5	–2.6 to –0.5
Mortality after introduction (short term)	18.9	–	18.1	–	–	–	–	–
Change from short term to long term	–	–1.8	–	–2.3	0.7	0.3 to 1.2	1.9	1.0 to 2.8
Mortality after introduction (long term)	17.1	–	15.8	–	–	–	–	–
Change from before to long term	–	–3.4	–	–3.1	–0.1	–0.6 to 0.3	0.4	–0.6 to 1.3
AMI								
Mortality before introduction	11.2	–	10.5	–	–	–	–	–
Change from before to short term	–	–1.1		–0.9	–0.1	–0.9 to 0.6	–0.8	–2.0 to 0.3
Mortality after introduction (short term)	10.0	–	9.6	–	–	–	–	–
Change from short term to long term	–	–1.0		–1.4	0.4	–0.3 to 1.0	1.6	0.5 to 2.6
Mortality after introduction (long term)	9.0	–	8.3	–	–	–	–	–
Change from before to long term	–	–2.2	–	–2.3	0.2	–0.5 to 0.9	0.7	–0.4 to 1.8

continued

TABLE 6 Risk-adjusted mortality for the conditions included in the P4P and those not included in the programme, before and after the introduction of the programme in the north-west of England (*continued*)

	North West region		Rest of England		Between-region difference in differences		Triple difference	
Mortality comparisons	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Heart failure								
Mortality before introduction	16.9	–	15.2	–	–	–	–	–
Change from before to short term	–	–1.1	–	–0.9	–0.2	–1.1 to 0.7	–0.9	–2.1 to 0.3
Mortality after introduction (short term)	15.8	–	14.3	–	–	–	–	–
Change from short term to long term	–	–1.6	–	–1.6	0.2	–0.6 to 1.0	1.4	0.2 to 2.5
Mortality after introduction (long term)	14.2	–	12.7	–	–	–	–	–
Change from before to long term	–	–2.7	–	–2.5	0	–0.9 to 0.8	0.5	–0.7 to 1.7
Pneumonia								
Mortality before introduction	26.9	–	24.8	–	–	–	–	–
Change from before to short term	–	–2.2	–	–0.7	–1.5	–2.3 to –0.7	–2.2	–3.4 to –1.0
Mortality after introduction (short term)	24.7	–	24.1	–	–	–	–	–
Change from short term to long term	–	–1.9	–	–3.2	1.1	0.4 to 1.8	2.3	1.3 to 3.4
Mortality after introduction (long term)	22.8	–	20.9	–	–	–	–	–
Change from before to long term	–	–4.1	–	–3.9	–0.4	–1.1 to 0.3	0.1	–1.0 to 1.2

Note
The short-term period covers the first 18 months of the programme. The long-term period includes months 19–42 of the programme. The between-region difference in differences are the changes over time in the North West region minus the changes over time in the rest of England. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors. The results are robust to other specifications of standard errors and weights.

The triple-difference analyses show an overall short-term effect of –1.5 percentage points (95% CI –2.6 to –0.5 percentage points), comprising a statistically significant reduction for pneumonia (–2.2 percentage points, 95% CI –3.4 to –1.0 percentage points) and non-significant reductions in the other two conditions.

Between the short- and the long-term periods, risk-adjusted mortality for the incentivised conditions fell by 2.3 percentage points in the rest of England and 1.8 percentage points in the North West region. This reduction in the rest of England is significantly larger (0.7 percentage points, 95% CI 0.3 to 1.2 percentage points) than in the North West region, and is again concentrated in pneumonia (1.1 percentage points, 95% CI 0.4 to 1.8 percentage points). However, the reductions in mortality were larger for the non-incentivised conditions in the North West region than in the rest of England between these periods. The triple-difference analysis shows a larger reduction in mortality (1.9 percentage points, 95% CI 1.0 to

2.8 percentage points) for the rest of England than in the North West region between the short-term and long-term periods.

We verified that the trends in mortality were similar in the two regions before the introduction of the programme (see *Appendix 4*) and confirmed that our findings were unaffected when we also included out-of-hospital deaths (which were less than 1 percentage point for all conditions) (see *Appendix 4*), and baseline mortality (see *Appendix 4*), excluded a small group of control hospitals for whom incentives for the same conditions were introduced in the long term (see *Appendix 4*), and when we examined 90-day rather than 30-day in-hospital mortality (see *Appendix 4*). We do not present these results in the main analysis, as the data were incomplete for the final 3 months of the study period.

In the long-term period of the study (19–42 months), two other regions outside the North West region (the South Central and South East Coast regions) adopted the AQ process measures which had already been used in North West region, although without the full supporting mechanisms in the AQ programme. For these two regions (which named their programmes 'Enhancing Quality' and 'Improving Quality'), in the long term, there was a financial incentive for performance on the AQ process measures, with money being withheld from hospitals in these regions if they failed to perform to negotiated standards on the AQ process measures.

To test whether or not the adoption of the AQ measures had an effect on mortality in these two additional regions, we conducted a between-region difference-in-differences analysis similar to our main analysis, but identifying the South Central region and South East Coast region as a separate group (labelled 'late-adopter regions').

Table 7 shows that in the first 18-month period (the short term), when AQ was introduced in the North West region but before the process measures were introduced in the late-adopter regions, the reductions in mortality were similar in the rest of England and the late-adopter regions. This was as expected given the lack of incentives in the late-adopter regions at that time.

In the longer term (when the AQ process measures were used in the late-adopter regions), reductions in 30-day in-hospital mortality were greater in the late-adopter regions than in the rest of England, but this difference was statistically significant only for AMI (–0.7 percentage points, 95% CI –1.3 to –0.1 percentage points). This suggested that the adoption of the AQ indicators in the two late-adopting regions was associated with a reduction on mortality attributable to AMI.

To investigate the possibility that the apparent loss of effect of AQ on the incentivised conditions in the north-west of England in the long term was a result of improvements in care in the non-incentivised conditions in the intervention hospitals, we first examined the between-region difference in differences separately for each of the non-incentivised conditions. This was similar to the separate analysis of the incentivised conditions but carried out for non-incentivised conditions.

TABLE 7 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme before and after the introduction of the programme in the north-west of England and the adaptation of the programme quality metrics in the late-adopter regions

Mortality comparisons	North West region		Late-adopter regions		Rest of England		Between-region difference in differences			
							North West region vs. rest of England		Late-adopter regions vs. rest of England	
	Rate	Change	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Non-incentivised conditions										
Mortality before introduction	14.0	–	11.8	–	12.5	–	–	–	–	–
Change from before to short term	–	–0.5	–	–0.9	–	–1.2	0.8	–0.1 to 1.7	0.3	–0.4 to 1.0
Mortality after introduction (short term)	13.5	–	10.9	–	11.3	–	–	–	–	–
Change from short term to long term	–	–2.9	–	–1.9	–	–1.6	–1.3	–2.0 to –0.5	–0.2	–0.8 to 0.4
Mortality after introduction (long term)	10.5	–	9	–	9.7	–	–	–	–	–
Change from before to long term	–	–3.5	–	–2.8	–	–2.8	–0.5	–1.3 to 0.4	0.1	–0.5 to 0.8
Incentivised conditions combined										
Mortality before introduction	20.5	–	18.6	–	19	–	–	–	–	–
Change from before to short term	–	–1.6		–0.8	–	–0.8	–0.9	–1.4 to –0.4	–0.1	–0.6 to 0.3
Mortality after introduction (short term)	18.9	–	17.8	–	18.2	–	–	–	–	–
Change from short term to long term	–	–1.8	–	–2.6	–	–2.2	0.7	0.2 to 1.1	–0.2	–0.6 to 0.2
Mortality after introduction (long term)	17.1	–	15.2	–	16	–	–	–	–	–
Change from before to long term	–	–3.4	–	–3.4	–	–3.0	–0.2	–0.7 to 0.2	–0.3	–0.8 to 0.1

TABLE 7 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme before and after the introduction of the programme in the north-west of England and the adaptation of the programme quality metrics in the late-adopter regions (*continued*)

Mortality comparisons	North West region		Late-adopter regions		Rest of England		Between-region difference in differences: North West region vs. rest of England		Between-region difference in differences: Late-adopter regions vs. rest of England	
	Rate	Change	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
AMI										
Mortality before introduction	11.2	–	10	–	10.7	–	–	–	–	–
Change from before to short term	–	–1.1	–	–0.6		–1.0	–0.1	–0.8 to 0.7	0.3	–0.4 to 0.9
Mortality after introduction (short term)	10.0	–	9.4	–	9.7	–	–	–	–	–
Change from short term to long term	–	–1.0	–	–1.9		–1.1	0.2	–0.5 to 0.8	–0.7	–1.3 to –0.1
Mortality after introduction (long term)	9.0	–	7.5	–	8.6	–	–	–	–	–
Change from before to long term	–	–2.2	–	–2.5		–2.1	0.1	–0.7 to 0.8	–0.5	–1.1 to 0.2
Heart failure										
Mortality before introduction	16.9	–	15	–	15.3	–	–	–	–	–
Change from before to short term	–	–1.1	–	–0.7		–1.1	–0.1	–1.0 to 0.8	0.3	–0.6 to 1.1
Mortality after introduction (short term)	15.8	–	14.3	–	14.2	–	–	–	–	–
Change from short term to long term	–	–1.6	–	–2.1	–	–1.3	0	–0.8 to 0.8	–0.7	–1.4 to 0.1
Mortality after introduction (long term)	14.2	–	12.2	–	12.9	–	–	–	–	–
Change from before to long term	–	–2.7	–	–2.8	–	–2.4	–0.1	–1.0 to 0.7	–0.4	–1.2 to 0.4

continued

TABLE 7 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme before and after the introduction of the programme in the north-west of England and the adaptation of the programme quality metrics in the late-adopter regions (*continued*)

Mortality comparisons	North West region		Late-adopter regions		Rest of England		Between-region difference in differences: North West region vs. rest of England		Between-region difference in differences: Late-adopter regions vs. rest of England	
	Rate	Change	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Pneumonia										
Mortality before introduction	26.9	–	24.6	–	24.9	–	–	–	–	–
Change from before to short term	–	–2.2	–	–1.1		–0.6	–1.7	–2.5 to –0.9	–0.5	–1.2 to 0.2
Mortality after introduction (short term)	24.7	–	23.5	–	24.3	–	–	–	–	–
Change from short term to long term	–	–1.9	–	–3.0	–	–3.2	1.2	0.5 to 1.9	0.2	–0.4 to 0.8
Mortality after introduction (long term)	22.8	–	20.5	–	21.1	–	–	–	–	–
Change from before to long term	–	–4.1	–	–4.1		–3.8	–0.5	–1.2 to 0.3	–0.2	–0.9 to 0.4

Note
Late-adopter regions are the two regions in England (the South Central region and the South East Coast region) that formally adopted the AQ metrics in months 19–42 of the programme running period in the North West region.
The short-term period covers the first 18 months of the programme. The long-term period includes months 19–42 of the programme. The between-region difference in differences are the changes over time in the North West region or South East Cost and South Central regions minus the changes over time in the rest of England. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

This analysis (reported in *Table 8*) shows that changes in mortality for non-incentivised conditions were similar in the North West region and the rest of England from before the programme to the short term. However, from the short term to the long term, reductions in mortality were greater for non-incentivised conditions in the North West region than in the rest of England, with statistically significant differences in acute renal failure (–2.7 percentage points, 95% CI –4.4 to –1.1 percentage points) and alcoholic liver disease (–2.0 percentage points, 95% CI –3.7 to –0.3 percentage points).

We then examined the extent to which patients with the non-incentivised conditions were treated in the same specialties as patients with the incentivised conditions in order to investigate the possibility of spillover of QI activities into the non-incentivised conditions in the intervention hospitals. We hypothesised that, if positive spillover was to occur, it would most likely affect conditions treated in the same specialties as those exposed to the QI measures in the AQ programme. We therefore examined the extent to which the non-incentivised conditions were treated in the same specialties and by the same specialists as the incentivised conditions. We did that by analysing two fields in the HES data that uniquely identify the specialist team responsible for each patient's care and under which specialty the lead specialist was employed.

TABLE 8 Risk-adjusted mortality for the non-incentivised conditions (not included in the programme), before and after the introduction of the P4P programme in the north-west of England

	North West region		Rest of England		Between-region difference in differences	
Mortality comparisons	Rate	Change	Rate	Change	Estimate	95% CI
Acute renal failure						
Mortality before introduction	20.1	–	17.3	–	–	–
Change from before to short term	–	–1.5	–	–1.6	0.2	–1.8 to 2.2
Mortality after introduction (short term)	18.6	–	15.7	–	–	–
Change from short term to long term	–	–5.4	–	–2.5	–2.7	–4.4 to –1.1
Mortality after introduction (long term)	13.2	–	13.2	–	–	–
Change from before to long term	–	–6.9	–	–4.1	–2.5	–4.4 to –0.7
Alcoholic liver disease						
Mortality before introduction	14.4	–	14.4	–	–	–
Change from before to short term	–	0.5	–	–0.9	1.7	–0.2 to 3.6
Mortality after introduction (short term)	14.9	–	13.5	–	–	–
Change from short term to long term	–	–3.2	–	–1.3	–2	–3.7 to –0.3
Mortality after introduction (long term)	11.7	–	12.2	–	–	–
Change from before to long term	–	–2.7	–	–2.2	–0.4	–2.2 to 1.5
Duodenal ulcer						
Mortality before introduction	7	–	7.5	–	–	–
Change from before to short term	–	–0.6	–	–1.6	1.1	–1.0 to 3.1
Mortality after introduction (short term)	6.4	–	5.9	–	–	–
Change from short term to long term	–	–1.3	–	–1.2	0	–1.9 to 1.9
Mortality after introduction (long term)	5.1	–	4.7	–	–	–
Change from before to long term	–	–1.9	–	–2.8	1.1	–0.8 to 2.9
Intracranial injury						
Mortality before introduction	14.2	–	12.7	–	–	–
Change from before to short term	–	–1.0	–	–1.4	0.7	–1.5 to 2.9
Mortality after introduction (short term)	13.2	–	11.3	–	–	–
Change from short term to long term	–	–1.9	–	–1.4	–0.2	–2.1 to 1.8
Mortality after introduction (long term)	11.3	–	9.9	–	–	–
Change from before to long term	–	–2.9	–	–2.8	0.6	–1.5 to 2.6

continued

TABLE 8 Risk-adjusted mortality for the non-incentivised conditions (not included in the programme), before and after the introduction of the P4P programme in the north-west of England (*continued*)

	North West region		Rest of England		Between-region difference in differences	
Mortality comparisons	Rate	Change	Rate	Change	Estimate	95% CI
Paralytic ileus and intestinal obstruction without hernia						
Mortality before introduction	9.4	–	7.8	–	–	–
Change from before to short term	–	–0.7	–	–0.6	–0.2	–1.6 to 1.3
Mortality after introduction (short term)	8.7	–	7.2	–	–	–
Change from short term to long term	–	–0.9	–	–1.2	0.5	–0.8 to 1.7
Mortality after introduction (long term)	7.8	–	6	–	–	–
Change from before to long term	–	–1.6	–	–1.8	0.3	–1.0 to 1.6

Note
The conditions are ordered from left to right in their proximity to the incentivised conditions. The short-term period covers the first 18 months of the programme. The long-term period includes months 19–42 of the programme. The between-region difference in differences are the changes over time in the North West region minus the changes over time in the rest of England. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

The results (*Tables 9* and *10*) show that patients with incentivised and some of the non-incentivised conditions were indeed being treated in the same specialties and by the same specialist teams. Specifically, *Table 9* shows that, at specialty level, 41–53% of patients with the incentivised conditions were treated in general medicine (approximately equivalent to general internal medicine in the USA). This is also the specialty in which 48% of patients with acute renal failure and 57% of patients with alcoholic liver disease were treated. These are the two conditions that showed the greater reductions in mortality from the short term to the long term in the North West region. There was less overlap between the specialties treating the other non-incentivised conditions, which were also those with non-significant reductions in mortality in the later period.

We also found overlap at the lower level of individual specialists' teams that treated patients with incentivised and non-incentivised conditions. *Table 10* shows that 55% of specialists treating at least one AMI patient had also treated at least one patient with acute renal failure. It can be seen that, although this level of overlap is not uncommon across all incentivised and non-incentivised conditions, the proportion of specialists with a substantial workload of both patients with incentivised and patients with non-incentivised conditions [defined here as treating at least 20% of patients with either condition (of the total of any two conditions)] remains high for the two non-incentivised conditions with statistically significantly larger mortality reductions in the North West region and is lower for patients with the other non-incentivised conditions.

In summary, these findings can be interpreted as modest evidence of a potential mechanism through which AQ could have affected the care of patients for some of the non-incentivised conditions (in particular acute renal failure and alcoholic liver disease). Further investigation into such potential positive spillovers is an important area for further research.

TABLE 9 Percentage of patients treated under each specialty

	Incentivised conditions			Non-incentivised conditions				
Specialties and incentivisation	AMI	Heart failure	Pneumonia	Acute renal failure	Alcoholic liver disease	Duodenal ulcer	Intracranial injury	Paralytic ileus and intestinal obstruction without hernia
Specialties in which incentivised conditions are concentrated								
General medicine	41	48	53	48	57	38	18	9
Geriatric medicine	9	16	17	15	6	6	6	3
Cardiology	35	16	3	2	2	1	1	0
Respiratory medicine	3	5	8	3	3	2	1	1
Specialties in which non-incentivised conditions are concentrated								
Gastroenterology	3	4	4	4	18	20	2	2
Accident and emergency	2	2	3	3	3	2	24	3
General surgery	0	1	1	4	3	24	11	74
Neurosurgery	0	0	0	0	0	0	18	0
Nephrology	1	1	1	8	1	1	0	0
Trauma and orthopaedics	0	0	0	0	0	0	9	0
Total	93	92	90	87	93	93	90	93

Figures are presented for specialties treating at least 5% of patients for one of the incentivised or non-incentivised conditions.

TABLE 10 Percentage of specialists treating at least one patient with an incentivised condition who also treated at least one patient with a non-incentivised condition

AQ conditions	Acute renal failure	Alcoholic liver disease	Duodenal ulcer	Intracranial injury	Paralytic ileus and intestinal obstruction without hernia
AMI, % [% treating at least 20% of patients from either condition (of the total of any two conditions)]	55 (38)	48 (25)	43 (17)	42 (16)	39 (12)
Heart failure, % [% treating at least 20% of patients from either condition (of the total of any two conditions)]	55 (32)	47 (19)	42 (13)	40 (10)	38 (8)
Pneumonia, % [% treating at least 20% of patients from either condition (of the total of any two conditions)]	53 (16)	38 (9)	40 (13)	38 (10)	42 (9)

Chapter 5 Qualitative methods and findings

Data collection

This report is based on qualitative data collected between April 2009 and March 2014. The first phase of the study included 300 interviews with staff members from participating trusts, 91 interviews with PCT and SHA executive officers (including AQuA staff) and over 50 observations of local AQ meetings, AQ Programme Leads meetings and SHA collaborative events and in-depth observations in trusts. We did not use a standard interview schedule but varied our questions according to the role and background of interviewee. In some cases, our questions were informed by observations undertaken prior to interview. The interviews were focused on understanding what was happening in relation to AQ, as well as the interviewee's role and background. We also asked about how people had come to be involved in AQ.

In terms of the types of people we interviewed, not all of them slot neatly into categories. For example, many AQ leads had a clinical background and some continued clinical duties alongside their role. We classified people who were AQ leads as leads, rather than clinicians. Similarly, in some cases staff who undertook data quality work were part of QI teams in some hospitals, whereas they were in clinical audit departments in others. Additionally, these staff may or may not have been involved in data analysis depending on the local context. We classified all of these staff as being data/audit personnel. For people at director or deputy director level, we classified these as senior managers and other managers as operational managers. According to this classification of the hospital staff, 89 of the interviews were with nurses, 53 were with data audit staff, 52 were with AQ leads, 44 were with operational managers, 31 with doctors, 26 with senior managers and five with other clinicians.

The initial intention of the qualitative component was to focus over time on a subset of sites, after initial qualitative data collection in all sites. However, owing to the difficulty in identifying consistently high and low performers (see *Appendices 2* and *3*), combined with the fact that systems and processes were continuously evolving, we continued to collect data across all provider sites.

For the second phase (after the first 36 months) we planned to conduct interviews with all AQ programme leads and a staff member from each CCG.

The introduction of CCGs made it difficult to identify relevant personnel in CCGs who could comment on AQ. This meant that only 11 CCG staff were interviewed in the final phase of the evaluation. In addition to interviews with AQ leads, we also spoke to staff participating in the AQ programme more generally and observed AQ meetings ($n = 10$). In total, therefore, we conducted 126 'interviews' in the second phase. Only 20 of these were full interviews. The nine interviews with hospital staff were conducted with five data/audit staff, three leads and one nurse.

The remainder were relatively short, involving conversations at AQ events rather than lengthy semistructured interviews of the sort that had been undertaken in the first phase. Of these informal interviews, 33 were conducted with leads, 33 with audit/data staff, 25 with operational managers, 10 with nurses and five with doctors. This informal approach was partly because saturation had been reached in the earlier phase with regard to many issues and these conversations confirmed earlier findings. Additionally, as agreed with the funder, greater emphasis was placed on quantitative analysis in the second phase, which meant that the funding that had been used in phase 1 for a dedicated qualitative researcher was used to support quantitative analysis in phase 2.

Interviews were digitally recorded and transcribed verbatim. In addition to mapping the journeys of participating providers, in order to provide a coherent structure for making sense of the data and to identify common themes and differences, we used the framework developed by Paul Bate and colleagues in their study of QI journeys in health-care organisations.[84] This framework identifies six common challenges facing organisations, which are as follows:

- 'Structural (organising, planning and co-ordinating quality efforts)
- Political (addressing and dealing with the politics of change surrounding any QI effort)
- Cultural (giving quality a shared, collective meaning, value and significance within the organisation)
- Educational (creating a learning process that supports improvement)
- Emotional (engaging and mobilising people by linking QI efforts to inner sentiments and deeper commitments and beliefs)
- Physical and technological (the design of technological infrastructure that supports and sustains quality efforts)'.[84]

We used these as part of a framework to analyse our data. Additionally, we had originally intended to use realist evaluation[102] to identify context–mechanism–outcome configurations to explain our findings (i.e. this mechanism, in this context, produced this outcome). As we progressed with qualitative data collection, we developed an initial understanding and related programme theory to explain how the programme was intended to work in practice. This enabled us to start to identify mechanisms that were intended to contribute to the achievement of AQ outcomes.

However, when attempting to use this approach to explain how AQ operated in practice, we had difficulty in distinguishing between mechanisms and outcomes in some cases. For example, various aspects of the AQ programme contributed to its gaining legitimacy. This legitimacy might be viewed as an outcome of the AQ programme, but it also contributed to universal participation, which in turn could be viewed as a mechanism that strengthened legitimacy. In addition to blurring the distinction between mechanisms and outcomes, the relationship between factors was dynamic rather than a linear causal chain with context/mechanism configurations generating outcomes.

We retained the aim of elucidating how changes occurred but for phase 2 of the evaluation we revisited the data collected in phase 1. This led us to draw on the Beckert[94] framework outlined in our literature review, which reflects the dynamic nature of change and facilitates consideration of the broader field in which change initiatives are situated.

Developing a programme theory

We outline below the programme theory developed before comparing and contrasting what happened in practice. Throughout this description we provide more detail in relation to the challenges presented by AQ implementation as categorised by Bate *et al.*[84] We also draw a more general interpretation, using the Beckert[94] framework outlined in our literature review, of how and why the programme produced the impacts we identified in *Chapter 5*.

There was never an explicit programme theory expressed in terms of AQ, but it is possible to identify implicit programme theory from the data collected over the course of this research. Having decided that a HQID-style approach met their needs, AQ programme leaders began making plans to put this into practice in the region. There was no protocol outlining in detail how the overall programme was intended to work at the outset. However, various discussions, documents and pronouncements contained (albeit often implicit) theory about the ways in which AQ was intended to work.

Given that the whole process was one of an emerging and evolving programme theory, the boundaries between initial and subsequent theory are sometimes blurred. At the same time, we found it helpful to try to distinguish between the initial theory and what happened subsequently, in order to elucidate the how and why of the programme and to highlight important lessons for programme roll-out and learning more generally. The distinction drawn between initial and subsequent programme theory may therefore at times be somewhat artificial. Additionally, the compartmentalisation of theory components in the way we present this may suggest a coherence which belies the fuzzy nature of this process and ignores the dynamic relationship between programme aspects. However, we articulate this aspect of the process using an alternative framework, drawing on Beckert,[94] as outlined in our literature review.

The first three elements of programme theory relate to the establishment and infrastructure for AQ and these relate to the 'structural' challenges identified by Bate *et al.*[84]

Adapting a tried-and-tested initiative

Looking around for targeted ways of spending growth money, which would enable them to measure impact, the programme leaders came across the HQID programme. This was chosen as the preferred solution after the programme leaders had investigated a range of such initiatives. It was hypothesised that opting for an existing programme would be preferable to developing a local solution from scratch, but this theory assumed that adapting this to local contexts was important in order to increase the chances of success.

Piloting

The intention was to use piloting to test feasibility and identify any factors likely to hinder implementation to increase the likelihood of success. First-wave organisations were to receive twice as much in set-up costs as second-wave trusts (£60,000 vs. £30,000). It was envisaged that second-wave hospitals would experience fewer teething troubles, hence the reduced allocation. It is not clear whether or not the theory at this stage assumed that no additional resources would be required beyond this or whether or not there existed sufficient 'organisational slack'[85,103] to enable participating organisations to implement AQ.

Investing in dedicated support and infrastructure

Programme leaders believed that a supporting infrastructure would be required, possibly as a result of their extensive interaction with Premier, which had provided similar support to the US initiatives. Programme leaders hypothesised that dedicated support and infrastructure from people with direct experience of QI initiatives from outside the NHS were needed and undertook a tendering process to appoint a partner organisation.

They also established an AQ team comprising experienced NHS managers, based in the SHA and committed to developing a governance structure to underpin AQ. The theory here seems to have been that it was important to create places and spaces to conduct work and dialogue about AQ issues outside the day-to-day workings of the providers participating in AQ. In addition, the pledge to ensure transparency, making the workings of these various groups publicly available, appeared to be aimed at enhancing legitimacy and increasing participant commitment. It seems clear that the programme leaders reasoned that merely putting incentives out there and expecting people to respond was unlikely to achieve the desired effect.

However, this dedicated support was soon also required at a local level, with all organisations eventually investing some of their AQ resources (in some cases, the set-up funds) in dedicated local staff to support AQ within their own organisations. There was little initial explicit programme theory supporting these type of appointments, as evidenced by the staged approach and variation in the nature of the roles across the organisations.

Using data to get attention

The programme leaders hypothesised that it was necessary to use quantitative performance data (as opposed to anecdote or gut feeling) to get attention from providers about current performance, prior to the commencement of AQ. In the absence of a focusing event[104] such as a crisis, the intention was to highlight the gap between perceived and actual service delivery as a means of alerting providers and commissioners to the problem of suboptimal service delivery. This was a means of addressing the political challenge for which it was felt that quantitative data were less likely to be debated politically than other types of data or anecdote.

Using voluntarism and peer pressure to drive participation

The political challenges associated with gaining commitment to change were also clear in the implicit theory, where it was hypothesised that a voluntary approach would result in higher levels of commitment compared with mandatory participation. However, it was hoped that peer pressure would act to encourage organisations to join the programme, leading to participation of all 24 eligible organisations.

Pragmatically using evidence to get buy-in, programme leaders drew on peer-reviewed studies of the HQID and also used Premier's own publicity material (which had not been subject to peer review) to publicise the benefits of the programme and get participating organisations and key individuals on board.

Comparing apples with apples

The emphasis on data continued into the programme itself, in which leaders emphasised the importance of providing clear data definitions and compliance with these definitions to ensure comparisons were on a like-with-like basis. This was intended to enable valid benchmarking but also to ensure commitment of staff to what they saw as a fair and legitimate process. In addition to a data dictionary, clear written guidance and training sessions delivered and supported by Premier, the programme also incorporated an assurance process which subjected participants' data to scrutiny by the Audit Commission, a national body external to the NHS North West. This also contributed towards addressing the cultural challenge of giving a shared meaning to quality throughout the organisations involved.

Providing strategic leadership by sustaining senior-level commitment

Sign-up to participate in AQ involved Chief Executive commitment to the process, as this was hypothesised as being likely to increase success through demonstration of the importance of quality within organisations – one of the cultural challenges. In a context in which well-intentioned initiatives quickly become yesterday's news, this senior leadership was seen by programme leaders as important in sustaining participation.

Institutionalising behaviour change

For chief executives of AQ organisations, the implicit hypothesis appeared to be that, over a relatively short period of time, AQ would become embedded as part of routine practice, and dedicated AQ staff within the hospitals could be moved on to pastures new. The idea of replacing one set of practices and beliefs with another and for these to become taken for granted, institutionalised, was central to their theory and appeared to be addressing some of the cultural challenges.

Using feedback to inform learning

The theory here was that feeding back data to participants would enable them to act on that feedback by identifying poorly performing areas, investigate reasons and trial solutions. Feeding back comparative data was intended to encourage participants to improve care, but there were no correct solutions to accompany these data. Instead the theory seems to have been that data would encourage learning within organisations, as well as learning from good practice elsewhere, thus addressing one element of the educational challenge identified in QI.

Using competition to drive performance

From the outset, the programme leaders pledged to report publicly the results of the programme. This reflects a commitment to transparency generally, as part of a process of securing legitimacy, but it was also intended to act as a spur to improvement. In terms of other mechanisms, in addition to using annual public reporting, the intention was to feed comparative data back to all participating hospitals on a regular basis, tapping into the competitive spirit of participants and underpinned by the tournament-style AQ design. A combination of normative and coercive pressure embodied in the quarterly and annual performance ranking reports was hypothesised to prompt action in a context in which poor performance on AQ measures would be collectively frowned upon. This approach to addressing some of the emotional challenges identified was particularly important for clinical professionals for whom peer comparison is a fundamental part of their own professional development.

Using collaboration to drive improvement

Programme leaders were, however, anxious to avoid competition driving out collaboration. Alongside the competitive elements of AQ, it was also seen as important to encourage the development of an AQ community, which was hypothesised to facilitate collaborative learning and maintain improvement momentum. Implicitly, they were attempting to develop a movement for improvement, thus addressing another emotional challenge.

Using money to spur improvement

Underpinning all financial incentive initiatives is a relatively simple theory of cause and effect. Linking money to desired outcomes is hypothesised to improve performance compared with the absence of such monetary rewards. Whether this is viewed as addressing a political or an emotional challenge will depend on the context and the individuals and professions involved, but it was clear that programme leaders believed that the financial incentives were important.

In year 1, AQ rewarded trusts performing in the top two quartiles of all participating organisations in the North West region. In subsequent years, the intention was to reward improvement, as well as absolute levels of performance. There were no financial penalties for trusts whose performance was below average. The design that rewarded relative performance in year 1 and subsequently improvement too, was seen as fair and was hypothesised to get greater buy-in relative to other potential designs (e.g. low performers lose money, all participants win money). AQ leaders also stressed that rewards would flow to clinical teams to reinvest in services, in an effort to focus money and attention where action was needed most, as this was hypothesised to act as a motivator for the staff working in those areas and might be viewed as addressing an emotional challenge by obtaining frontline buy-in.

Facilitating standardisation

Promotion of AQ measures was intended to standardise processes of care, with the intention of improving outcomes. According to institutional theory, the pursuit of legitimacy leads to organisations yielding to pressures within the field. The various components of the AQ programme (e.g. standard metrics, data feedback, building a community and so on) were hypothesised to lead to compliance with the policy and convergence within the organisational field.[105] The latter refers to the way organisations within the field increasingly come to resemble one another as part of the process of securing legitimacy. Pressures towards this institutional isomorphism can be mimetic (organisations copy other organisations), coercive (responses to external pressures such as governmental action) and normative (reflecting accepted norms of the professional community). The emphasis on standardisation within the AQ programme theory appeared to involve a combination of all three of these types of pressures.

Putting the theory into practice

Adapting a tried-and-tested initiative

The programme was adapted to suit the local context throughout the first 3 years.

Although the AQ measures used were based on relatively good evidence of efficacy,[106] their application was also intended to take account of local contexts and national policies. For example, the HQID scheme measure for pneumonia patients used antibiotics within 4 hours of arrival. But, in recognition of a national NHS 4-hour wait target for accident and emergency department patients, AQ used 6 hours. Although admitting patients a few minutes before they breached the 4-hour ED target was not ideal, it was common practice at the time and, rather than trying to change this practice, the leaders reasoned that 6 hours would be long enough to enable hospital staff to cope with a last-minute admission to meet the 4-hour target and give 2 hours thereafter to administer antibiotics. Taking account of other constraints in the system to produce measures that would be regarded as reasonable and feasible was seen as important in securing initial commitment and sustained participation therefore.

Early on in the AQ process, concerns were raised about the inclusion of beta-blocker at arrival for AMI patients, which led to this being dropped from the measures from July 2009 (i.e. 9 months after the start of the programme). Similarly, the requirement to document a CURB-65 score (a clinical indicator used to assess severity of pneumonia) was introduced with effect from April 2011 with agreement from the AQ community. This was intended to help participating organisations systematise care and identify patients early on in their journey, to facilitate prompt delivery of care in accordance with AQ metrics.

It is important that adaptation should not be viewed solely as a technical process of refining measures. A key aspect of this evolution and development involved listening to AQ participants and responding to their concerns. Changes to AQ were made on the basis of dialogue with participants. Reflecting the pragmatic approach to implementation, the requirement to take blood cultures prior to starting antibiotics was dropped with effect from November 2012. Many providers had experienced difficulties with this measure, in part because of policies to reduce taking blood cultures generally, which were designed to reduce hospital-acquired infection and, more recently, to cut costs. The need to listen resulted in what at times was a protracted process, but this was important in securing agreement for any changes made. Once agreement was reached, however, participants had to abide by decisions taken, albeit with the proviso that as with all aspects of the programme, the potential and mechanisms existed to review this at a future date if concerns were raised.

Objections about the requirement to provide smoking cessation advice to patients who had not smoked for many months (the target requires this for anybody who has smoked within the last 12 months) were a recurring theme, with this issue creating tensions among staff, many of whom regarded counselling non-smokers as a waste of time. However, this measure was not altered.

With regard to CABG, there were indications that insufficient adaptation had taken place in transferring HQID to AQ. Providers had little difficulty with compliance, scoring highly from the first quarter onwards. Given that there were only four providers within the region that performed this procedure, distinguishing between providers and allocating rewards on the basis of tiny differences in performance was perceived as problematic by some participants (see Bhattacharyya *et al.*[47] for similar issues with hip and knee measures in HQID). Furthermore, clinicians viewed the CABG measures as insufficiently challenging, particularly in a context in which collecting and reporting data on cardiac surgery performance on a range of indicators was introduced long before the AQ programme arrived in England. In some cases this resulted in clinicians feeling distanced from the programme, as well as rewards being allocated for relatively little additional effort.

Over the final 2 years the team supporting AQ extended its application to other clinical areas – starting with stroke, mental health and others. This is not a part of this evaluation but showed the enthusiasm for this type of approach (identifying common measures, reporting on them, comparing data, providing financial reward, etc.) based on the perceived success of the original five conditions.

Piloting

Seven trusts participated in a first-wave pilot. Site selection was informed by geographical location (bacause the intention was to obtain geographical spread), willingness to participate, readiness of systems, good evidence of partnership between primary and secondary care and organisation type. Among pilot sites, participation in a dry run AQ exercise was perceived as a shock to the system. Various problems were identified as part of this process. Rather than expressing their gratitude for the piloting opportunity, participants were anxious and risked being overwhelmed by the enormity of the task they had taken on. They were thankful for the additional resources, but at this stage many had chosen to keep those in reserve until they had worked out where best to invest them. The fact that these resources were non-recurrent was also a source of worry, because in most cases these were viewed as likely to be required on a recurring basis. Within participating organisations, there was very little in the way of organisational slack to allow organisations to adapt successfully to pressures for adjustment created by AQ and to initiate required changes in a sustainable manner.

Piloting also had a somewhat negative impact upon those not involved in the pilot. In a context in which rewards were based on competition and relative performance, these concerns were exacerbated. From their perspective, providers that has not participated in the pilot would receive less money and start the race from a position far behind the sites that had, which seemed unfair to both. Furthermore, the fact that not all organisations that volunteered to participate in the pilot were chosen was not well received by some participants.

The pilots revealed a number of challenges which persisted beyond the pilot and remain in many sites at the time of writing. These chime with the challenges identified by Bate *et al.*[84] and we discuss these in more detail.

Investing in dedicated support and infrastructure

Central support

A governance structure was established for the AQ process across the north-west of England and, as part of this process, formal channels for feeding back concerns and issues were established. In general, comments suggested that these channels worked well, although the time pressures on staff in AQ providers meant that participation in processes outside their trust sometimes suffered.

The support from a central NHS AQ team, together with Premier, involved provision of technical advice, data systems and training. In addition, in a context where AQ programme leads and lead clinicians were often under great pressure to improve performance and in danger of being a voice in the wilderness in their own organisation, the support provided appeared to have an important role in helping staff cope with pressure. The face-to-face contact in the early stages of AQ with Premier trainers and data personnel, who had been flown over to England to meet NHS staff, contributed to the development of personal relationships based on trust and respect. These continued when most of the Premier employees returned to the USA to provide support at a distance.

The fact that Premier employees had experience of the HQID programme was also helpful, as AQ participants saw them as seasoned and wise campaigners who could sympathise in times of trouble and offer helpful advice. When most of the Premier employees returned to the USA to provide telephone support to AQ participants, the service was delivered in accordance with UK working hours, which meant that US personnel were working unsocial hours. It also meant that AQ participants had a very responsive service and one that was based primarily on personal relationships. Providers had specific Premier staff who

covered their organisation, as opposed to these being a general pool across all 24 organisations. Participants were very appreciative of the Premier service and the fact that AQ provided this had the added benefit of making them feel valued. This structural support enabled a significant emotional challenge to be addressed for those who had specific responsibility for AQ in their organisations and were therefore able to access this support.

The SHA AQ team were also viewed as friends, and various participants contrasted this invaluable support with their previous perceptions of the SHA as being nameless bureaucrats.

> *. . . somewhere in the ether . . . I'd be depressed, you know. But they are absolutely lovely . . . they do inspire you, and they do give you confidence.*
>
> *AQ lead, ID7, T23*

> *[names two members of the SHA AQ team] I email them or I ring them up, and I think they've been brilliant, I really do. Historically, you always felt like the SHA were sort of somewhere in the ether. But, they're so hands-on and so practical. Not just the AQ leads meetings, but the subcommittees that I've been to were really, really helpful. So, yeah, I think they've been invaluable.*
>
> *AQ lead, ID18, T22*

Furthermore, in the context of English cultural norms that may tend to focus on shortcomings and problems, the positive language of the team from the USA also appeared to help. When scrutinising monthly performance reports, participants were encouraged to use the phrase 'opportunities' (as in opportunities to do the right thing) instead of 'failures' and 'challenges' when referring to patients identified as eligible for AQ but missing some or all of the AQ measures.

> *. . . when you look at patients where they haven't met all the measures, where you're starting to look at your opportunities . . .*
>
> *AQ lead, ID13, T13*

> *It's a marathon not a sprint, as my friend, Christy, from the States continues to tell me.*
>
> *AQ lead, ID9, T7*

Participants both recognised and referred to the implications (a typically American positive spin on negative news) but, at the same time, accepted the phrase, initially using it in some cases in an ironic and knowingly self-conscious way. Over time (and immediately in some cases) it became part of the accepted language of AQ participants, which also helped contribute to an emerging lexicon shared by and specific to the AQ community. This shared language emerged from and added to the development of an AQ organisational identification, which provided support and was also part of the process of socialising new entrants[107,108] to the AQ community, again addressing an 'emotional' challenge.

Support from staff at Premier and the SHA AQ team was also seen as helpful in bringing down stress levels in ways that were much more about supportive relationships with individuals and personal contact than formal structures to facilitate communication between providers and the SHA and Premier AQ teams (as the word 'friend', used by some participants when referring to Premier staff, indicates).

> *we thought that was really good, we'd worked really hard . . . And then to see that [big mismatch between the number of pneumonia measures for whom AQ data had been collected and the number of eligible patients identified by Premier] it's quite devastating. But, I mean, speaking with Christy, who's from Premier, we had some training with her last week and we were talking about December and she said 'Everybody in the region, everybody's data went off the high end' It's obviously that time of year isn't it?*
>
> *AQ specialist nurse, ID8, T23*

Technical support and infrastructure

The challenge of 'the design and use of a physical, informational and technological infrastructure that improves service quality and the experience of care'[84] also required dedicated resources at a local level, although this was not always recognised from the start.

For AQ participants, data collection was a burdensome process owing to excessive reliance on paper-based records and the location of these records in different departments. Often, data collection involved a significant amount of trawling, as AQ patients could be located in a range of different wards across the hospital.

Even when electronic data-collection systems were used, they were often unable to provide solutions to the data requirements of AQ and meant that much time and effort was spent on doing retrospective trawls of case notes every month.

> *What we need to do is get away from this mammoth case note audit . . . We can't carry on forever, pulling case notes permanently. So, that's something that we are addressing, we need to work around.*
>
> *AQ lead, ID28, T17*

One provider with an electronic workable solution in the early stages was highly unusual, although several providers had moved to electronic data collection by the end of the evaluation.

> *we've got handheld devices . . . that's why it's worked with us, especially with nurses. They will have that on their walk around the emergency floor. Follow the patient and start inputting that way . . . And that's why we've not had a resource to sit down and put anything in retrospectively . . . That's what happens in reality sometimes, but that's not the system.*
>
> *AQ lead, ID36, T24*

> *we don't have the ability to embed within those systems currently. The data captures mechanism easily for stuff that isn't . . . for data items that aren't already defined.*
>
> *AQ executive lead, ID35, T3*

The departure of Premier from AQ and the introduction of AQuA was also a cause for concern regarding available technical expertise connected with data input. The many positive comments about Premier also highlight how important the nature of the service is, as opposed to merely the technological interface, to AQ staff.

Advancing Quality leads

Although identifying an AQ programme lead was a standard approach, these individuals varied enormously in terms of their background experience and disciplinary training, as well as their seniority and departmental location within the trust. Although some leads were a dedicated full-time AQ resource ($n = 3$), most were attempting to combine their AQ duties with their existing role. Some were seconded on a temporary basis and were uncertain about their future, and, as a result of staff turnover, others had recently come into post following the departure of the previous incumbent.

A considerable part of their duties involved liaising with people across the various components of AQ-related work and ensuring routine recording and coding tasks were complete and that data submission deadlines were met.

> *Advancing Quality it is done by our divisions. Everything is given to the divisions to say you need to sort this out and what I do is I overall for each of the divisions and the trust as a whole . . . send out performance reports, tell them what the targets are, if they are up to date. Give them ideas where*

> *they need to work on but generally the divisions are the ones who say well this is what we need to put in place.*
>
> *AQ lead, ID109, T2*

> *They gave us money to help appoint the clinical project manager, who's [name], who you'll meet at some point. Who really does a lot of the donkey work and she runs around. She actually does most of the work and I do nothing but just sit around talking to her.*
>
> *AQ clinical information officer, ID123, T14*

The amount of time available to leads varied, and most leads were attempting to combine their AQ duties with other roles. AQ leads (alongside auditing and coding staff) appeared to bear a large part of the additional workload involved in the everyday implementation and sustainability of AQ. Overall, this appeared to be a very demanding role and some turnover was observed during the period covered by this report. In some cases, this was because these posts were initially filled on a 1-year secondment basis rather than a more permanent arrangement. However, given the knowledge accumulated over the first year of AQ, it would seem to be important to ensure continuity in the role wherever possible.

The perceived effectiveness of AQ implementation in provider organisations was very closely linked to an AQ lead's ability to act as a link between the executive and clinical levels as well as a focal point for frontline staff. AQ leads frequently found themselves under considerable pressure to explain the reasons behind a trust's weaknesses and to initiate interventions in those areas. In-house structures were established in each provider to facilitate regular meetings with clinical, administrative and executive staff involved in AQ work. In these meetings, the topics of discussion often revolved around performance updates, challenging areas and suggestions for future work. However, progress often required action by committed and diligent individuals. Furthermore, these AQ groups often stood outside any organisation-wide infrastructure, and awareness of AQ among staff across the organisation was often limited.

Local support

The impact of competitive pressure and the timescales for reporting led, in many sites, to investment in staff, particularly AQ specialist nurses, focused on data collection. Some providers had taken a strategic decision, in the early stages of AQ to invest in additional staff at levels above and beyond the initial start-up costs, whereas others worked within their allocated start-up cost budgets. Among some providers, the appointment of specialist nurses for a fixed term had been actively resisted, partly because this was seen as unlikely to contribute to sustainable or system-wide solutions and also partly because using specialist nurses to collect data was viewed as an inappropriate use of resources. For others, actions were constrained by available resources and specialist nurses employed to improve AQ performance were seen as unaffordable.

Pneumonia and heart failure were the areas for which most problems arose, and in one trust a decision was taken to employ specialist nurses to work on these areas. When, at a collaborative learning event, participants heard how the input of specialist nurses appeared to be improving care (and performance against related metrics), other providers took an interest and many began to copy this strategy. Over time, most providers invested additional resources in AQ, particularly in relation to recruitment of specialist nurses, because existing resources proved inadequate to meet the AQ challenge. In many cases, these posts were funded for a time-limited period, but this was subsequently extended as it became clear that institutionalising behaviour change was proving very difficult.

Advancing Quality leads were quick to grasp that early identification of patients who were potentially eligible for AQ-related care was crucial to meeting and improving care, especially in the context of a 6-hour target for antibiotics for pneumonia patients. This led to various initiatives at sites to identify patients, but having identified them it was necessary to ensure that their diagnosis was confirmed and treatment given or that they were treated as if they required care which was in line with AQ measures. As described previously, this led to recruitment of specialist nurses at one site to join up gaps in care and

maintain a sustained focus on patients as they moved through the hospital, which subsequently was adopted by an increasing number of participating sites as the programme evolved.

At this stage, the competitive element of AQ was implicitly prioritised over longer-term sustainability, despite the initial notion from the programme designers that new ways of working would become institutionalised and that this was central to the theory underpinning AQ.

In some cases AQ was explicitly part of ongoing QI processes, with support structures in place which were used for previous and current QI programmes. In addition to any trust-wide general QI arrangements, there were specific AQ teams, which involved the AQ lead. Membership and ways of working evolved over time. In most cases, team members had AQ duties added to their existing workload.

> *Although people say, 'Oh, well, yeah. Quality has to be owned by the divisions. In 12 months' time, when it's all up and running, it will just be integral to what we do'. The reality is you need somebody there to be in long enough for support.*
>
> *AQ lead, ID12, T11*

> *When we were first put into post, it was as a commitment to put in sustainable processes so that we [extra staff to facilitate AQ initial implementation] would withdraw after 12 months. But . . . everyone has recognised it's not as easy as that, and that there's a lot more involved in it than I think maybe what was anticipated in the early stages.*
>
> *AQ lead, ID7, T23*

As the scale of the task became apparent, concerns were raised about the ability to sustain AQ without ongoing dedicated support. The increasing realisation that a complete devolution of responsibility to staff without a dedicated responsibility for AQ delivery would be very challenging, made the local provider AQ teams invaluable in providing assistance and support in the various organisational levels where AQ work tasks were carried out. Constant vigilance was required to ensure that AQ requirements were adhered to, with data recording continuing to present problems throughout our study.

> *It's the workload. They've all been given so much to do, all of them. This will just get lost in the familiar, of all the stuff. Which is the reason why if you want to do something properly you have to target or label someone whose job it is to do that thing. Which is why it's working at the moment. So, what's happened is, because [specialist nurses have] been here for 12 months, we just applied for an extension [of funding for the nurses]. And we're already recognising that it's not embedded. So, will it ever be? Tell me. So, we're applying for another 12 months.*
>
> *AQ consultant lead, heart failure ID5, T23*

> *Clinical work is not cancelled, you keep on doing the clinical work at the same time. I ended up finishing late and completing my things the very next day to catch up with things . . . It's big challenge yeah. At some stages I thought I'll just leave it because why I should spend half my time every week, and every week is a nightmare . . . but the thing is it was dumped in my way so I have to be . . . because it was something . . . a project I had to do. So I thought we have to be successful.*
>
> *AQ consultant lead, pneumonia ID185, T18*

The challenge of 'structuring, planning and co-ordinating the quality and service improvement effort, and embedding it within the organisational fabric'[84] was ongoing throughout the study and we found no evidence that QI through AQ became embedded in the organisational fabric.

Using data to get attention

Given that there were no baseline measures of performance on AQ measures, programme leaders used data from a small-scale audit of a sample of local hospitals to highlight deficiencies in the current system. In addition to presenting these as percentages for each measure (i.e. 'x%' of patients received 'y'), the

data highlighted that patients did receive care which followed pathways corresponding to HQID process measures (e.g. for pneumonia patients – oxygenation assessment within 24 hours prior to or after hospital arrival, blood cultures performed in the emergency department prior to initial antibiotic received in hospital, initial antibiotic consistent with current recommendations, receipt of antibiotics within 4 hours of arrival, smoking cessation advice/counselling provided). However, the percentage of patients who received all of these interventions was disappointing, particularly in relation to pneumonia and heart failure.

One of the programme leaders explained that, at a meeting for potential evaluators prior to implementation, the results shocked hospital staff who on seeing the indicators initially had responded that 'we're already doing this'. It certainly helps for a problem to be countable and indicators of performance can help move issues up the agenda. Yet, at the same meeting, the statement by this programme leader that there were no 'antibodies' in the system, suggests an underestimation of the amount of effort required to move from getting attention to changing behaviour. Having grabbed people's attention, according to the theory, it was important to feed data back in a timely fashion and on a regular basis to sustain involvement and attention, although this was not what happened in practice, as we discuss in *Data reporting and verification*.

Using voluntarism and peer pressure to drive participation

The AQ community, which evolved over time, was helpful in maintaining voluntary participation via a process that involved normative pressures, together with support mechanisms for staff directly involved. However, as the programme developed, it was pressure to perform rather than to participate that became the dominant feature. Participation at a local level was also important and appeared to be affected by clinical/managerial relationships.

The introduction of AQ into CQUIN and the punitive (i.e. payment is withheld as opposed to being a bonus) aspect that AQ performance entailed, was seen by some as having the potential to contribute to an increase of clinical engagement. Given that AQ was now seen as part of a trust's funding contract and missed AQ measures would result in financial penalties (i.e. reduced budgets), there has been a move over time, by several trusts to resort to naming and shaming in order to present more personalised versions of AQ results and performance to clinical teams. The aim was to appeal to professional competitiveness and engage a wider array of clinicians and it was reported by some leads as resulting in improvements.

Comparing apples with apples

Compliance with the guidance provided in the detailed data dictionary was intended to ensure standardisation within and between providers. The dictionary (or 'the Bible' as it was often referred to) was seen as very helpful and informative, and participants described having learned much during the process of working with the dictionary. In the early stages of implementation, despite training to supplement the process and e-mail advice and support from Premier and SHA AQ team staff, there were different interpretations of the rules and guidance, which led some staff to panic or feel disheartened. At a training session we observed in April 2009, for example, one attendee was horrified to learn that she had been incorrectly interpreting the guidance.

Another in an interview described confusing all the discharges for a particular month with all the admissions for that month at the start of the process because of a lack of understanding about what was required. Other examples included not collecting PROMs data for all eligible patients and general uncertainty around reasons for exclusion from the measures. The learning process around AQ data collection and submission was a lengthy process, although over time as providers built up a stock of knowledge around this it became less burdensome.

The result was that difficulties with the interpretation of AQ measurement definitions were, to a considerable degree, overcome.

The competitive nature of AQ may have contributed to some distrust in the early stages of the programme, with a small number of low performers suggesting that performance was more a reflection of inconsistency in data interpretation between trusts (with high performers being less stringent in their application of AQ rules) than of real differences in service delivery. Staff from the Audit Commission, who were responsible for assuring the data, confirmed that in some trusts the guidance had been interpreted too narrowly, leading these organisations to under-report their performance, although following feedback this has been largely rectified.

The involvement of the Audit Commission, which undertook a number of data assurance processes, helped reassure participants over time that all were playing by the rules of AQ and comparisons were valid. Despite welcoming the input, participants also complained that assurance processes were burdensome, but, as these evolved, a differential approach with greater focus on providers where problems had been identified was implemented. Many of the issues identified related to participants applying rules too strictly, as opposed to being too lax, which helped reassure doubters.

The process of comparing and assuring data highlighted the issue of data completeness. It would be possible for providers to declare success in meeting all measures for 100% of patients, if they failed to report on all eligible patients. Rules were developed, consulted on and implemented to ensure that only providers meeting minimum standards with regard to completeness and accuracy were eligible to be considered for rewards. These processes appeared to be important in securing and sustaining buy-in, with discussions increasingly being focused on strategies and tactics for improvement and sharing learning. Over time, distrust around data began to disappear from the AQ community discourse. Overcoming the cultural challenge by systematic ongoing attempts to align definitions of quality was one of the key aspects of the AQ programme, and one which led the approach to be applied to other areas.

Providing strategic leadership by sustaining senior-level commitment

In terms of 'structuring, planning and co-ordinating'[84] AQ, in all providers accountability structures were established identifying a member of the executive team with lead responsibility for AQ. Lead directors spanned a range of disciplines including finance, information, nursing and medicine. In many cases the choice of executive lead appeared to reflect the way in which AQ had been conceptualised by executive teams (e.g. an informatics project, a quality initiative), in terms of the nature of the programme and the relevant skills set required, although this was not always the case.

There were instances of AQ executive leadership responsibility being reassigned in a pragmatic way (who might be available to fill a gap, as opposed to who would be suited to the nature of the task) following the departure of an executive lead. In some cases, executive leadership was seen by those charged with delivering AQ on the ground as lacking, insufficiently engaged and/or integrated, with this adding to pressures on other members of staff involved in AQ. A small number of people reported that, despite having a designated executive lead and formal accountability structures, senior staff were too busy with other priorities to actively engage with AQ.

> *I was sat there as an information manager . . . And [other trust CEO] was sat there representing [other trust] . . . well which one of these is going to succeed? The one that's got the Chief Executive sat there or the one that's just sent the information manager?*
>
> *Information manager, ID219, T15*

> *Now there is a big push to say right let's hit these targets . . . be in the top 25% . . . from the very top from the Chief Executive all the way down now there is a massive push to do it.*
>
> *AQ lead, ID55, T2*

> *Without the executives saying, yes we support you 100%, lots of projects die. You could put a lot of effort into it, but until somebody says, yeah okay, we'll give you the money, it will never happen. So we very early on established that we had executive buy-in, certainly from the medical director that was*

> *interested in quality. We had a quick chat with the Chief Executive so yes, we really want to make sure that we get this going right . . . there was an executive direct lead who was interested, wanted to know what was going on but didn't do any real hands on work.*
>
> *AQ clinical information officer, ID123, T14*

However, leadership meant different things to different people. In some sites this involved having a designated executive lead who responded positively to requests for resources, rather than to a nominated person at executive level driving forward QI initiatives. In some cases executive leadership was seen as lacking, insufficiently engaged and/or integrated, with this adding to pressures on other members of staff involved in AQ.

> *Get management and exec engagement straight away or as early on as possible and then it would work in the trust . . . [here] . . . suddenly panic, it missed out a load of people, it just suddenly went to somebody who had time to implement it and then onto somebody else who had time to do the data entry.*
>
> *Clinical information analyst, ID26, T4*

> *The Chief Exec doesn't really get involved or the Trust Board doesn't get involved.*
>
> *Clinical audit co-ordinator, ID222, T12*

Although the lack of formal structures for promoting executive leadership was reported as detracting from efforts to implement AQ, the mere presence of such structures was no guarantee of success. The extent to which executive leadership was sustained over time varied between providers, with executive leads in some sites only becoming actively involved when performance was poor. At one point, when a study was published which showed no impact, based on a large US data set,[9] one of the Chief Executives who had initially been very vocal in his support for AQ, raised concerns at the AQ Reference Board meeting. The fact that this provider had not been performing as well as expected may have been a contributing factor here. This highlights that evidence can be used as part of a political process, a point that the rather underspecified nature of the initial theory tends to overlook. However, it also highlights how the normative pressures of AQ ensured that participants remained within the programme, despite any reservations they might have about doing so. It is clear that, although there are structural challenges in improving quality, addressing them is likely to be necessary but not sufficient for improvement.

Institutionalising behaviour change

Although not specified explicitly in the initial programme theory, in relation to institutional change, we might frame AQ as attempting to move to a new dominant logic. Institutional theory places great emphasis on legitimacy and shared norms and AQ can be seen as attempting to institutionalise AQ pathways, which are seen as good practice. This is the political challenge described by Bate *et al.*: 'Negotiating the politics of change associated with implanting and sustaining the improvement process, including securing stakeholder buy-in and engagement, dealing with conflict and opposition, building change relationships, and agreeing and committing to a common agenda for improvement'.[84]

At the same time, what counts as good practice and, more generally, knowledge are highly influenced by political and cultural factors (in other words, cognitive frames are hugely important, as we discuss in *Chapter 6*, *Impact on mortality in the longer term*). This makes knowledge brokering by AQ leads and/or other staff difficult and is a key cultural challenge. Part of the problem related to power differentials[109] between core staff and clinicians. For example, many coding and administrative staff felt that they could not challenge clinicians. However, in common with Currie and White,[110] we found that the challenge of knowledge brokering beyond one's professional affiliation and hierarchical status was not insurmountable. This was related in part to professional roles, with some non-medical clinical staff less inhibited than their administrative counterparts. In some cases, specialist nurses were based relatively close to other groups of staff with whom they interacted on a regular basis. This facilitated the building of relationships, which enabled discussion and some modification of views. Even where this was not the case, the role undertaken

by specialist nurses meant that they interacted regularly with staff involved in care delivery. In some sites medically qualified clinical champions acted to reinforce the AQ message in a way that allowed the nurses and/or AQ leads to punch above their hierarchical weight in terms of their knowledge brokerage role.

Clinical–managerial partnering was a key element of brokering. There were reports of clinician resistance to AQ from some quarters in all trusts. Often this conflict concerned the process of data recording, as opposed to the content of AQ quality metrics.

The issue of junior doctors failing to adhere to AQ treatment metrics was a recurring theme. The rotation of junior doctors across different sites at regular intervals as part of their training, in combination with the time lag in AQ results, was seen as contributing to the persistence of resistance to change among this group of staff. In all trusts there were reports of clinicians failing to record activity; however, although not easily resolved, some participants reported that improvements on this had been made on this over time. The most frequent obstacles in sustained clinical–managerial partnerships were reported to be the differences of perceived priorities between the two groups and the assumption that managers lack knowledge and direct experience of frontline services.

The actions of core AQ staff and, where relevant, medical champions suggested that knowledge brokering is as much a group process as an individual one. However, even where groups worked well together to perform this function, getting clinicians to change behaviour was often done by following them around, rather than trusting them to comply with AQ requirements. Junior (and often more senior) doctors failed to act in accordance with AQ requirements and at times complained that AQ was merely an audit that got in the way of patient care. More generally, workload pressures contributed to staff forgetting to take action. The absence of data-recording technologies, to prompt action and embed behaviour change, compounded the problem. Instead of creating institutionalised change in ways of doing things, often the old ways persisted, with a relatively small number of dedicated individuals participating in an ongoing process of educating, prompting and cajoling, as well as rolling up their sleeves and plugging gaps where necessary. It is therefore unclear whether or not this cultural challenge was addressed in a sustainable way.

Advancing Quality requires the spanning of boundaries between levels, professional groups and divisions within the organisation as well as internally and externally. In theory at least, the designated AQ programme lead in each trust could perform such a role. The extent to which leads acted as boundary spanners using influence and negotiation to achieve AQ goals varied between providers. For example, in one trust much of the time of the AQ lead was spent closely monitoring data collection and entry, ensuring that these activities happened and that they did so in a timely fashion. Covering for data collection and data entry when the staff members responsible (one for each clinical area) were on sick leave or overloaded with other duties was also part of their role. Alongside this, the AQ lead in this provider was directly accountable for performance and was often pressured (in one-to-one meetings or through electronic correspondence) by trust executives to improve AQ results and secure more substantial financial rewards for the organisation. The limited clinical engagement here contributed to a situation in which much of the implementation of AQ appeared to be sustained by four data collectors and an AQ lead.

This contrasts with another provider where the AQ lead gradually moved to a supervisory role, acting as a point of reference for clinical and executive staff but not being held directly and solely accountable for AQ results. Performance was discussed openly in bimonthly AQ meetings, with regular executive and clinical attendance, where each clinical department presented their AQ work and expected performance. Data collection and data entry were undertaken by several clinical staff members in each area allowing sufficient cover and sustained performance when staff members are absent or carrying a heavy clinical workload. The AQ lead in this organisation provided input as the person who has the most in-depth knowledge of the AQ programme (data requirements, financial incentive changes and future developments) as well as a comprehensive overview of the implementation process in the organisation (past and present). Although even in this case, it took a lot of time to get to this position.

In some cases challenges surfaced and were conceptualised in terms of a presenting problem,[111] which turned out to be a symptom of a deeper issue. For example, in one trust where we observed meetings over a period of months, junior doctors were failing to record data required for AQ. Various steps were taken to improve recording including amending the data-collection tool (responding to a perceived physical and technological challenge), raising awareness among junior doctors (responding to a perceived educational challenge) and trying to build a shared understanding and commitment to AQ (responding to a perceived cultural challenge). However, despite attempts to address perceived challenges, and, although these staff were made aware of AQ and its requirements, they were largely resistant to it, which suggests that the challenge may also be about the politics of change surrounding AQ. Although it may be necessary to raise awareness and increase understanding and knowledge, where other challenges persist, such actions may not be sufficient to obtain engagement and improve AQ performance.

Nurses were also failing to take action on AQ measures. These staff were extremely unlikely to be unaware of the need to advise patients on smoking cessation, but a combination of work pressures and forgetfulness meant that many breaches occurred with regard to this measure. For many nurses, despite education on the importance of delivering this advice, recording this activity was not seen as important, which appeared to be indicative of a lack of a shared understanding and commitment to AQ. The fact that nurses needed constant reminders and nagging by AQ specialist nurses and other AQ staff is indicative of attitudes among many clinicians that AQ was not part of their core responsibilities.

A failure to record may sound like a minor transgression, on the grounds that delivering care is what matters. However, the act of recording involves interaction with technologies (electronic and paper medical records) in a way that helps change to become embedded and self-reproducing. Recording facilitates the institutionalisation of new ways of working so that over time (and this may be a relatively short time, e.g. see Checkland *et al.*[112]) they become taken for granted. Rather than seeing data-recording tools and templates as helping to reinforce the systematic delivery of high-quality care, however, a frequent complaint from clinicians was that AQ was an audit and, therefore, a paper exercise of little value.

Bate *et al.*[84] suggest that an important part of the process of overcoming political challenges and institutionalising change is strong and close partnerships with relevant external stakeholders. In the early phase of interviews there were occasional reports of PCT commissioners helping to support AQ by providing financial support for activities relating to AQ, but this was rare. We also found reports of increased partnership working with PCT providers in the delivery of services (prior to Transforming Community Services) related to AQ metrics, particularly around smoking cessation.

Interviews undertaken with PCT commissioners in Autumn 2009 indicated that many had little involvement in and/or not much detailed knowledge of AQ. These interviews were undertaken following the receipt of reports within the PCT (normally by the Chief Executive) on provider AQ performance, but many interviewees had not seen these reports. Among those who had seen them, there were mixed views about their content, which in part appear to reflect the capacity of commissioners to understand and engage with AQ. These reports presented the first formal results on performance to PCTs and commissioners reported little involvement in the process up to this point. In interviews with commissioners during the second half of 2010, most participants reported that they had little long-term involvement with the programme in the implementation process, as they mostly considered it an acute trust initiative. Nevertheless, with hindsight commissioners accepted that future implementation of AQ in regions outside the North West region would require a more hands-on approach from a commissioner's perspective in learning more about the internal processes and microcultures of the acute trusts to allow more effective monitoring and intervention.

The incorporation of AQ into the CQUIN payment framework might have been expected to encourage commissioners to play an active part in the monitoring of provider performance as part of regular meetings at which trust performance on CQUIN goals was discussed. However, there was little evidence that commissioners were more engaged following the introduction of CQUIN. Some participants suggested that introducing AQ into CQUIN in April 2010 did not allow enough time to assess the impact of AQ and

learn lessons from the implementation process. Furthermore, although in 2009 some commissioning staff did start to attend AQ events, the announcement in July 2010 of large-scale NHS reforms to include the abolition of PCTs meant that PCT staff interest and participation was severely reduced.

In the final round of commissioner interviews with CCG leads, the high costs of AQ was mentioned repeatedly, with most interviewees suggesting that they were uncertain about their continued support for the programme. Shortly after these interviews, a small number of commissioners withdrew from AQ.

Using feedback to inform learning

Data collection

'Formal data collection and information processing systems for monitoring, measuring and benchmarking of performance'[84] are viewed as key components of successful QI strategies. The establishment and development of systems to support the AQ data collection and monitoring process has been a significant challenge for participating Trusts and in some case it continues to be one. In many cases problems related to incomplete or inconsistent documentation from clinicians (doctors and nurses). However, the lengthy wait for data from AQ was also reported as making it difficult to assess and respond to performance monitoring information in a timely manner.

From the outset, the collection and monitoring of relevant data was a significant challenge for participating trusts and in most cases it continues to be one. Particularly difficult metrics were the pneumonia and heart failure ones, partly because of their diverse pathways and the extent to which participants reported being able to overcome these problems varied. Challenges also concerned the low documentation standards – the latter was a persistent obstacle which complicated and prolonged all subsequent data-collection and data-entry procedures. A continual problem arose from the rotation of junior medical staff.

More generally, although progress was made on this issue over the time, the difficulty of getting clinicians engaged and willing to alter existing documentation practices was a key challenge. Owing to, in many cases, long-established habits, even where clinicians were more vigilant in relation to care delivery covered by AQ measures, they often failed to see the importance of documenting their actions, or else viewed this as a distraction from care delivery in a context where time was highly constrained.

The evolving approach to data collection at the different participating sites resulted in considerable variation in data capture and entry which largely depends on the availability and/or standards of the above areas (i.e. existing resources, documentation standards and clinical engagement). This kind of flexibility also provided space for local innovation in the search for more efficient techniques of data collection.

Data reporting and verification

The lengthy delay between submission and feedback of validated information was a big problem. Initially providers found that there was a sizeable mismatch between Premier's data and the number of patients identified by AQ teams. These patients had slipped through the net, often as a result of being on outlying wards, and this meant that hospitals were failing in relation to AQ measures for these patients. The long wait for AQ data was contrasted with local Plan Do Study Act (PDSA) initiatives[113] and audits in other areas of the hospitals, using small-scale sample data which allowed quick feedback to assess the impact of any changes made.

> *In December, if I employed a new heart failure nurse . . . I would have to wait 12 months to know whether or not the implementation of that new nurse had made a difference. Because you'd have to wait 4 months for the AQ data to catch up with the fact that you've appointed somebody. Then you'd need eight data points to show a significant change in the process . . . So you make your one improvement, potentially, a year. As opposed to very rapid improvements. I don't quite understand at*

> *the moment how AQ is designed to drive quality as opposed to improve performance and hit targets. I think lots of clinicians don't quite understand that.*
>
> *Specialist registrar, ID38, T15*

> *If I was in a hospital which was used to getting a rapid data analysis back, and using PDSA cycles to improve quality of care, then this would feel very disengaged from frontline. We're not that sophisticated . . . But what is an issue, and does feel as though it's taken a very long time is the fact that we're putting data into QMR .. . . And then after it's been validated it gets copied over to Advancing Quality. Then they try and match it up. And then there's data queries. It's all running months in retrospect.*
>
> *Associate medical director, ID39, T1*

Many staff reported feeling under pressure to meet data collection and input timetables, so that attempts to reduce the turnaround time, by shortening the submission timetable, would not necessarily be welcome. At the same time, the process of data collection helped identify issues within hospitals that needed addressing. Additionally, it brought together people who would not normally talk to each other such as coders and clinicians and the latter began to appreciate the role and contribution to AQ of the former, as well as the need for better communication between the two groups. These meetings enabled coders to put faces to names (as opposed to dealing with anonymous and distant doctors) and helped to build relationships based on mutual respect. Furthermore, meetings involving coders and clinicians often highlighted differences of opinion and diverse approaches to the same condition among clinicians. These issues covered a spectrum of topics from data entry and coding, through clinical training to pathway design and redesign and helped with the process of standardisation.

Over time, the discrepancy between patient data reported by Premier and patient data identified by AQ staff diminished, as hospital staff became more adept at identifying patients and working within the requirements of the Premier data-collection tool. The feedback delay continued but participants began to trust their own data, rather than expecting it to be deficient and waiting to receive validated data from Premier. This enabled them to compare the current month with previous performance and, in some cases, to collaborate with other providers to share data and benchmark performance.

The delay (reduced to around 3 months eventually) continued to be a problem. In particular, it made it difficult, if not impossible, to feed back validated performance data to show the impact of any changes made to junior doctors who spend relatively short periods of time at a number of hospital sites as part of their training. This was viewed an important obstacle to AQ delivery, as these staff play a key part in the care of AQ patients, and was also a missed opportunity to address educational challenges.

In terms of learning by comparing feedback on current activity with historical trends prior to AQ, there appeared to be very few providers engaged in this process in the early stages of AQ. Many people reported being taken by surprise by a surge of activity in December 2008, having based their expectations on admissions for October and November rather than comparing activity for these conditions with trends based on previous years. The absence of a baseline meant that most programme leads were content to monitor progress from October when the programme went live, and programme leads were mostly not trained or experienced in the analysis of historical data trends. However, some participants expressed concern at the lack of baseline data.

> *So, we've got no baseline, but we've been doing improvement work since September. The first data submission was for October patients. So we'd already started. We have no idea whether our AQ data, which might be say 58 per cent compliance for heart failure, is actually 30 per cent higher than it was already. Or whether that's actually at baseline; we've made no improvement at all. We don't know that. We also don't know whether our ongoing work is making any difference because we're only up to February status. In order to have a look to see whether the stuff we're getting out of AQ is where we're at, or an improvement on where we were, we need to collect baseline, pre-AQ baseline data.*

> *Which will involve going back and sampling at least ten sets of notes for discharge. Every condition for the 6 months prior to AQ stuff.*
>
> *Specialist registrar, ID38, T15*

Furthermore, because seasonal variations may mean that admissions are not constant in each month, sampling for the 6-month period prior to AQ may not provide an accurate picture for a 12-month baseline so that, in one sense, the first year of AQ might be regarded as a baseline against which to monitor subsequent performance.

The educational challenge

'Embedding and nurturing a continuous learning process in relation to quality and service improvement issues, including both formal and informal mentoring, instruction, education and training, and the acquisition of relevant knowledge, skills and experience'[84] has been a key focus throughout the AQ programme at all levels. A large part of the work undertaken by AQ leads and specialist nurses recruited to improve AQ performance has involved raising awareness of AQ, educating staff members about treatment pathways for AQ patients and feeding back performance data to staff involved in AQ pathways.
These activities have not been easy. Although good progress has been made in educating staff in relation to activities necessary to deliver pathway compliant care, data recording to demonstrate compliance continues to be a problem, with many clinicians seeing this as low priority compared with care delivery.

In most cases it was reported that there was very little knowledge of AQ and associated processes outside the teams and staff members who were directly involved in it during the first 12 to 18 months of the programme. In some cases providers had focused on establishing structures but had neglected some of the educational aspects, which were necessary (although not sufficient) to contribute to a shared cultural commitment to and understanding of AQ.

> *The notes are requested from medical records. The list goes down . . . to [the clerk who] gets all the notes out and then I go and pick them up and go through them. I tend to pick them up on a specialty basis. I try to do all pneumonias at once and then all the hips and knees . . . I'm not sure where they [Premier] get their information from. I just wait on this list . . . I have not been told too much about the purpose of looking for the information and so on. I just flipped through the profile material and had a look at what information I needed to find from the notes . . . I don't know why they want antibiotics . . . and I don't know why it's time specific all the time.*
>
> *Clinical audit facilitator, ID41, T15*

In terms of other potential sources of learning, as part of the support process from Premier, AQ participants had access to the Premier Performance Improvement Portal. This was a source of performance-improvement advice and best-practice information. The portal was intended to provide a forum for sharing knowledge and good practice. Almost all participants who were aware of the portal reported being too busy to use it.

Learning from experience

Bate *et al.*[84] describe a culture that values 'risk-taking and experimentation', constantly encouraging people to do more and differently, developing and sharing new knowledge, skills and expertise. All respondents described processes of adaptation and experimentation. In many, but not all, cases this led to learning and developing new understanding related to AQ. There were some examples of learning from evidence and experience which highlighting the challenges involved.

> *For heart failure . . . patients weren't given any instructions at the point of discharge and nothing was documented in the notes about what they'd been told . . . the heart failure team . . . they've been spending a lot of time researching what's going to be best for patients and developing their own ideas, and we're finally just now at the stage of a patient information leaflet that's been developed jointly by the community heart failure nurses and the ward nurses and signed off by the consultants,*

which just gives patients the basic information that they need. But they'll supplement that with information from the British Heart Foundation, and also one of our staff nurses has designed a patient-held alert card that patients will be given on discharge. So it's taken them a while to work through their ideas and agree what they wanted and then design things . . . but it takes time in large organisations with complex relationships.

AQ lead, ID12, T11

We're having trouble with the nurses . . . especially around the heart failure patients, with some of the measures there. It's such a wide, such a big group to get to. We try to do awareness on the ward, we've done workshops where we just had drop-in sessions. But, obviously, the ward's just so big . . . It's getting people off to do education. We feel as though, we've done quite a lot, but there's definitely a long way to go.

AQ lead, ID6, T23

Using competition to drive performance

Participants were acutely aware of the competitive nature of AQ. Although it was many months before initial first-quarter comparative data were made available, the commitment to public reporting and the tournament nature of the incentive system encouraged participants to compare their progress with that of other organisations. Striving to be near (or at) the top of the league table motivated people but it also placed them under pressure. Comparative data were presented at meetings where all AQ leads were present and the excitement, delight and disappointment were often palpable.

One participant described feeling 'relieved but not complacent' on seeing the figures because 'it wouldn't take much for you to slip just a tiny fraction and lose the incentive payment'. So, even among high performers, the results did nothing to diminish the feeling of being under constant pressure.

Advancing Quality programme leads used data sharing within their hospitals to try to use competition to motivate local clinicians, with varying degrees of success. As hypothesised, normative and coercive pressures prompted action in response to performance deficits.

Among those in each organisation who were responsible for delivering some aspect of the AQ programme on a day-to-day basis, AQ appeared to have its own momentum. The constant round of monthly data collection and reporting was a stressor. For AQ programme leads, participation in the process meant taking an emotional roller-coaster ride as they grappled with the highs and lows of the programme. None of the AQ leads had envisaged that the process would be so difficult and time-consuming, and poor performance on AQ metrics could leave participants feeling very low. Early on in the process, we interviewed one AQ programme lead whose office was full of mugs and pens that he was 'too busy' to distribute, and the practice of adding AQ to existing workload may explain such attitudes. In such cases, staff were replaced by others with a more positive orientation to AQ.

The competitive nature of AQ also resulted in AQ leads coming under increased pressure, as opposed to executive staff taking action, which translated into behaviour change among frontline clinicians. The use of evidence, combined with public reporting, was used to attempt to engage clinicians or, in extreme cases, to shame them. As one exasperated AQ lead recounted 'And I've used phrases like, "Do you want to be viewed as the worst evidenced surgeons in this region?"'. However, the challenge of getting clinicians to follow evidence may also have been a presenting problem not solely amenable to resolution by such pragmatic approaches.

Achieving culture change and whole systems working takes time, yet the competitive nature of AQ, with results to be publicly reported, meant that short-term improvements were needed. This is one reason why AQ involved going around people rather than relying on persuading them to change their practice.

Using collaboration to drive improvement

Cross-site collaboration

Communities of practice are cross-organisational and occupational networks and groups that come together regularly to debate, share knowledge and take forward the QI agenda.[84] There were identifiable attempts to develop a community of practice in all the trusts, supported by the SHA AQ team and Premier.

In addition to bringing together AQ leads at regular leads meetings, the programme soon introduced collaborative learning events for the areas of greatest challenge – pneumonia, heart failure and AMI. In part the initial idea for this was informed by experience of Premier staff who had been involved in QI collaboratives[114,115] to support the HQID initiative. In AQ, these events involved taking people out of their work environment and bringing them together to discuss common problems in a structured way but also providing ample opportunity for informal networking over a pleasant lunch and various coffee breaks. This also helped people to make contact with their opposite numbers outside meetings. As this evolved, participants brought storyboards which were displayed around the venue and which various members of the AQ team manned, while other AQ staff from all providers circulated and listened to short (5 minutes) presentations. After a quick exchange, a buzzer sounded moving people on. This helped facilitate serious discussion but in the context of a somewhat frenetic and light-hearted manner, which made it easier for staff at the storyboards to present failures (or 'opportunities') as well as success.

Early on in the programme a residential event was held for AQ programme leads in the picturesque surroundings of the English Lake District to enable people to interact in an informal way and also to reward the efforts of those charged with managing AQ on a day-to-day basis. More generally, having fun was seen as important for participants. For example, at one collaborative event with Christmas approaching, in addition to mince pies with their coffee, participants were given tubs of Premier play foam and allowed to play at making models of snowmen, fir trees and other relevant yuletide symbols during the serious business of the meeting, with prizes being awarded for the best team at the close of the day.

The Premier portal, a virtual learning forum, was created to enable participating staff to share knowledge, good practice, problems and solutions online without having to leave their desks. Almost all participants reported being too busy to use it, yet the collaborative events were whole-day sessions away from the office but were always well attended. The portal appeared to fail in part because the busy, reactive behaviour of AQ staff did not fit well with proactively engaging on an ongoing basis in a virtual forum. This suggests that taking people out of their working environment was necessary to facilitate shared learning. Beyond this, our observations and interview data suggest that people found more value and pleasure in face-to-face exchanges than virtual conversations.

There was widespread agreement by participants that the SHA AQ team (subsequently provided by AQuA) collaborative events facilitated the establishment of a region-wide network which was reported to provide valuable emotional support and sustained motivation. Collaborative events were also reported to have encouraged learning and the sharing of good practice despite the competitive element of AQ.

The extent to which participants valued the collaborative events was reflected in the negative responses to the switch to half-day (instead of full-day) collaborative events in the third year of the programme. These changes were made to reduce time demands on participating staff and excluded the use of storyboards where each provider explained progress to date, mistakes made and lessons learned. Several participants reported feeling very disappointed by this move on the grounds that they expected it to significantly limit their opportunity for learning and networking with colleagues. Furthermore, in some cases, this was perceived as threatening future engagement of medical staff, as some AQ leads reported that having to invite doctors to such half-day events (without storyboards) would be 'a waste of their time'.

On-site collaboration

A group collaborative culture refers to a strong 'we group' culture that promotes teamwork and cooperation between staff and places a premium on values such as respect, integrity, trust, pride, honesty, inclusion and openness.[84] Participating trusts reported varying experiences in this area. Some reported difficulties in engaging different groups mainly because of staff members' time limitations and existing heavy workload, whereas others described AQ as improving collaboration and communication between previously disparate groups of staff within the organisation.

But there were also providers who reported the engagement of different staff groups in AQ work across different levels, something that previous QI initiatives had not managed to do in the past.

Both types of collaboration enable the emotional challenge to be addressed, 'Energising, mobilising and inspiring staff and other stakeholders to want to join in the improvement effort by their own volition and sustain its momentum through individual and collective motivation, enthusiasm and movement',[84] although the limits of success in addressing this are clear.

Using money to spur improvement

Economic theory would suggest that providers need to be compensated for the costs of investing in service changes related to P4P,[33] and at some sites it was necessary to make a business case for additional staff to demonstrate that this would be at least cost neutral. However, in many cases, investments were made without such information, largely on the grounds that poor performance was deemed unacceptable by participating organisations in a context in which the normative pressures surrounding AQ performance were strong.

The initial theory was based on a commitment to rewarding clinical teams, but ad hoc investment in coding, audit, information and clinicians (specialist nurses) meant that there was often no clear link between rewards and teams as envisaged in the theory. However, this did not appear to detract from the commitment of core AQ staff and a combination of evolving routines and practices together with feedback of relative performance ensured that they maintained their focus.

Facilitating standardisation

Although the implicit theory appears to assume compliance and convergence, with organisations coming to resemble each other in terms of standardised approaches to care, some organisational characteristics may be more open to change than others. Ashworth *et al.*[116] hypothesise that the impact of these pressures on organisational structures and processes is stronger than on strategy and culture (although their results are not entirely in accordance with this hypothesis). This resonates with the evidence on the difficulties of securing culture change in health organisations[117] and with the AQ experience.

As the programme evolved, structures and processes were developed to plug gaps in care pathways and improve care coordination. The AQ measures represented requirements which, in the sense that they were non-negotiable (on a day-to-day basis), were coercive in nature. We observed two types of standardisation:

Standardisation within sites

Advancing Quality placed additional demands on clinical areas that formed part of emergency admissions (such as pneumonia, AMI and heart failure) at a time when these services were already under pressure and often the focus of a range of improvement activity. Early on in the evaluation, hospitals were struggling to achieve high rates of compliance with many of the AQ measures, particularly in relation to heart failure and CAP. As providers investigated the process of care, it became apparent that the fragmented nature of care for these patients was leading to delays, omissions and substandard service delivery. Part of the problem related to the absence of a clear diagnosis for some patients, which led to delays in treatment and poor or inadequate care on arrival. For pneumonia, uncertainty concerning diagnosis also added to the difficulties, particularly with regard to the requirement that these patients received antibiotics within 6 hours of hospital arrival. Over time, in many trusts, clinicians came together to examine antibiotic

protocols, care pathways and prescribing behaviour, resulting in more standardised prescribing and better compliance with AQ metrics in these areas.

Although structures had been developed aimed at ensuring that care was consistently in line with the AQ measures, these structures were sometimes constrained by other policies and processes intended to improve care. For example, in several trusts, in order to reduce infection transmission, the taking of blood cultures has been reduced, in terms of both the number of cultures taken and the number of people permitted to take blood. This made taking blood cultures prior to the first antibiotic administration difficult. However, these challenges brought staff together to ensure alignment between AQ requirements and other policies.

There was also a failure to follow up patients in a timely manner once they were admitted, in part as a result of poor handovers on admission and patients being placed on a range of wards depending on bed availability rather than appropriateness of care. Analysis of patient flows, based on patients for whom AQ processes were not delivered effectively, highlighted gaps in coordination and knowledge, which providers have made efforts to fill. As part of this process, providers began to systematise care by formalising pathways. In addition to more standardised approaches to care, examination of patient journeys also led to care being delivered in a timelier manner than had hitherto been the case. It is likely that this increased internal standardisation will have also affected patients whose conditions were not part of AQ.

Cross-site standardisation

Learning from others at collaborative events resulted in mimetic processes as stories of success and failure encouraged avoidance of some tactics and adoption of others. At all sites there was some element of staff training aimed at increasing awareness of AQ and thereby changing behaviours, although this varied between sites.

Similarly, initiatives developed at one site spread to others when these appeared to improve performance. These included:

- providing clinicians in the accident and emergency department laminated cards with the CURB-65 tool to guide prescribing of antibiotics
- provision of patient information leaflets intended to remind clinicians to deliver discharge instructions and cover all relevant aspects
- standardising antibiotic protocols to simplify prescribing and increase its appropriateness
- electronic alerts to notify specialist nurses when AQ patients are admitted.

(The CURB-65 uses a clinical prediction rule validated for predicting mortality in CAP and is recommended by the British Thoracic Society for the assessment of severity of CAP.)

These mechanisms were reported as changing behaviour and thereby improving care for patients in many cases. However, achieving and maintaining behaviour change was an ongoing process requiring constant vigilance on the part of AQ leads and/or specialist nurses.

In many sites, action was taken to improve the timeliness of data coding, given that tackling clinicians about AQ opportunities was much more meaningful when undertaken a day or two after patient discharge, rather than a month later. Despite this and other measures adopted, the extent to which clinicians complied with AQ requirements varied between sites, which meant that, even where specialist nurses were recruited, the ways in which changes were implemented differed. In other words, apparent standardisation often concealed rather different processes and contexts.

Chapter 6 Discussion and conclusions

Impact on mortality in the short term

Our evaluation of AQ suggests that the programme was associated with a significant reduction in patient mortality during the first 18 months of the programme. Risk-adjusted mortality rates for all three of the conditions we studied (pneumonia, heart failure and AMI) decreased over the study period in both the North West region and the rest of England. The reduction in mortality for incentivised conditions was significantly different and greater in the North West region than in the rest of England. The reduction in mortality over the 18-month period studied for non-incentivised conditions was not significantly different between the North West region and the rest of England.

Cost-effectiveness of Advancing Quality

Based on the first 18 months, we found AQ to be a cost-effective use of resources. The total cost of the AQ programme was just over £13M over the initial 18-month period, with only £5M of this consisting of the financial incentives. The ongoing running costs of the scheme exceeded the bonus payments, making up the majority of the costs, at just over £7M. We estimated a gain of 6700 QALYs as a result of the reduction in mortality for the programme as a whole. At a QALY value of £20,000, this equals an estimated health gain worth £134 million. Our estimates suggest that AQ also resulted in a reduction of 22,700 bed-days in the first 18 months. This is equivalent to a £5M reduction in costs.

Relationship between performance on process measures and short-term outcomes

The average performance reported by the participating hospitals on all of the quality measures improved in the first 18 months and improved further in the following 24 months, particularly for heart failure and pneumonia. Some of the process quality measures were significantly associated with better health outcomes at a trust level, but the magnitudes of the estimated coefficients were too high to represent clinically plausible direct consequences of these process measures. The findings suggest that these financial incentives to improve quality only weakly led to improved patient outcomes through their direct effects on the process measures that were incentivised.

Advancing Quality appears to have also led to improved patient outcomes by inducing positive spillovers in terms of wider improvements in care quality across unmeasured dimensions and improvements in care for all patients. Our qualitative data provide support for this explanation, highlighting developments at sites (e.g. recruitment of specialist nurses to join up gaps in care and maintain a sustained focus on patients as they moved through the hospital) to improve care quality for patients in AQ clinical areas. They also suggest that clinician compliance with data recording requirements varied between clinicians and across sites. Performance on process measures reflects what is recorded, as opposed to the care that was delivered and failure to record care delivery in a systematic fashion was a persistent problem. This further complicates the issue of quantifying relationships between performance on process measures and relevant outcomes.

Impact on mortality in the longer term

When we looked over the longer term from 18 to 42 months, risk-adjusted mortality rates continued to decrease in both the North West region and the rest of England for both incentivised and non-incentivised conditions. The reduction in the rest of England was significantly larger than in the North West region and was concentrated in pneumonia. However, the reductions in mortality were also larger for the non-incentivised conditions in the North West region compared with the rest of England between these periods. For incentivised conditions, there was a larger reduction in mortality for the rest of England than in the North West region between the short- and long-term periods.

We considered various explanations for the smaller reduction in mortality for the incentivised conditions in the North West region in the long term (i.e. at 42 months) compared with the rest of England. The first potential explanation we considered is the effect of the change in incentive structure, as the AQ programme switched from a tournament scheme with bonuses to a scheme involving penalties for failure to reach quality benchmarks (CQUIN). We did not find significant effects of the change in incentive structure on achievement of the process measures. The continued improvement in performance on incentivised process measures in the AQ hospitals suggests that the incentives may still have been effective, but we have no data from control hospitals for these measures.

A further possible explanation is that there was a positive spillover from the adopting region (i.e. the North West region) to other regions. We considered the possibility that the loss of effect might be because of improvements in care in the control regions or in non-incentivised conditions in the intervention hospitals. We found limited evidence of a positive spillover effect for both of these. In particular, the early results had been widely disseminated in England and two regions had adopted a form of AQ incentives; these regions showed a greater reduction in mortality in the long term compared with other control regions which did not incentivise the AQ indicators, although the reduction was only statistically significant for AMI. We also found limited evidence for positive spillover effects within the AQ hospitals in that non-incentivised conditions which were treated by specialists who also treated the incentivised conditions in AQ hospitals showed the largest reductions in mortality among the non-incentivised conditions in the long term. It is also possible that initial improvements had the greatest effect on the sickest patients, leaving less room for subsequent improvement in mortality.

An additional possible explanation is that there were positive spillovers in quality of care from participating to non-participating hospitals and from incentivised to non-incentivised conditions in the participating hospitals. We found some modest evidence for both these hypotheses. After the early results showed reductions in mortality, two other English regions began to incentivise the AQ measures during the long-term period of our study, albeit with none of the supporting mechanisms of the AQ programme. These regions had a larger reduction in mortality for the incentivised conditions in the long term compared with other English regions, although the reduction was only statistically significant for AMI. Our finding that the largest reductions in mortality for non-incentivised conditions were for those treated by the same specialists as those treating patients with the incentivised conditions also lends some support to the hypothesis that there might have been positive spillovers in the AQ hospitals.

Despite continued improvements in the process measures that were incentivised, our findings provide no evidence of a long-term impact of the incentives on outcome as judged by 30-day mortality. Possible reasons for this include short-term effects which were not sustained, early impact on outcomes which were easier to achieve (low-hanging fruit), changes in the incentives from bonuses to withholds and unintended but desirable spillover effects into other geographical and clinical areas.

The qualitative findings provide some explanation of how the benefits quantified previously were delivered as well as highlighting challenges, many of which persisted throughout the evaluation period. In broad terms, AQ can be conceptualised as introducing new rules into the field inhabited by the 24 participating acute trusts. However, rules are not self-implementing. Furthermore, health-care fields are characterised by

rules that encourage some behaviours and discourage, or sometimes outlaw, others. This meant that AQ rules did not always sit easily with existing rules. In some cases, there were rule conflicts that had not been anticipated.

Many of the features of AQ can be conceptualised in terms of creating new network structures and attempting to influence existing network structures. For example, the central regional AQ team, comprising experienced NHS managers together with dedicated support and infrastructure from people with direct experience of a similar QI initiative (Premier staff), linked participating hospitals and facilitated communication between them. Connections between people in network structures were influenced by, and were part of the process of influencing, the formal AQ rules. For example, the AQ governance framework included groups established to review and develop rules as part of an ongoing process of programme evolution. At the same time, in order to ensure that formal rules were complied with, it was necessary to introduce assurance processes and the Audit Commission, a national body external to the NHS, was contracted to conduct site visits and data quality audits. Connections between people in network structures were influenced by, and were part of the process of influencing, the formal AQ rules. For example, the AQ governance framework included groups established to review and develop rules as part of an ongoing process of programme evolution.

Making connections between relevant people in network structures also occurred within hospitals. AQ leads were instrumental in attempting to modify network structures as part of processes to enable data input, monitoring and feedback. At the same time, the process of data collection was useful in helping to identify clinical and coding issues within hospitals that needed addressing. More generally, it highlighted dependency relationships between staff, as well as the absence of mechanisms to bring these staff together to facilitate shared understanding. Adding specialist nurses to existing structures meant that these nurses in collaboration with core AQ staff (the programme lead and whoever else could be co-opted) often worked around obstacles such as intransigent or forgetful staff when attempts to change behaviour appeared to be ineffective.

Changes to rules and network structures were necessary, but, in order to create change, it was important for participants to perceive this change as desirable and legitimate. Developing shared learning and understanding involves discussion as ideas are exchanged and refined. Cognitive frames do not exist in some free-floating state, instead ‘the individuals who participate in the collective experience having developed differing interpretations of this experience, [must] . . .. refine [them] into a collective interpretation through a process of discussion and argument (a process pervaded with power)’.[118]

Within hospitals, team meetings involving staff whose work was directly related to the achievement of AQ goals contributed to the development of shared learning and understanding to some extent. However, for busy clinicians and for staff in rotational roles particularly, the opportunities to participate in the development of collective understanding were limited or non-existent. Furthermore, relying on one-off training events to change cognitive frames produced disappointing results. This meant that, rather than AQ becoming institutionalised as AQ staff hoped, constant vigilance from a relatively small number of AQ staff at each site was required to ensure that action and data entry were maintained. This contrasts with events at the broader programme level, where various forums were created which enabled staff working in AQ areas to discuss the programme and engage in discussion and develop shared understanding.

The story of AQ might be seen in terms of a linear narrative with rule changes at the beginning. However, the reason these rules were adopted was a result of the existence of networks which connected relevant stakeholders who came together to discuss and develop collective understanding. In other words, the relationship between the different aspects of the field (rules, network structures and cognitive frames) are interlinked and act together in a dynamic way. We, therefore, should not conceptualise the change in terms of a linear process, with a beginning, middle and end.

Furthermore, although many QI initiatives draw on collaborative approaches, most involve relatively short time frames and are conceptually underpinned by a series of short-term PDSA cycles.[119] This suggests a view of QI as involving an initial (short) implementation phase after which new ways of working become routine. Such an approach runs the risk that participants are encouraged to overcome resistance or lack of knowledge as part of a one-off process. Instead, although organisations are quasi-stable entities, they are also sites of ongoing change. It makes sense, therefore, to view 'implementation' as an ongoing process requiring continuous effort.[120]

Our findings suggest that creating new network structures and changing cognitive frames (see *Box 3* examples) takes time and effort. AQ was not a knee-jerk response to a perceived crisis. Instead, ideas were floated and refined in discussion over time. Furthermore, although money may not always be necessary to create new network structures, the AQ reward money (actual and potential) helped hospital staff who were engaged with AQ to make the case for additional staff members such as specialist nurses. Bringing people together to solve problems in comfortable venues with appetising food and a dedicated regional AQ team all required funding, without which new structures and events to facilitate shared cognitive frames would be difficult to maintain. In the context of economic recession and a continual drive for efficiency, organisations need to preserve 'organisational slack',[85] as this is hugely important for QI processes.

Our findings show that competition did not drive out collaborative learning, despite financial incentives to compete rather than co-operate. Cohesive network relationships support the social enforcement of anticompetitive norms.[121] The creation of connections between relevant people in network structures in AQ, combined with rules that encouraged participants to compete, helped to reap the benefits of collaboration and competition while maintaining good network relationships. In other words, rather than having to choose between competitive or collaborative approaches, it is possible to design programmes to encompass both. However, design extends beyond consideration of formal rules, to encompass consideration of network structures and changes in cognitive frames.

We considered various explanations for the smaller reduction in mortality for the incentivised conditions in the North West region in the long term (i.e. at 42 months) compared with the rest of England. The first is the possibility that the scheme became less effective with the change in incentive structure, as the AQ programme switched from a tournament scheme with bonuses to a scheme involving penalties for failure to reach quality benchmarks (CQUIN). The continued improvement in performance on incentivised process measures in the AQ hospitals suggests that the incentives may still have been effective, but we have no data from control hospitals for these measures. However, as described previously, we did not find a significant relationship between performance on process measures and outcomes.

A second possible explanation is that there was a positive spillover from the adopting region (i.e. the North West region) to other regions. The early results of AQ had been widely disseminated in England and two other regions had adopted a form of AQ incentives. These regions showed a greater reduction in mortality in the long term compared with other control regions which did not incentivise the AQ indicators,

BOX 3 Examples of changes to the organisational field following AQ introduction

Formal rules: new AQ rules to reward quality.

Network structures: additional staff (e.g. central AQ staff team, administrative/data-collection personnel, specialist nurses).

Cognitive frames: communication forums (e.g. collaborative learning events, within hospital meetings, AQ lead meetings) to develop individual and collective perceptions of field values and activities.

although the reduction was only statistically significant for AMI. We also found limited evidence for positive spillover effects within the AQ hospitals in that non-incentivised conditions which were treated by specialists who also treated the incentivised conditions in AQ hospitals showed the largest reductions in mortality among the non-incentivised conditions in the long term. It is also possible that initial improvements had the greatest effect on the sickest patients, leaving less room for subsequent improvement in mortality.

A number of factors appeared to contribute to the success (as measured by improving performance on process measures and mortality at 18 months) of the scheme. These include in-person collaborative learning events, dedicated infrastructure support, financial rewards to invest in additional staff and a combination of competition to spur improvement and collaboration to facilitate learning. Additionally, programme participants were able to contribute to shaping the programme as it evolved, enhancing legitimacy and buy-in.

At the same time, there were a number of barriers to implementation. In the context of heavy workloads and competing priorities, frontline staff did not always adhere to AQ requirements. Furthermore, data collection was burdensome in a context in which AQ was not part of existing electronic patient information systems. AQ did not become institutionalised and embedded into routine behaviours. Instead, there was a reliance on core AQ staff to cajole and persuade, which often resulted in going around obstacles, rather than resolving enduring problems.

Although there were some common themes in the approach taken (in particular, the employment of specialist nurses), more generally, hospitals implemented AQ using a range of activities tailored to and developed in their local context. This suggests that there was no one blueprint for implementing AQ in each site.

In terms of impact on commissioners, input from staff in commissioning organisations was relatively limited in the first year of AQ. Although some commissioner staff had begun to engage with AQ by year 2, the subsequent reorganisation of NHS commissioning functions during the study period meant that input from commissioners was limited or non-existent for most of the study period.

The AQ scheme design incorporated features of what our literature review identifies as good practice. It did not involve penalties and it rewarded relative, as well as absolute, performance. The fact that participation was on a voluntary basis and was universal (i.e. all 24 eligible organisations took part) appeared to add to AQ's legitimacy. Additionally, the competitive nature of the scheme did not crowd out knowledge sharing and collaboration more generally. However, our findings which highlight implementation challenges and a failure to embed change in routine practice suggest that, although scheme design is important, there are other aspects relating to implementation which require attention if financial incentive schemes are to fulfil their potential.

Concluding remarks

Based on the first 18 months, AQ was a relatively cost-effective intervention. The findings at 42 months are open to interpretation. Our failure to find a relationship between process and outcome measures at 18 months suggests that there were positive spillovers beyond the changes in relation to AQ measures. An alternative interpretation, however, is that short-term improvements were not sustained and that the observed improvements in mortality in the non-incentivised conditions within hospitals participating in AQ were unrelated to AQ.

The first explanation is supported by changes to care delivery identified by our evaluation. It may be that there were further positive spillovers in quality of care both from participating to non-participating hospitals and from incentivised to non-incentivised conditions in the participating hospitals. We found

some modest evidence for both of these hypotheses. However, we did not explicitly focus on non-incentivised conditions. Furthermore, because we collected qualitative data from a large number of sites ($n = 24$), we were unable to conduct detailed, in-depth research to explore these issues in a comprehensive manner. Further research to investigate the relationship between AQ and changes in incentivised and non-incentivised conditions would shed light on this area. Linked to this, research exploring changes in rest of England sites would also add to our knowledge.

The study highlights the importance of considering costs beyond the incentive payments of financial incentive programmes intended to improve care quality. It also suggests that, contrary to economic theory, competition did not inhibit collaboration, with providers keen to share learning within the AQ community of practice. Instead, cohesive network relationships appeared to support the social enforcement of anticompetitive norms. In-person collaborative learning events were an important part of building and sustaining such relationships.

We found no evidence of changes in care resulting from AQ being institutionalised. Instead, modifications to practice were generally not systematised and behaviour change was still largely reliant on prompting by particular individuals. The success of AQ seems to have been a result of persistent and focused individuals working to remind staff and to plug gaps in data collection and/or care pathways. Furthermore, far from being everybody's business and part of organisation-wide change, AQ was delivered in a context in which many staff were unaware of its existence.

Acknowledgements

We acknowledge the contribution of our dear colleague Dr Helen Lester, who died. Helen was a valuable and productive member of the team from the outset of the evaluation until her death in March 2013.

We are also extremely grateful to all of the people who gave their time generously to assist in our study.

This project was funded by the National Institute for Health Research Health Services and Delivery Research programme, project number 08/1809/250, and we are grateful for this funding, without which the study would not have been possible.

Contributions of authors

Ruth McDonald was the principal investigator. She was involved in the design of the study, the collection of data in phases 1 and 2 and the data analysis, and contributed to the final report writing.

Ruth Boaden was involved in designing the research project, contributing to the analysis of qualitative data and writing the final report.

Martin Roland was involved in designing the research project, contributing intellectual input with regard to analysis and interpretation of data as well as bringing a clinical perspective. He also contributed to the writing of the final report.

Søren Rud Kristensen was involved in the analysis of quantitative data and contributing to the writing of the final report.

Rachel Meacock was involved in the analysis of quantitative data and contributing to the writing of the final report.

Yiu-Shing Lau was involved in the analysis of quantitative data and contributing to the writing of the final report.

Tom Mason was involved in the analysis of quantitative data and contributing to the writing of the final report.

Alex J Turner was involved in the analysis of quantitative data and contributing to the writing of the final report.

Matt Sutton took lead responsibility for the quantitative components of the study. He was involved in designing the research project, analysing data and contributing to the writing of the final report.

Publications

Kristensen SR, Meacock R, Turner AJ, Boaden R, McDonald R, Roland M, *et al.* Long-term effect of hospital pay for performance on outcomes in England. *N Eng J Med* 2014;**371**:540–8.

Meacock R, Kristensen S, Sutton M. The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. *Health Econ* 2013;**23**:1–13.

Sutton M, McDonald R, Roland M. Hospital pay for performance in England. *N Eng J Med* 2013;**368**:968–9.

Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Eng J Med* 2012;**367**:1821–8.

References

1. Paris V, Devaux M, Wei L. *Health Systems Institutional Characteristics: A Survey of 29 OECD Countries*. OECD Health Working Papers, No. 50. Paris: OECD Publishing; 2010.
2. Eijkenaar F. Pay for performance in health care an international overview of initiatives. *Med Care Res Rev* 2012;**69**:251–76. http://dx.doi.org/10.1177/1077558711432891
3. Mehrotra A, Damberg CL, Sorbero MES, Teleki SS. Pay for performance in the hospital setting: what is the state of the evidence? *Am J Med Qual* 2009;**24**:19–28. http://dx.doi.org/10.1177/1062860608326634
4. Grossbart SR. What's the return? Assessing the effect of 'pay-for-performance' initiatives on the quality of care delivery. *Med Care Res Rev* 2006;**63**:29S–48S. http://dx.doi.org/10.1177/1077558705283643
5. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, *et al.* Public reporting and pay for performance in hospital quality improvement. *N Engl J Med* 2007;**356**:486–96. http://dx.doi.org/10.1056/NEJMsa064964
6. Glickman SW, Ou F-S, DeLong ER, Roe MT, Lytle BL, Mulgund J, *et al.* Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA* 2007;**297**:2373–80. http://dx.doi.org/10.1001/jama.297.21.2373
7. Damberg C, Raube K, Teleki S, de la Cruz E. Taking stock of pay for performance: a candid assessment from the front lines. *Health Affairs* 2009;**28**:517–25. http://dx.doi.org/10.1377/hlthaff.28.2.517
8. Ryan AM, Blustein J. The effect of the masshealth hospital pay-for-performance program on quality. *Health Serv Res* 2011;**46**:712–28. http://dx.doi.org/10.1111/j.1475-6773.2010.01224.x
9. Ryan AM. Effects of the premier hospital quality incentive demonstration on medicare patient mortality and cost. *Health Serv Res* 2009;**44**:821–42. http://dx.doi.org/10.1111/j.1475-6773.2009.00956.x
10. Flodgren G, Eccles MP, Shepperd S, Scott A, Parmelli E, Beyer FR. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. *Cochrane Database Syst Rev* 2011;**7**:CD009255. http://dx.doi.org/10.1002/14651858.CD009255
11. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 2012;**367**:1821–8. http://dx.doi.org/10.1056/NEJMsa1114951
12. Middleton E, Baker D. Comparison of social distribution of immunisation with measles, mumps and rubella vaccine, England, 1991–2001. *BMJ* 2003;**326**:854. http://dx.doi.org/10.1136/bmj.326.7394.854
13. Smith P. On the unintended consequences of publishing performance data in the public sector. *Intl J Pub Admin* 1995;**18**:277–310. http://dx.doi.org/10.1080/01900699508525011
14. Davies C, Anand P, Artigas L, Holloway J, McConway K, Newman J, *et al. Links Between Governance Incentives and Outcomes: A Review of the Literature*. London: National Co-ordinating Centre for NHS Service Delivery and Organisation R&D (NCCSDO); 2004.
15. Christianson J, Leatherman S, Sutherland K. *Financial Incentives for Healthcare Providers and Quality Improvement: A Review of the Evidence*. London: Health Foundation; 2007.

16. McDonald R, Cheraghi-Sohi S, Tickle M, Roland M, Doran T, Campbell S *et al. The Impact of Incentives on the Behaviour and Performance of Primary Care Professionals*. London: Service Delivery Organisation; 2010.

17. Berenson R, Rich E. US Approaches to physician payment: the deconstruction of primary care. *J Gen Intern Med* 2010;**25**:613–18. http://dx.doi.org/10.1007/s11606-010-1295-z

18. Emanuel E, Fuchs V. The perfect storm of overutilization. *JAMA* 2008;**299**:2789–91. http://dx.doi.org/10.1001/jama.299.23.2789

19. Kerr E, Mittman B, Hays R, Leake B, Brook R, *et al.* Quality assurance in capitated physician groups. Where is the emphasis? *JAMA* 1996;**276**:1236–9. http://dx.doi.org/10.1001/jama.1996.03540150038028

20. Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal M, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 2010;**247**:1–13. http://dx.doi.org/10.1186/1472-6963-10-247

21. Roland M. Linking physician pay to quality of care: a major experiment in the UK. *N Engl J Med* 2004;**351**:1448–54. http://dx.doi.org/10.1056/NEJMhpr041294

22. Kristensen SR, McDonald R, Sutton M. Should pay-for-performance schemes be locally designed? Evidence from the Commissioning for Quality and Innovation (CQUIN) Framework. *J Health Serv Res Policy* 2013;**18**:38–49. http://dx.doi.org/10.1177/1355819613490148

23. Rowe J. Pay-for-performance and accountability: related themes in improving health care. *Ann Intern Med* 2006;**145**:695–9. http://dx.doi.org/10.7326/0003-4819-145-9-200611070-00013

24. Williams T, Raube K, Damberg C, Mardon R. Pay for performance: its influence on the use of IT in physician organizations. *J Med Pract Manage* 2006;**21**:301–6.

25. McDonald R, Zaidi S, Todd S, Konteh F, Hussain K, Roe J, *et al. A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs*. Nottingham: University of Nottingham; 2012.

26. McDonald R, Zaidi S, Todd S, Konteh F, Hussain K, Brown S, *et al. Evaluation of the Commissioning for Quality and Innovation Framework Final Report*. Nottingham: University of Nottingham; 2013.

27. McDonald R, White J, Marmor T. Paying for performance in primary medical care: learning about and learning from 'success' and 'failure' in England and California. *J Health Polit Policy Law* 2009;**34**:747–76. http://dx.doi.org/10.1215/03616878-2009-024

28. Sautter K, Bokhour BG, White B, Young GJ, Burgess JF, Berlowitz D, *et al.* The early experience of a hospital-based pay-for-performance program. *J Healthcare Manag* 2007;**52**:95–107.

29. Deci E. Effects of externally mediated rewards on intrinsic motivation. *J Pers Soc Psychol* 1971;**18**:105–15. http://dx.doi.org/10.1037/h0030644

30. Productivity Commission. *Behavioural Economics and Public Policy: Roundtable Proceedings*. Canberra: Australian Government Productivity Commission; 2008.

31. Wynia M. The risks of rewards in health care: how pay-for-performance could threaten, or bolster, medical professionalism. *J Gen Intern Med* 2009;**24**:884–7. http://dx.doi.org/10.1007/s11606-009-0984-y

32. Gaynor M, Gertner P. Moral hazard and risk spreading in partnerships. *RAND J Econ* 1995;**26**:591–613. http://dx.doi.org/10.2307/2556008

33. Rosenthal M, Dudley R. Pay-for-performance: will the latest payment trend improve care? *JAMA* 2007;**297**:740–4. http://dx.doi.org/10.1001/jama.297.7.740

34. Landon B, Normand S, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. *JAMA* 2003;**290**:1183–9. http://dx.doi.org/10.1001/jama.290.9.1183

35. Kuhn M. *Quality in Primary Care: Economic Approaches to Analysing Quality-Related Physician Behavior*. London: Office of Health Economics; 2003.

36. Boyd C, Darer J, Boult C, Fried LP, Boult L, Wu A, *et al.* Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *JAMA* 2005;**294**:716–24. http://dx.doi.org/10.1001/jama.294.6.716

37. Ratto M, Propper C, Burgess S. Using financial incentives to promote teamwork in health care. *J Health Serv Res Policy* 2002;**7**:69–70. http://dx.doi.org/10.1258/1355819021927683

38. Hernández-Quevedo C, Llano R, Mossialos E. Paying for integrated care: an overview. *Eurohealth Observer* 2013;**19**:3–6.

39. Werner R, Goldman L, Dudley R. Comparison of change in quality of care between safety-net and non-safety-net hospitals. *JAMA* 2008;**299**:2180–7. http://dx.doi.org/10.1001/jama.299.18.2180

40. Gilmore A, Zhao Y, Kang N, Ryskina K, Legorreta A, Taira D, *et al.* Patient outcomes and evidence-based medicine in a preferred provider organization setting: a six-year evaluation of a physician pay-for-performance program. *Health Serv Res* 2007;**42**:2140–59. http://dx.doi.org/10.1111/j.1475-6773.2007.00725.x

41. Rosenthal M, Frank R, Li Z, Epstein A. Early experience with pay for performance: from concept to practice. *JAMA* 2005;**294**:1788–93. http://dx.doi.org/10.1001/jama.294.14.1788

42. Campbell S, Reeves D, Kontopantelis E, Sibbald B, Roland M, *et al.* Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009;**361**:368–78. http://dx.doi.org/10.1056/NEJMsa0807651

43. Hasnain-Wynia R, Baker D, Nerenz D, Feinglass J, Beal A, Landrum M, *et al.* Disparities in health care are driven by where minority patients seek care: examination of the hospital quality alliance measures. *Arch Intern Med* 2007;**167**:1233–9. http://dx.doi.org/10.1001/archinte.167.12.1233

44. Deci E, Koestner R, Ryan R. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol Bull* 1999;**125**:627–68. http://dx.doi.org/10.1037/0033-2909.125.6.627

45. Prendergast C. The provision of incentives in firms. *J Econ Lit* 1999;**37**:7–63. http://dx.doi.org/10.1257/jel.37.1.7

46. Pope G. Overview of Pay for Performance Models and Issues. In Cromwell J, Trisolini MG, Pope GC, Mitchell JB, Greenwald LM, editors. *Pay for Performance in Health Care: Methods and Approaches*. Research Triangle Park, NC: RTI Press; 2011. pp. 33–75.

47. Bhattacharyya T, Freiberg A, Mehta P, Katz J, Ferris T. Measuring the report card: the validity of pay-for-performance metrics in orthopedic surgery. *Health Aff (Millwood)* 2009;**28**:526–32. http://dx.doi.org/10.1377/hlthaff.28.2.526

48. Kautter J, Pope G, Trisolini M, Grund S. Medicare physician group practice demonstration design: quality and efficiency pay-for-performance. *Health Care Financ Rev* 2007;**29**:15–29.

49. Tversky A, Kahneman D. Loss aversion in riskless choice: a reference dependent model. *Q J Econ* 1991;**107**:1039–61. http://dx.doi.org/10.2307/2937956

50. Ariely D, Huber J, Wertenbroch K. When do losses loom larger than gains? *J Mark Res* 2005;**42**:134–8. http://dx.doi.org/10.1509/jmkr.42.2.134.62283

51. Opelka F, Brown C. Understanding pay for performance. *Bull Am Coll Surg* 2005;**90**:12–17.

52. Beaulieu N, Horrigan D. Putting smart money to work for quality improvement. *Health Serv Res* 2005;**40**:1318–34. http://dx.doi.org/10.1111/j.1475-6773.2005.00414.x

53. Pelonero A, Johnson R. Economic grand rounds: a pay-for-performance program for behavioral health care practitioners. *Psychiat Serv* 2007;**58**:442–4. http://dx.doi.org/10.1176/ps.2007.58.4.442

54. Cromwell J. Financial Gains and Risks in Pay for Performance Bonus Algorithms. In Cromwell J, Trisolini MG, Pope GC, Mitchell JB, Greenwald LM, editors. *Pay for Performance in Health Care: Methods and Approaches*. Research Triangle Park, NC: RTI Press; 2011. http://dx.doi.org/10.3768/rtipress.2011.bk.0002.1103

55. Department of Health. *Using the Commissioning for Quality and Innovation (CQUIN) payment framework: For the NHS in England 2009/10*. London: Department of Health; 2008.

56. Zaslavsky A, Hochheimer JN, Schneider EC, Cleary PD, Seidman JJ, McGlynn EA, *et al.* Impact of sociodemographic case mix on the HEDIS measures of health plan quality. *Med Care* 2000;**38**:981–92. http://dx.doi.org/10.1097/00005650-200010000-00002

57. Casalino L, Elster A, Eisenberg A, Lewis E, Montgomery J, Ramos D. Will pay-for-performance and quality reporting affect health care disparities? *Health Aff (Millwood)* 2007;**26**:w405–14. http://dx.doi.org/10.1377/hlthaff.26.3.w405

58. Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. *N Engl J Med* 2008;**359**:274–84. http://dx.doi.org/10.1056/NEJMsa0800310

59. Gravelle H, Sutton M, Ma A. Doctor behaviour under a pay for performance contract: treating, cheating and case finding? *Econ J* 2010;**120**:F129–56. http://dx.doi.org/10.1111/j.1468-0297.2009.02340.x

60. Casalino L, Alexander G, Jin L, Konetzka R. General internists' views on pay-for-performance and public reporting of quality scores: a national survey. *Health Aff (Millwood)* 2007;**26**:492–9. http://dx.doi.org/10.1377/hlthaff.26.2.492

61. Wodchis W, Ross J, Detsky A. Is P4P really FFS? *JAMA* 2007;**298**:1797–9. http://dx.doi.org/10.1001/jama.298.15.1797

62. de Bruin S, Baan C, Struijs J. Pay-for-performance in disease management: a systematic review of the literature. *BMC Health Serv Res* 2011;**11**:272. http://dx.doi.org/10.1186/1472-6963-11-272

63. Sutton M, Elder R, Guthrie B, Watt G. Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. *Health Econ* 2010;**19**:1–13. http://dx.doi.org/10.1002/hec.1440

64. Fairbrother G, Hanson K, Friedman S, Butts G. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. *Ambul Pediatr* 2001;**1**:206–12. http://dx.doi.org/10.1367/1539-4409(2001)001<0206:IOFIOD>2.0.CO;2

65. Stulberg J. The Physician Quality Reporting Initiative: a gateway to pay for performance: what every health care professional should know. *Qual Manag Health Care* 2008;**17**:2–8. http://dx.doi.org/10.1097/01.QMH.0000308632.74355.93

66. Institute of Medicine. *Performance Measurement: Accelerating Improvement (Vol. 11517)*. Washington, DC: The National Academies Press; 2006.

67. Topol E, Califf R. Scorecard cardiovascular medicine: its impact and future directions. *Ann Intern Med* 1994;**120**:65–70. http://dx.doi.org/10.7326/0003-4819-120-1-199401010-00011

68. Trisolini M. Introduction to Pay for Performance. In Cromwell J, Trisolini MG, Pope GC, Mitchell JB, Greenwald LM, editors. *Pay for Performance in Health Care: Methods and Approaches*. Research Triangle Park, NC: RTI Press; 2011.

69. Mehrotra A, Bodenheimer T, Dudley R. Employers' efforts to measure and improve hospital quality: determinants of success. *Health Aff (Millwood)* 2003;**22**:60–71. http://dx.doi.org/10.1377/hlthaff.22.2.60

70. Gagné M, Deci E. Self-determination theory and work motivation. *J Organ Behav* 2005;**26**:331–62. http://dx.doi.org/10.1002/job.322

71. Kluger A, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull* 1996;**119**:254–84. http://dx.doi.org/10.1037/0033-2909.119.2.254

72. Conrad D, Perry L. Quality-based financial incentives in health care: can we improve quality by paying for it? *Annu Rev Publ Health* 2009;**30**:357–71. http://dx.doi.org/10.1146/annurev.publhealth.031308.100243

73. NHS England. *Review of Incentives, Rewards and Sanctions: Discussion Paper for Stakeholders*, London: NHS England/Commissioning Policy and Resources; 2013.

74. Center for Health Care Strategies, Inc. *Physician Pay-for-Performance in Medicaid: A Guide for States*. Hamilton, NJ: Center for Health Care Strategies; 2007.

75. Kahn C, Ault T, Isenstein H, Potetz L, Van Gelder S. Snapshot of hospital quality reporting and pay-for-performance under Medicare. *Health Aff (Millwood)* 2006;**25**:148–62. http://dx.doi.org/10.1377/hlthaff.25.1.148

76. Cromwell J, Smith K. Evaluating Pay for Performance Interventions. In Cromwell J, Trisolini MG, Pope GC, Mitchell JB, Greenwald LM, editors. *Pay for Performance in Health Care: Methods and Approaches*. Research Triangle Park, NC: RTI Press: 2011. http://dx.doi.org/10.3768/rtipress.2011.bk.0002.1103

77. RTI International. *Evaluation of the Premier Hospital Quality Incentive Demonstration*. Research Triangle Park, NC: Premier, Inc.; 2011.

78. Hibbard J, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood)* 2003;**22**:84–94. http://dx.doi.org/10.1377/hlthaff.22.2.84

79. Reeves D, Doran T, Valderas JM, Kontopantelis E, Trueman P, Sutton M, *et al.* How to identify when a performance indicator has run its course. *BMJ* 2010;**340**:899–901. http://dx.doi.org/10.1136/bmj.c1717

80. Centers for Medicare & Medicaid Services. *Evaluation of the Premier Hospital Quality Incentive Demonstration – Executive Summary: Impacts on Quality of Care, Medicare Reimbursements, and Medicare Beneficiaries' Length of Stay during the First Three Years of the Demonstration*. Baltimore, MD: Centers for Medicare & Medicaid Services; 2009.

81. Giuffrida A, Gravelle H, Roland M. Measuring quality of care with routine data: avoiding confusion between performance indicators and health outcomes. *BMJ* 1999;**319**:94–8. http://dx.doi.org/10.1136/bmj.319.7202.94

82. Department of Health. *Developing the Quality and Outcomes Framework: Proposal For a New, Independent Process*. London: Department of Health; 2008.

83. Isles V, Sutherland K. *Managing Change in the NHS: Organisational Change, a Review for Health Care Managers, Professionals and Researchers*. London: Service Delivery Organisation; 2001.

84. Bate P, Mendel P, Robert G. *Organizing for Quality. The Improvement Journeys of Leading Hospitals in Europe and the United States*. Oxford: Radcliffe; 2008. URL: www.radcliffehealth.com/shop/organizing-quality-improvement-journeys-leading-hospitals-europe-and-united-states

85. Bourgeois L. On the measurement of organizational slack. *Acad Manag Rev* 1981;**6**:29–39. http://dx.doi.org/10.2307/257138

86. Mintzberg H. *The Structuring of Organizations: A Synthesis of the Research*. Englewood Cliffs, NJ: Prentice-Hall; 1979.

87. Haas PM. Introduction: epistemic communities and international policy co-ordination. *Int Organ* 1992;**46**:1–36. http://dx.doi.org/10.1017/S0020818300001442

88. Burt RS. *Structural Holes: the Social Structure of Competition*. Cambridge, MA: Harvard University Press; 1992.

89. Burt RS. The Network Structure of Social Capital. In Staw BM, Sutton RI, editors. *Research in Organizational Behavior* 2000;**22**:345–431.

90. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. *Milbank Q* 2004;**82**:581–629. http://dx.doi.org/10.1111/j.0887-378X.2004.00325.x

91. Pawson, R. *The Science of Evaluation: A Realist Manifesto*. London: Sage; 2013. http://dx.doi.org/10.4135/9781473913820

92. Greenhalgh T, Humphrey C, Hughes J, Macfarlane F, Butler C, Pawson R. How do you modernize a health service? A realist evaluation of whole-scale transformation in London. *Milbank Q* 2009;**87**:391–416. http://dx.doi.org/10.1111/j.1468-0009.2009.00562.x

93. Fligstein, N. Organizational Fields. In Bevir M, editors. *Encyclopaedia of Governance*. New York, NY: Sage; 2007. http://dx.doi.org/10.4135/9781412952613.n371

94. Beckert, J. How do fields change? *Organ Stud* 2010;**31**:605–27. http://dx.doi.org/10.1177/0170840610372184

95. Powell, WW, DiMaggio PJ. *The New Institutionalism in Organizational Analysis*. Chicago, IL: University of Chicago Press; 1992.

96. Reay T, Hinings CR. Managing the rivalry of competing institutional logics. *Organ Stud* 2009;**30**:e629–52. http://dx.doi.org/10.1177/0170840609104803

97. Granovetter M. Economic institutions as social constructions: a framework for analysis. *Acta Sociol* 1992;**35**:3–11. http://dx.doi.org/10.1177/000169939203500101

98. Fuhse JA. The Meaning Structure of Social Networks. *Sociol Theor* 2009;**27**:51–73. http://dx.doi.org/10.1111/j.1467-9558.2009.00338.x

99. Advancing Quality Alliance. *About AQuA*. Advancing Quality Alliance. URL: http://www.aquanw.nhs.uk (accessed 5 March 2015).

100. World Health Organization (WHO). *International Statistical Classification of Diseases and Related Health Problems*. 10th edn. Geneva; WHO; 2010.

101. Maynard A. The powers and pitfalls of payment for performance. *Health Econ* 2012;**21**:S3–12. http://dx.doi.org/10.1002/hec.1810

102. Pawson R, Tilley N. *Realistic Evaluation*. London: Sage; 1997.

103. Cyert RM, March JGA. *Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall; 1963.

104. Kingdon J. *Agendas, Alternatives and Public Policies*. New York, NY: HarperCollins; 1994.

105. DiMaggio PJ, Powell WW. The Iron Cage Revisited – Institutional Isomorphism and Collective Rationality in Organizational Fields. *Am Sociol Rev* 1983;**48**:147–60. http://dx.doi.org/10.2307/2095101

106. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. *JAMA* 1995;**274**:1800–4. http://dx.doi.org/10.1001/jama.1995.03530220066035

107. Chao G, O'Leary-Kelly A, Wolf S, Klein H, Gardner P. Organizational socialization: its content and consequences. *J Appl Psychol* 1994;**79**;730–43. http://dx.doi.org/10.1037/0021-9010.79.5.730

108. Schieffelin BB, Ochs E, editors. *Language Socialization across Cultures*. New York, NY: Cambridge University Press; 1986.

109. Fernandez R, Gould R. A dilemma of state power: brokerage and influence in the national health policy domain. *Am J Sociol* 1994;**99**:1455–91. http://dx.doi.org/10.1086/230451

110. Currie G, White L. Inter-professional barriers and knowledge brokering in an organizational context: the case of healthcare. *Organ Stud* 2012;**33**:1333–61. http://dx.doi.org/10.1177/0170840612457617

111. McDonald, R. Everything you wanted to know about anxiety but were afraid to ask. *J Health Serv Res Policy* 2008;**3**:249–50. http://dx.doi.org/10.1258/jhsrp.2008.008067

112. Checkland K, McDonald R, Harrison S. Ticking boxes and changing the social world: Data collection and the new UK general practice contract. *Soc Policy Admin* 2007;**41**:693–710. http://dx.doi.org/10.1111/j.1467-9515.2007.00580.x

113. Berwick D. Developing and testing changes in delivery of care. *Ann Int Med* 1998;**8**:651–6. http://dx.doi.org/10.7326/0003-4819-128-8-199804150-00009

114. Kilo C. A framework for collaborative improvement: lessons from the Institute for Healthcare Improvement's Breakthrough Series. *Qual Manag Health Care* 1998;**6**:1–13. http://dx.doi.org/10.1097/00019514-199806040-00001

115. Øvretveit J, Bate P, Cleary P, Cretin S, Gustafson D, McInnes K, *et al.* Quality collaboratives: Lessons from research. *Qual Saf Health Care* 2002;**11**:345–51. http://dx.doi.org/10.1136/qhc.11.4.345

116. Ashworth RE, Boyne GA, Delbridge R. Escape from the Iron Cage? Organizational change and isomorphic pressures in the public sector. *JPART* 2009;**19**:165–87.

117. Mannion R, Davies H. Will prescriptions for cultural change improve the NHS? *BMJ* 2013;**346**:f1305. http://dx.doi.org/10.1136/bmj.f1305

118. Joas H. *The Genesis of Values*. Chicago, IL: University of Chicago Press; 2000.

119. Nadeem E, Olin S, Campbell Hill L, Eaton Hoagwood K, McCue Horwitz S. Understanding the components of quality improvement collaboratives: a systematic literature review. *Milbank Q* 2013;**91**:354–94. http://dx.doi.org/10.1111/milq.12016

120. Tsoukas H, Chia R. On Organizational becoming: rethinking organizational change. *Organ Sci* 2002;**13**:567–82. http://dx.doi.org/10.1287/orsc.13.5.567.7810

121. Ingram P, Roberts PW. Friendship among competitors in the Sydney hotel industry. *Am J Sociol* 2000;**106**:387–423. http://dx.doi.org/10.1086/316965

Appendix 1 Advancing Quality measures

Community-acquired pneumonia

1. Percentage of patients who received an oxygenation assessment within 24 hours prior to or after hospital arrival.
2. Initial antibiotic selection.
3. Blood culture collected prior to first antibiotic administration.
4. Antibiotic timing: the percentage of pneumonia patients who received their first dose of antibiotics within 6 hours after hospital arrival.
5. Smoking cessation advice/counselling.
6. CURB-65 score (in effect from April 2011).

Hip and knee replacement

1. Prophylactic antibiotic received within 1 hour prior to surgical incision.
2. Prophylactic antibiotic selection for surgical patients.
3. Prophylactic antibiotics discontinued within 24 hours after surgery end time.
4. Recommended venous thromboembolism prophylaxis ordered.
5. Appropriate venous thromboembolism prophylaxis within 24 hours prior to surgery to 24 hours after surgery.
6. Readmission rate.

Acute myocardial infarction

1. Aspirin at arrival.
2. Aspirin prescribed at discharge.
3. Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for left ventricular systolic dysfunction.
4. Smoking cessation advice/counselling.
5. Beta-blocker at arrival (withdrawn with effect from 1 July 2009).
6. Beta-blocker prescribed at discharge.
7. Thrombolytic received within 30 minutes of hospital arrival.
8. Percutaneous coronary intervention received within 90 minutes of hospital arrival.
9. Inpatient mortality rate.

Coronary artery bypass graft

1. Aspirin prescribed at discharge.
2. Prophylactic antibiotic received within 1 hour prior to surgical incision.
3. Prophylactic antibiotic selection for surgical patients.
4. Prophylactic antibiotics discontinued within 48 hours after surgery end time.
5. Inpatient mortality rate.

Heart failure

1. Left ventricular systolic assessment.
2. Detailed discharge instructions.
3. Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for left ventricular systolic dysfunction.
4. Smoking cessation advice/counselling.

Appendix 2 Changes in trust performance over the first 12 months of Advancing Quality

As outlined in the report, the AQ programme aims to achieve better health outcomes for patients in five domains: AMI, pneumonia, hip and knee replacements, heart failure and CABG. Each of the five clinical areas that AQ currently focuses on encompasses key activities (process and outcome measures) that should happen to all patients. These measures have equal weighting.

A CQS is calculated by looking at the number of opportunities for delivering the AQ processes and outcomes and the number of times these were delivered. For example, if there were four measures and five patients, then there are 20 opportunities that should be met. If all five patients received all four measures, then 20 out of 20 is achieved; equating to a CQS of 100%. If one of the five patients only received two of the process measures, and another patient received only three, then 15 out of 20 is achieved; equating to a CQS of 75%.

The composite scores for these five domains are derived from a total of 26 indicators.

We obtained the indicator-specific quarterly performance data for each trust on these 26 indicators for each quarter of year 1 of AQ. This enabled us to compare changes over time and between trusts.

We used the indicator-specific quarterly data to examine:

- the distribution of bonus payments across trusts
- the distribution of domain CQS across trusts over time
- transitions of trusts between quartiles of performance on each domain
- how changes in achievement between quarters relate to baseline achievement
- changes in performance on individual indicators
- how differences in performance on individual indicators relate to differences in CQS.

Findings

We found differences between trusts in terms of their responses to challenges and their overall approach to AQ implementation. We also observed changes over time within trusts as participants identified problems and modified systems and behaviours in an attempt to overcome deficiencies. These findings from our qualitative analysis are consistent with the results of our quantitative analyses.

Performance and distribution of bonus payments across trusts

There is a risk with schemes that reward only top performance that improvement will be concentrated among participants who are likely to achieve the bonus. This did not occur in the first year of the AQ programme. From quarter to quarter, the largest improvements in performance tended to be observed among the trusts that did not receive a bonus in the previous quarter.

Across the four quarters one trust consistently delivered a bonus-earning performance. During the first year of the programme each of the 24 participating trusts received at least one bonus. In both of the first two quarters, three trusts received a bonus for their performance in each domain and three trusts received no

bonus. In the third quarter, four trusts received a bonus for their performance in each domain, while four trusts did not receive a bonus in any domain. In the fourth quarter, each trust received at least one bonus and two trusts were given a reward for their performance in each domain.

In most domains there has been considerable mobility of trusts between quartiles of achievement. However, the mobility of trusts between quartiles on the heart failure domain is relatively low. Trusts ranked in the lowest quartile did not cross the median cut-off point and receive a bonus. Trusts that ranked in the top quartile did not slip in performance below the median in any of the quarters.

Appendix 3 Changes in the distribution of composite quality score over time

Table 11 shows how the values of the CQS at the 50th and 75th percentiles have evolved over the four quarters in each domain. These are the values at which trusts receive the 2% and 4% revenue bonuses, respectively. These cut-off values have risen for all domains, most noticeably for heart failure. The 50th percentile value for pneumonia has increased by just over 1%.

In most domains we did not observe large differences in the CQS between the trusts just above and just below the bonus-earning thresholds. Only in the case of heart failure are there substantial differences in performance that determine which individual trusts earn bonuses.

Some participants suggested that patient numbers would influence performance against AQ measures and that it may not be appropriate therefore to compare trusts without taking this into account. We studied the variability in quality estimates as a function of the number of patients eligible for treatment to investigate whether or not there was a relationship between this and the number of patients treated at a participating provider but found that patient volumes did not account for the heterogeneity of the outcomes in any of the clinical domains.

TABLE 11 Composite quality scores at the bonus cut-offs, by domain and quarter

Quarter	50th percentile	Difference at 50th percentile	75th percentile	Difference at 75th percentile
AMI				
1	93.181	1.186	96.112	0.742
2	92.024	0.078	97.103	0.326
3	94.706	0.165	97.422	0.785
4	96.183	1.597	98.763	0.005
Hip and knee				
1	89.908	0.781	94.682	0.097
2	93.896	1.169	96.986	0.247
3	95.350	0.389	97.304	0.138
4	93.782	0.761	96.145	0.145
Heart failure				
1	57.831	0.339	72.519	4.296
2	55.301	3.460	74.154	4.152
3	62.198	1.319	78.372	0.293
4	65.850	0.375	77.963	5.049
Pneumonia				
1	78.555	0.589	80.912	1.110
2	77.049	1.298	82.432	0.963
3	77.500	0.634	83.582	0.436
Q4	79.775	1.271	85.158	0.124

Changes in performance on individual indicators

The most substantial changes between the first and fourth quarters of the first year are improvements in the provision of smoking cessation advice (*Table 12*).

TABLE 12 Average quarter 1 and quarter 4 performance on each indicator

Indicator	Q1	Q4	Q4–Q1
AMI			
Aspirin at arrival	95.1	97.0	1.8
Aspirin at discharge	98.0	97.6	–0.5
Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for left ventricular systolic dysfunction	94.3	97.6	3.3
Smoking cessation advice	73.6	86.1	12.5
Beta-blocker at discharge	91.2	93.7	2.6
Beta-blocker at arrival	76.6	78.9	2.3
Fibrinolytic therapy	78.2	77.1	–1.0
Hip and knee replacement			
Antibiotic 1 hour prior to incision	76.9	83.5	6.5
Antibiotic selection for surgical patients	85.4	88.1	2.7
Antibiotic discontinued within 24 hours post surgery	89.7	95.5	5.8
Venous thromboembolism prophylaxis	85.9	94.9	9.0
Received appropriate venous thromboembolism prophylaxis	82.9	90.9	8.0
Heart failure			
Evaluation of left ventricular systolic function	86.4	89.8	3.4
Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for left ventricular systolic dysfunction	91.2	93.3	2.2
Discharge instructions	22.8	34.0	11.2
Adult smoking cessation advice	42.5	68.3	25.8
Pneumonia			
Oxygenation assessment	94.8	98.8	4.0
Initial antibiotic selection for immunocompetent patients	81.7	83.7	2.0
Blood cultures performed in accident and emergency	60.9	61.3	0.3
Initial antibiotic received within 6 hours	63.8	65.3	1.4
Adult smoking cessation advice/counselling	32.4	43.3	11.0

How differences in performance on individual indicators relate to the composite quality score

In the AMI domain, achievements on the indicators aspirin at arrival, 'aspirin prescribed at discharge', and 'angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for left ventricular systolic dysfunction' are very similar and do not distinguish between low- and high-performing trusts. In the hip and knee replacement domain the top-performing trusts achieved the highest mean value on all indicators and the trusts in the bottom quartile had the lowest attainment on all indicators. In the heart failure domain the 'discharge instructions' and 'adult smoking cessation advice' indicators are the most discriminating indicators separating the trusts that performed worst and best on the CQS. In the pneumonia domain the 'oxygenation assessment' and (by the fourth quarter) 'adult smoking cessation' indicators do not differentiate between trusts with low and high CQSs.

With regard to discharge instructions for heart failure patients, a number of providers had adopted a strategy of producing leaflets containing these instructions. This may account in part for the gap between high- and low-performing trusts in relation to this indicator. If this is the case, we would expect this gap to close over time as more trusts adopt this approach.

Appendix 4 Sensitivity analysis

To confirm the robustness of the results of our main analysis, we undertook a number of sensitivity analyses which we report in the following sections.

First, we verified that the conclusions of our main analysis were not sensitive to our use of observations weights or choice of variance estimator.

Second, we used pre-trends tests to confirm that hospitals in the north-west of England did not have a different trend to those in the rest of England prior to the introduction of the programme.

We also repeated the main analysis from the paper but using total (in-hospital and out-of-hospital) mortality rates rather than in-hospital mortality rates only and confirmed that the results remained stable to using total mortality rates.

We verified that our results were not generated by regression towards the mean by including baseline mortality instead of hospital fixed effects.

We showed that our results were unaffected by the removal of the small group of hospitals that introduced financial incentives for the incentivised or non-incentivised conditions in the long term.

We confirmed that our results remain stable when analysing 90-day in-hospital mortality rates rather than 30-day in-hospital mortality rates.

Effects of weighting and clustering on standard errors

TABLE 13 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England, with different weight and standard-error specifications

	Unweighted regressions				Weighted[a] regressions			
		95% CI				95% CI		
		Standard errors				Standard errors		
Incentivisation and conditions	Coefficient	Ordinary least squares	Robust[b]	Cluster[c]	Coefficient	Ordinary least squares	Robust[b]	Cluster[c]
Non-incentivised conditions								
Change from before to after P4P (short term)	0.38	–0.63 to 1.38	–0.66 to 1.41	–0.79 to 1.55	0.70	–0.06 to 1.46	–0.19 to 1.58	–0.28 to 1.68
Change from short term to long term	–1.25	–2.19 to –0.30	–2.14 to –0.35	–2.31 to –0.18	–1.21	–1.97 to –0.46	–1.98 to –0.44	–2.29 to –0.13
Change from before to after P4P (long term)	–0.87	–1.81 to 0.07	–1.84 to 0.11	–2.08 to 0.34	–0.51	–1.27 to 0.24	–1.36 to 0.33	–1.72 to 0.69
Incentivised conditions (total)								
Change from before to after P4P (short term)	–0.79	–1.36 to –0.22	–1.33 to –0.25	–1.38 to –0.19	–0.86	–1.29 to –0.42	–1.34 to –0.37	–1.45 to –0.27
Change from short term to long term	0.71	0.18 to 1.25	0.24 to 1.18	0.05 to 1.38	0.71	0.27 to 1.15	0.27 to 1.16	0.04 to 1.38
Change from before to after P4P (long term)	–0.08	–0.61 to 0.46	–0.60 to 0.45	–0.85 to 0.70	–0.15	–0.58 to 0.29	–0.61 to 0.31	–0.96 to 0.67
AMI								
Change from before to after P4P (short term)	–0.36	–1.24 to 0.53	–1.27 to 0.55	–1.38 to 0.67	–0.13	–0.82 to 0.56	–0.85 to 0.59	–1.11 to 0.86
Change from short term to long term	0.67	–0.16 to 1.50	–0.05 to 1.39	–0.37 to 1.71	0.35	–0.34 to 1.05	–0.30 to 1.00	–0.58 to 1.28
Change from before to after P4P (long term)	0.31	–0.52 to 1.14	–0.61 to 1.23	–0.95 to 1.58	0.22	–0.47 to 0.92	–0.50 to 0.95	–0.94 to 1.38

Incentivisation and conditions	Unweighted regressions				Weighted[a] regressions			
		95% CI				95% CI		
		Standard errors				Standard errors		
	Coefficient	Ordinary least squares	Robust[b]	Cluster[c]	Coefficient	Ordinary least squares	Robust[b]	Cluster[c]
Heart failure								
Change from before to after P4P (short term)	–0.18	–1.24 to 0.89	–1.13 to 0.78	–1.12 to 0.77	–0.20	–1.06 to 0.66	–1.06 to 0.66	–1.03 to 0.63
Change from short term to long term	0.34	–0.66 to 1.34	–0.55 to 1.23	–0.73 to 1.41	0.17	–0.71 to 1.04	–0.65 to 0.99	–0.77 to 1.10
Change from before to after P4P (long term)	0.17	–0.84 to 1.17	–0.77 to 1.10	–1.03 to 1.36	–0.03	–0.90 to 0.84	–0.86 to 0.80	–1.14 to 1.07
Pneumonia								
Change from before to after P4P (short term)	–1.83	–2.83 to –0.83	–2.73 to –0.93	–2.86 to –0.80	–1.52	–2.24 to –0.81	–2.30 to –0.75	–2.46 to –0.59
Change from short term to long term	1.12	0.19 to 2.05	0.31 to 1.93	–0.18 to 2.42	1.14	0.41 to 1.86	0.43 to 1.84	0.02 to 2.25
Change from before to after P4P (long term)	–0.71	–1.64 to 0.22	–1.55 to 0.13	–2.08 to 0.65	–0.39	–1.11 to 0.33	–1.10 to 0.33	–1.70 to 0.92

a The weight of each hospital–quarter–condition observation is the product of the share of each condition of all incentivised and control conditions and the share of each hospitals admissions in a given quarter of all admissions for that condition in the time periods before the introduction of AQ and in the short and long term.
b Uses the Huber/White/sandwich variance estimator robust to unspecified heteroscedasticity.
c Uses the clustered sandwich estimator robust to unspecified heteroscedasticity and within-hospital variation not captured by the hospital fixed effects.

Pre-trends tests

We tested whether or not the risk-adjusted mortality rates in the North West region had a different trend to those in the rest of England prior to the introduction of the programme using a pre-trends test.

Using data from before the introduction of the programme, we estimated the following regression model for each condition:

$$y_{jt} = \alpha_5 + u_j + \beta \times t + \rho \times G_j \times t + \varepsilon_{jt}; \quad (7)$$

in which t represents the quarter since the start of the data period, β is an estimate of the quarterly trend in the rest of England and ρ is the difference between the quarterly trend in the North West region of England and the quarterly trend in the rest of England.

We were able to accept the null hypothesis of equal pre-trends for each condition. The estimated values for ρ were as follows:

- AMI, –0.34 (95% CI –0.98 to 0.29)
- heart failure, 0.19 (95% CI –0.52 to 0.90)
- pneumonia, –0.23 (95% CI –0.81 to 0.36)
- non-incentivised conditions, –0.66 (95% CI –1.40 to 0.09).

Analysis using total (in-hospital and out-of-hospital) mortality rates

TABLE 14 Risk-adjusted total mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England

	North West region		Rest of England		Between-region difference in differences	
Incentivisation and conditions	Rate	Change	Rate	Change	Estimate	95% CI
Non-incentivised conditions						
Mortality before introduction	14.9	–	13.1	–	–	–
Change from before to short term	–	–0.6		–1.1	0.6	–0.3 to 1.6
Mortality after introduction (short term)	14.3	–	12	–	–	–
Change from short term to long term	–	–4.9		–3.7	–1.3	–2.3 to –0.2
Mortality after introduction (long term)	9.4	–	8.3	–	–	–
Change from before to long term	–	–5.5	–	–4.8	–0.6	–1.8 to 0.6
Incentivised conditions combined						
Mortality before introduction	21.1	–	19.4	–	–	–
Change from before to short term	–	–1.5	–	–0.8	–0.7	–1.3 to –0.1
Mortality after introduction (short term)	19.6	–	18.6	–	–	–
Change from short term to long term	–	–4.6	–	–5.3	0.6	–0.2 to 1.4
Mortality after introduction (long term)	15	–	13.3	–	–	–
Change from before to long term	–	–6.1	–	–6.1	–0.1	–1.1 to 0.9

TABLE 14 Risk-adjusted total mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England (*continued*)

	North West region		Rest of England		Between-region difference in differences	
Incentivisation and conditions	Rate	Change	Rate	Change	Estimate	95% CI
AMI						
Mortality before introduction	11.5	–	10.9	–	–	–
Change from before to short term	–	–1.2	–	–1.0	–0.1	–1.1 to 0.9
Mortality after introduction (short term)	10.3	–	9.9	–	–	–
Change from short term to long term	–	–2.2	–	–2.8	0.4	–0.5 to 1.3
Mortality after introduction (long term)	8.1	–	7.1	–	–	–
Change from before to long term	–	–3.4	–	–3.8	0.3	–0.8 to 1.4
Heart failure						
Mortality before introduction	17.7	–	15.9	–	–	–
Change from before to short term	–	–1.0	–	–0.9	–0.1	–0.9 to 0.8
Mortality after introduction (short term)	16.7	–	15	–	–	–
Change from short term to long term	–	–4.0	–	–3.8	0	–1.0 to 1.0
Mortality after introduction (long term)	12.7	–	11.2	–	–	–
Change from before to long term	–	–5.0	–	–4.7	–0.1	–1.4 to 1.1
Pneumonia						
Mortality before introduction	27.5	–	25.2	–	–	–
Change from before to short term	–	–1.9	–	–0.5	–1.3	–2.3 to –0.4
Mortality after introduction (short term)	25.6	–	24.7	–	–	–
Change from short term to long term	–	–5.9	–	–7.2	1	–0.1 to 2.1
Mortality after introduction (long term)	19.7	–	17.5	–	–	–
Change from before to long term	–	–7.8	–	–7.7	–0.4	–1.9 to 1.1

The total mortality rate includes both in- and out-of-hospital deaths, but the out-of-hospital mortality data are incomplete for the final 3 months of the study which explains the large decrease in mortality in the long term. The short-term period covers the first 18 months of the programme. The long-term period includes months 19–42 of the programme. The between-region difference in differences are the changes over time in the North West region minus the changes over time in the rest of England. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

Effects of including baseline mortality

TABLE 15 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England

Incentivisation and conditions	Between-region difference in differences		Between-region difference in differences with baseline		Triple differences		Triple differences with baseline	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Non-incentivised conditions								
Change from before to short term	0.7	–0.2 to 1.6	1.1	0.5 to 1.8	–	–	–	–
Change from short term to long term	–1.2	–2.0 to –0.4	–1.3	–2.1 to –0.5	–	–	–	–
Change from before to long term	–0.5	–1.4 to 0.3	–0.2	–0.7 to 0.4	–	–	–	–
Incentivised conditions combined								
Change from before to short term	–0.9	–1.3 to –0.4	–0.4	–0.7 to –0.0	–1.5	–2.6 to –0.5	–1.3	–2.0 to –0.5
Change from short term to long term	0.7	0.3 to 1.2	0.8	0.3 to 1.2	1.9	1.0 to 2.8	2.1	1.1 to 3.0
Change from before to long term	–0.1	–0.6 to 0.3	0.4	0.1 to 0.7	0.4	–0.6 to 1.3	0.8	0.2 to 1.4
AMI								
Change from before to short term	–0.1	–0.9 to 0.6	0	–0.5 to 0.6	–0.8	–2.0 to 0.3	–1	–1.8 to –0.2
Change from short term to long term	0.4	–0.3 to 1.0	0.4	–0.3 to 1.1	1.6	0.5 to 2.6	1.7	0.6 to 2.7
Change from before to long term	0.2	–0.5 to 0.9	0.4	–0.1 to 0.9	0.7	–0.4 to 1.8	0.6	–0.1 to 1.4
Heart failure								
Change from before to short term	–0.2	–1.1 to 0.7	0.4	–0.2 to 1.1	–0.9	–2.1 to 0.3	–0.5	–1.4 to 0.4
Change from short term to long term	0.2	–0.6 to 1.0	0.2	–0.7 to 1.1	1.4	0.2 to 2.5	1.5	0.3 to 2.7
Change from before to long term	0	–0.9 to 0.8	0.6	0.0 to 1.2	0.5	–0.7 to 1.7	1	0.2 to 1.8
Pneumonia								
Change from before to short term	–1.5	–2.3 to –0.7	–0.9	–1.5 to –0.4	–2.2	–3.4 to –1.0	–1.7	–2.6 to –0.9
Change from short term to long term	1.1	0.4 to 1.8	1.2	0.4 to 2.0	2.3	1.3 to 3.4	2.5	1.4 to 3.6
Change from before to long term	–0.4	–1.1 to 0.3	0.3	–0.3 to 0.8	0.1	–1.0 to 1.2	0.8	0.0 to 1.5

The short-term period covers the first 18 months of the programme. The long-term period includes months 19–42 of the programme. The between-region difference in differences are the changes over time in the North West region minus the changes over time in the rest of England. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

Analysis without hospitals with any financial incentives for the conditions incentivised by Advancing Quality or the non-incentivised conditions included in this paper

TABLE 16 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England, excluding hospitals with incentives for incentivised or non-incentivised conditions

	North West region		Rest of England		Between-region difference in differences	
Incentivisation and conditions	Rate	Change	Rate	Change	Estimate	95% CI
Non-incentivised conditions						
Mortality before introduction	13.9	–	11.7	–	–	–
Change from before to short term	–	–0.3	–	–0.6	0.7	–0.2 to 1.5
Mortality after introduction (short term)	13.6	–	11.1	–	–	–
Change from short term to long term	–	–2.5	–	–1.3	–1.2	–1.9 to –0.4
Mortality after introduction (long term)	11.1	–	9.8	–	–	–
Change from before to long term	–	–2.8	–	–1.9	–0.5	–1.4 to 0.3
Incentivised conditions combined						
Mortality before introduction	24.5	–	19.8	–	–	–
Change from before to short term	–	–1.1	–	0.1	–0.7	–1.2 to –0.2
Mortality after introduction (short term)	23.4	–	19.9	–	–	–
Change from short term to long term	–	–1.3	–	–1.7	0.6	0.2 to 1.0
Mortality after introduction (long term)	22.1	–	18.2	–	–	–
Change from before to long term	–	–2.4	–	–1.6	–0.1	–0.5 to 0.4
AMI						
Mortality before introduction	14.7	–	9.4	–	–	–
Change from before to short term	–	–0.7	–	–0.9	0.1	–0.6 to 0.8
Mortality after introduction (short term)	14	–	8.5	–	–	–
Change from short term to long term	–	–1.0	–	–0.9	0.1	–0.5 to 0.8
Mortality after introduction (long term)	13	–	7.6	–	–	–
Change from before to long term	–	–1.7	–	–1.8	0.2	–0.5 to 1.0
Heart failure						
Mortality before introduction	24.1	–	16.6	–	–	–
Change from before to short term	–	–0.4	–	–0.6	0	–0.9 to 0.8
Mortality after introduction (short term)	23.7	–	16	–	–	–
Change from short term to long term	–	–2.7	–	–1.4	0.1	–0.7 to 0.9
Mortality after introduction (long term)	21	–	14.6	–	–	–
Change from before to long term	–	–3.1	–	–2.0	0.1	–0.7 to 0.9

TABLE 16 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England, excluding hospitals with incentives for incentivised or non-incentivised conditions (*continued*)

Incentivisation and conditions	North West region		Rest of England		Between-region difference in differences	
	Rate	Change	Rate	Change	Estimate	95% CI
Pneumonia						
Mortality before introduction	26.4	–	24.1	–	–	–
Change from before to short term	–	–1.7	–	–0.4	–1.4	–2.2 to –0.6
Mortality after introduction (short term)	24.7	–	23.7	–	–	–
Change from short term to long term	–	–1.5		–2.7	0.9	0.3 to 1.6
Mortality after introduction (long term)	23.2	–	21	–	–	–
Change from before to long term	–	–3.2	–	–3.1	–0.5	–1.2 to 0.2

The short-term period covers the first 18 months of the programme. The long-term period includes months 19–42 of the programme. The between-region difference in differences are the changes over time in the North West region minus the changes over time in the rest of England. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

Effects of using 90-day mortality

TABLE 17 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England using 90-day in-hospital mortality

Incentivisation and conditions	North West region		Rest of England		Between-region difference in differences		Triple difference	
	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Non-incentivised conditions								
Mortality before introduction	17.9	–	15.8	–	–	–	–	–
Change from before to short term	–	–1.0	–	–1.5	0.6	–0.5 to 1.6	–	–
Mortality after introduction (short term)	16.9	–	14.3	–	–	–	–	–
Change from short term to long term	–	–3.6	–	–2.3	–1.3	–2.2 to –0.4	–	
Mortality after introduction (long term)	13.2	–	12.0	–	–	–	–	–
Change from before to long term	–	–4.7	–	–3.8	–0.7	–1.7 to 0.3	–	–

TABLE 17 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England using 90-day in-hospital mortality (*continued*)

Incentivisation and conditions	North West region		Rest of England		Between-region difference in differences		Triple difference	
	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Incentivised conditions combined								
Mortality before introduction	25.1	–	23.1	–	–	–	–	–
Change from before to short term	–	–2.0		–1.1	–1	–1.5 to –0.4	–1.5	–2.6 to –0.4
Mortality after introduction (short term)	23.1	–	21.9	–	–	–	–	–
Change from short term to long term	–	–2.8		–3.2	0.7	0.2 to 1.2	2	0.9 to 3.0
Mortality after introduction (long term)	20.4	–	18.7	–	–	–	–	–
Change from before to long term	–	–4.8	–	–4.3	–0.3	–0.8 to 0.2	0.4	–0.6 to 1.5
AMI								
Mortality before introduction	13.8	–	12.7	–	–	–	–	–
Change from before to short term	–	–1.3	–	–1.0	–0.1	–0.9 to 0.6	–0.7	–1.9 to 0.6
Mortality after introduction (short term)	12.5	–	11.7	–	–	–	–	–
Change from short term to long term	–	–1.7	–	–1.9	0.3	–0.4 to 1.1	1.6	0.4 to 2.8
Mortality after introduction (long term)	10.8	–	9.8	–	–	–	–	–
Change from before to long term	–	–3.0	–	–2.9	0.2	–0.6 to 1.0	0.9	–0.4 to 2.2
Heart failure								
Mortality before introduction	22.5	–	20.3	–	–	–	–	–
Change from before to short term	–	–1.6	–	–1.2	–0.4	–1.5 to 0.6	–1	–2.4 to 0.5
Mortality after introduction (short term)	20.9	–	19.1	–	–	–	–	–

TABLE 17 Risk-adjusted mortality for the conditions included in the P4P programme and those not included in the programme, before and after the introduction of the programme in the north-west of England using 90-day in-hospital mortality (*continued*)

Incentivisation and conditions	North West region		Rest of England		Between-region difference in differences		Triple difference	
	Rate	Change	Rate	Change	Estimate	95% CI	Estimate	95% CI
Change from short term to long term	–	–2.4	–	–2.6	0.3	–0.7 to 1.3	1.6	0.2 to 2.9
Mortality after introduction (long term)	18.5	–	16.5	–	–	–	–	–
Change from before to long term	–	–4.0	–	–3.8	–0.1	–1.2 to 0.9	0.6	–0.8 to 2.0
Pneumonia								
Mortality before introduction	32.2	–	29.6	–	–	–	–	–
Change from before to short term	–	–2.7		–1.1	–1.6	–2.4 to –0.8	–2.2	–3.5 to –0.9
Mortality after introduction (short term)	29.5	–	28.5	–	–	–	–	–
Change from short term to long term	–	–3.1		–4.2	1	0.3 to 1.8	2.3	1.1 to 3.5
Mortality after introduction (long term)	26.3	–	24.3	–	–	–	–	–
Change from before to long term	–	–5.9	–	–5.4	–0.6	–1.4 to 0.2	0.1	–1.1 to 1.4

The short-term period covers the first 18 months of the programme. The long-term period includes months 19–39 of the programme. The between-region difference in differences are the changes over time in the North West region minus the changes over time in the rest of England. The triple difference represents [(the change over time in mortality from the conditions incentivised in the North West region minus the change over time in mortality from the these conditions in the rest of England) minus (the change over time in mortality from the non-incentivised conditions in the North West region minus the change over time in mortality from the non-incentivised conditions in the rest of England)]. Estimates are from weighted least squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library