

PAY-FOR-PERFORMANCE: A LEVER FOR QUALITY IMPROVEMENT IN ACUTE HOSPITALS?



NICOLAS BOUCKAERT, JONATHAN STOKES, SØREN RUD KRISTENSEN, MARIJKE EYSSEN,
SOPHIE GERKENS, MEGGIE STAAL HØJGÅRD HANSEN, MÉLANIE LEFÈVRE, CARINE VAN DE
VOORDE

Title:	Pay-for-performance: a lever for quality improvement in acute hospitals?
Authors:	Nicolas Bouckaert (KCE), Jonathan Stokes (University of Glasgow, United Kingdom), Søren Rud Kristensen (University of Southern Denmark, Denmark), Marijke Eyssen (KCE), Sophie Gerkens (KCE), Meggie Staal Højgård Hansen (University of Southern Denmark, Denmark), Mélanie Lefèvre (KCE), Carine Van de Voorde (KCE)
Program manager:	Imgard Vinck (KCE)
Scientific director:	Roos Leroy (KCE)
Health Information Research Specialist:	Nicolas Fairon (KCE)
External experts:	Alice Benzimra (Direction Générale de l'Offre de Soins (DGOS), France), Stephen Duckett (University of Melbourne, Australia), Guy Durant (Cliniques universitaires Saint-Luc), Kristof Eeckloo (UZ Gent), Peter Fontaine (Europa Ziekenhuizen), Svin Deneckere (Vlaams Instituut voor Kwaliteit van Zorg (VIKZ)), Vera De Troyer (Zorgnet-Icuro), Donatienne Grégoire (Cliniques universitaires Saint-Luc), Denis Herbaux (Plateforme pour l'Amélioration continue de la Qualité des soins et de la Sécurité des patients (PAQS)), Didier Hoornaert (Clinique Saint-Pierre Ottignies), Grégoire Lefebvre (Centre Hospitalier de Mouscron), Annemie Lesage (FOD Volksgezondheid – SPF Santé publique), Mathieu Louiset (KU Leuven), Stéphan Mercier (ex-CHU HELORA), Alisha Lucas (Queensland Health, Australia), Samira Ouraghi (SPF Santé publique – FOD Volksgezondheid), Edward Norton (University of Michigan, United States), Johan Pauwels (RIZIV – INAMI), Quentin Schoonvaere (PAQS), Kris Vanhaecht (KU Leuven), Manfredi Ventura (Grand Hôpital de Charleroi), Joan Vlayen (Sint-Trudo Ziekenhuis)
Stakeholders:	Nicolas Böttcher (Santhea), Karin Cormann (Ministerium der Deutschsprachigen Gemeinschaft), Stephanie De Kempe (Zorgnet-Icuro), Vera De Troyer (Zorgnet-Icuro), Philippe Devos (UNESSA), Mieke Dolphens (Christelijke Mutualiteit), Anne Groswasser (Conférence des Hôpitaux Académiques de Belgique (CHAB)), Coralie Herijgers (Departement Zorg), Julie Peemans (Ligue des Usagers des Services de Santé (LUSS)), Geert Peuskens (Departement Zorg), Sabine Stordeur (SPF Santé publique – FOD Volksgezondheid), Laleh Taghipour Naghshi (CHAB)
External validators:	Jonas Brouwers (KU Leuven, UZ Leuven), Ricarda Milstein (OECD), Zeynep Or (IRDES, France)
Acknowledgements:	Yan Feng (Queen Mary University of London, United Kingdom), Sandrine Morin (Haute Autorité de Santé (HAS), France), Koen Van den Heede (KCE), Leen Verleye (KCE)
Reported interests KCE members:	KCE is a federal institution funded by INAMI/RIZIV, by the Federal Public Service of Health, Food Chain Safety and Environment, and by the Federal Public Service of Social Security. KCE's mission is to advise policymakers on decisions relating to health care and health insurance on the basis of scientific and objective research. It is expected to identify and shed light on the best possible solutions, in the context of an accessible, high-quality health care system with due regard for growing demand and budgetary constraints. KCE has no interest in companies (commercial or non-commercial, i.e. hospitals and universities), associations (e.g. professional associations, unions), individuals or organisations (e.g. lobby groups) that could be positively or negatively affected (financially or in any other way) by the implementation of the recommendations. All experts involved in the writing of the report or the peer-review process completed a declaration of interest form. Information on potential conflicts of interest is published in the colophon of this report. All



members of the KCE Expert Team make yearly declarations of interest and further details of these are available upon request.

Reported interests:

All experts and stakeholders consulted within this report were selected because of their involvement in the topic of P4P in hospitals. Therefore, by definition, each of them might have a certain degree of conflict of interest to the main topic of this report.

Layout:

Ine Verhulst

Disclaimer:

- Several external experts (cf. supra) were consulted about a (preliminary) version of the scientific report. They were selected based on their expertise in the topic of this report; potential conflicts of interest were considered in this process. Their comments were discussed during meetings and/or via email. They did not co-author the scientific report and do not necessarily agree with its content.
- Subsequently, the KCE research team invited three independent experts ('validators', cf. supra), with no prior involvement in the study, to conduct a peer review of the final version of the scientific report. They were selected based on their expertise in the topic of this report; potential conflicts of interest were considered in this process. The validators decided whether or not the report was scientifically valid. They did not co-author the scientific report and did not necessarily all three agree with its content. Their name and affiliation are included in the colophon.
- Finally, this report has been approved by a majority of votes by the Executive Board.
- KCE is exclusively responsible for the policy recommendations. Any (potential) persisting errors or omissions are the sole responsibility of KCE.

Publication date:

14 July 2026

Project number:

2025-02

Domain:

Health Services Research (HSR)

MeSH:

Reimbursement, Incentive; Quality of Health Care; Hospitals, General

NLM Classification:

W 84.4

Language:

English

Format:

Adobe® PDF™ (A4)

Legal depot:

D/2026/10.273/50

ISSN:

2466-6459

Copyright:

KCE reports are published under a "by/nc/nd" Creative Commons Licence <http://kce.fgov.be/content/about-copyrights-for-kce-publications>.



How to refer to this document?

Bouckaert N, Stokes J, Kristensen SR, Eyssen M, Gerkens S, Hansen MSH, Lefèvre M, Van de Voorde C. Pay-for-performance: a lever for quality improvement in acute hospitals?. Health Services Research (HSR) Brussels: Belgian Health Care Knowledge Centre (KCE). 2026. KCE Reports 424. <https://doi.org/10.57598/R424C>.

This document is available on the website of the Belgian Health Care Knowledge Centre.



■ TABLE OF CONTENTS

LIST OF FIGURES.....	3
LIST OF TABLES.....	3
LIST OF ABBREVIATIONS.....	4
■ SCIENTIFIC REPORT	8
1 INTRODUCTION	8
1.1 BACKGROUND	8
1.2 OBJECTIVE, RESEARCH QUESTIONS, SCOPE AND METHODS	10
2 PAY-FOR-PERFORMANCE IN BELGIUM: CURRENT SITUATION AND EVOLUTION	12
2.1 INTRODUCTION	12
2.2 ROLE OF QUALITY IN HOSPITAL FUNDING.....	13
2.3 DESIGN AND FUNCTIONING OF THE 2026 P4P PROGRAMME	15
2.4 EVOLUTION OVER TIME	16
3 PAY-FOR-PERFORMANCE INCENTIVES AND QUALITY OF CARE IN ACUTE HOSPITALS: A SYSTEMATIC REVIEW.....	20
3.1 INTRODUCTION	22
3.2 METHODS	22
3.2.1 Information sources	22
3.2.2 Search Strategy.....	22
3.2.3 Eligibility criteria.....	23
3.2.4 Study records	25
3.2.5 Data extraction	25
3.2.6 Risk of bias in individual studies.....	26
3.2.7 Data synthesis	26
3.3 RESULTS	31
3.3.1 Overview of selected studies.....	31
3.3.2 Summary of results by programme	32
3.3.3 Association between study outcomes, programme and study characteristics...	42
3.3.4 Intervention components, design features or contextual factors and P4P effectiveness (Programme theory)	46
3.3.5 Subgroup effects and equity.....	54
3.3.6 Cost or cost-effectiveness considerations.....	55
3.4 DISCUSSION.....	56
4 LESSONS LEARNED FROM THE DESIGN AND IMPLEMENTATION OF FIVE P4P PROGRAMMES IN ACUTE HOSPITALS	60



4.1	INTRODUCTION	62
4.2	METHOD	62
4.2.1	Selection of P4P programmes.....	62
4.2.2	Information collection process.....	63
4.3	OVERVIEW OF THE HEALTHCARE SYSTEM AND HOSPITAL FUNDING MECHANISMS.....	65
4.4	RATIONALE AND POLICY CONTEXT FOR THE INTRODUCTION OF P4P PROGRAMMES.....	66
4.5	FINANCING OF P4P PROGRAMMES: MACRO-LEVEL FRAMEWORK AND OPERATIONAL RULES	68
4.5.1	Integration with broader hospital payment systems	68
4.5.2	Type of financial incentive	69
4.5.3	Achievement or improvement incentives	70
4.5.4	Tournament- or threshold-based system	71
4.5.5	Performance measurement and scoring approaches	72
4.5.6	Allocation level and internal distribution of incentives	73
4.6	PROCESS FOR IDENTIFYING AND SELECTING QUALITY DOMAINS AND INDICATORS.....	75
4.6.1	Policy-driven prioritisation of quality domains	75
4.6.2	Clinical and expert-driven selection processes.....	76
4.7	TYPES OF QUALITY INDICATORS: STRUCTURE, PROCESS AND OUTCOME MEASURES.....	78
4.8	INTEGRATION OF P4P AND OTHER QUALITY INITIATIVES WITHIN HOSPITALS....	81
4.9	ADDITIONAL DESIGN AND IMPLEMENTATION FEATURES OF P4P PROGRAMMES	82
4.10	EVALUATIONS OF P4P PROGRAMMES	84
4.11	LIMITATIONS	88
5	IMPLEMENTATION OF A P4P PROGRAMME IN BELGIUM: POINT OF VIEW OF EXPERTS AND STAKEHOLDERS	89
5.1	INTRODUCTION	90
5.2	METHOD	90
5.2.1	Introduction.....	90
5.2.2	Targeted profiles for the modified nominal groups.....	91
5.2.3	Approach used during the nominal group meetings.....	91
5.3	RESULTS	92
5.3.1	Participants of the expert panels	92
5.3.2	Desirability of further pursuing a P4P programme for acute hospitals in Belgium	92



5.3.3	Feasibility to design a P4P programme that has a meaningful impact on quality of care.....	98
5.3.4	Development of a P4P programme in Belgium	103
5.4	LIMITATIONS	105
6	DISCUSSION.....	106
6.1	IS P4P AN EFFECTIVE AND APPROPRIATE STRATEGY FOR IMPROVING HOSPITAL CARE QUALITY?	106
6.1.1	P4P is not a universally applicable quality improvement strategy	106
6.1.2	But might work under narrow conditions	107
6.1.3	P4P to be used with caution and only selectively	110
6.2	LIMITATIONS OF THE BELGIAN P4P PROGRAMME	111
6.3	FUTURE ROLE OF P4P IN BELGIUM: A SELECTIVE AND STRUCTURED APPROACH	112
6.4	CONCLUSION	114
■	REFERENCES	116

LIST OF FIGURES

Figure 1 – Evolution of the P4P programme (2018-2026): budget, indicators and budgetary weight	18
Figure 2 – PRISMA flow diagram.....	32
Figure 3 – Programme theory relating P4P context-mechanisms-outcomes	50
Figure 4 – Cross tabulation of observed outcomes and design pathways by programme	54

LIST OF TABLES

Table 1 – Mixed-methods approach	11
Table 2 – Programme characteristics	34
Table 3 – Proportion of positive outcomes by programme	43
Table 4 – Associations between incentive and study design and positive P4P effects.....	45
Table 5 – Top ten mechanisms coded for explaining P4P programme results across studies	47
Table 6 – Mechanisms found more commonly in studies reporting overall beneficial outcomes for P4P	48
Table 7 – Mechanisms found more commonly in studies reporting overall detrimental or null outcomes for P4P	48
Table 8 – Mechanisms found more commonly in studies coded as using threshold-based payments and where overall headline finding beneficial	49
Table 9 – Programme theory alignment score – Programme level	52



Table 10 – Summary of study effects examining equity-relevant subgroups	55
Table 11 – Cost-effectiveness summary of findings	56
Table 12 – Template for programme assessment	64
Table 13 – Key features of selected P4P programmes (most recent status)	74

LIST OF ABBREVIATIONS

ABBREVIATION	DEFINITION
ACERIFH	Additional Charge for Emergency Reduction and Internal Fixation or Hemiarthroplasty for Hip Fracture
ACS	Acute Coronary Syndrome
AIM	Antimicrobial Stewardship, Infection Prevention and Control, and Microbiological Surveillance
AMI	Acute Myocardial Infarction
AMS	Antimicrobial Stewardship
API	Aligned Payment and Incentive
AQ	Advancing Quality
AR-DRG	Australian Refined Diagnosis Related Groups
BASE	Bielefeld Academic Search Engine
BC-P4P	Breast Cancer Pay-for-Performance Program
BFM	Budget of Financial Means
BHCS	Baylor Health Care System
BPT	Best Practice Tariffs
BSM	Be Service Minded
CABG	Coronary Artery Bypass Graft
CAUTI	Catheter-Associated Urinary Tract Infection
CBA	Controlled Before-After
CCG CQUIN	Clinical Commissioning Group Commissioning for Quality and Innovation
CEA	Cost-Effectiveness Analysis
CHF	Chronic Heart Failure
CI	Confidence Interval
CMO	Context-Mechanism-Outcome
CMS	Centers for Medicare & Medicaid Services
COPD	Chronic Obstructive Pulmonary Disease
CPIP	Clinical Practice Improvement Payment
CROM	Clinical-Reported Outcomes



CQUIN	Commissioning for Quality and Innovation
cTNM	clinical Tumor Nodes Metastasis
DGOS	Ministry of Health – ‘ <i>Direction générale de l'offre de soins, France</i> ’
DRG	Diagnosis Related Group
DSRIP	Delivery System Reform Incentive Payment
ED	Emergency Department
EDQIP	Emergency Department Quality Improvement Program
EPOC	Effective Practice and Organisation of Care
EQ-5D	EuroQol 5 Dimensions
EU	European Union
FPS	Federal Public Service
FRZV-CFEH	Federal Council for Hospital Facilities – <i>Federale Raad voor Ziekenhuisvoorzieningen / Conseil Fédéral des Etablissements Hospitaliers</i>
GHM	Homogeneous Patient Groups – ‘ <i>Groupes Homogènes de Malades, France</i> ’
GHS	Homogeneous Groups of Stays – ‘ <i>Groupes Homogènes de Séjours, France</i> ’
HAC	Hospital-Acquired Condition or Hospital-Acquired Complication
HAC-POA	Hospital-Acquired Condition Present on Admission
HACRP	Hospital-Acquired Condition Reduction Program
HAS	French National Authority for Health – ‘ <i>Haute Autorité de Santé, France</i> ’
HCAHPS	Hospital Consumer Assessment of Healthcare Providers and Systems
HEA	Health Equity Adjustment
HF	Heart Failure
HHS	Hospital and Health Services
HQID	Hospital Quality Incentive Demonstration
HRG	Health Resource Groups
HRRP	Hospital Readmissions Reduction Program
HSMR	Hospital Standardised Mortality Ratio
HSPA	Health System Performance Assessment
HVBP	Hospital Value-Based Purchasing
ICC	Intraclass Correlation Coefficient
ICU	Intensive Care Unit
IFAQ	Financial Incentives for Quality Improvement – ‘ <i>Incitation Financière à l'Amélioration de la Qualité, France</i> ’
IGAS	General Inspectorate of Social Affairs – ‘ <i>Inspection générale des affaires sociales, France</i> ’



IPC	Infection Prevention and Control
IPPS	Inpatient Prospective Payment System
IQR	Inpatient Quality Reporting
IQSS	Healthcare Quality and Safety Indicators – <i>'Indicateurs de Qualité et de Sécurité des Soins, France'</i>
ISQua	International Society for Quality in Health Care
ITS	Interrupted Time Series
LOK-GLEM	Local Quality Group – <i>'Lokale Kwaliteitskring / Groupe Local d'Evaluation Médicale'</i>
LOS	Length Of Stay
MOC	Multidisciplinary Oncology Consultation
MS-DRG	Medicare Severity Diagnosis Related Group
MZG-RHM	Minimal Hospital Data – <i>'Minimale Ziekenhuis Gegevens / Résumé Hospitalier Minimum'</i>
NEWS2	National Early Warning Score 2
NGT	Nominal Group Technique
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NQF	National Quality Forum
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
OQIP	Ontario Quality Improvement Plan
P4P	Pay-for-performance
P4Q	Pay-for-quality
P4R	Pay-for-results
PAQS	Platform for Continuous Improvement of Quality of Care and Patient Safety – <i>'Plateforme pour l'Amélioration continue de la Qualité des soins et de la Sécurité des patients'</i>
PbR	Payment by Results
PNU	Pneumonia
PREM	Patient-Reported Experience Measure
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PROGRESS-PLUS	Place, Race, Occupation, Gender, Religion, Education, Socioeconomic status, Social capital, personal characteristics associated with discrimination, features of relationships and time-dependent relationships
PROM	Patient-Reported Outcome Measure
PSS CQUIN	Prescribed Specialised Services Commissioning for Quality and Innovation



QALY	Quality-Adjusted Life Year
QIP	Quality Improvement Payment or Quality Improvement Plan
QOC	Quality-Oriented Consultation
RIZIV-INAMI	National Institute for Health and Disability Insurance – <i>‘Rijksinstituut voor Ziekte- en Invaliditeitsverzekering / Institut National d’Assurance Maladie-Invalidité’</i>
SDT	Self-Determination Theory
SE	Standard Error
TNM	Tumour, Node, Metastasis
UCSFMC	University of California at San Francisco Medical Center
UPMC	University of Pittsburgh Medical Center
US	United States
VBP	Value-Based Purchasing
VBR	Value-Based Reimbursement
VIKZ	Flemish Institute for Quality of Care – <i>‘Vlaams Instituut voor Kwaliteit van Zorg’</i>
VIP	Value Incentive Program
WHO	World Health Organization



■ SCIENTIFIC REPORT

1 INTRODUCTION

1.1 Background

QUALITY ISSUES FOSTERING QUALITY IMPROVEMENT INITIATIVES

Substantial progress has been made in improving several aspects of hospital care quality, both in Belgium and internationally, for example in cancer survival and mortality from cardiovascular diseases.¹⁻⁴ The most recent Health System Performance Assessment (HSPA) report shows that Belgium performs particularly well in terms of 5-year survival for breast and colorectal cancer and treatable mortality overall.²

However, progress in other areas has been slower and uneven, both nationally and internationally. This is, amongst others, the case for healthcare-associated infections, antimicrobial resistance, prevention of adverse events, and avoidable hospital admissions.^{2, 4-6}

In addition to these persistent quality issues, recent research in Belgium based on hospital administrative data has revealed substantial and unwarranted inter-hospital variation in risk-adjusted mortality, readmissions, prolonged length of stay, and patient safety indicators.⁷⁻¹¹ Similar variation has also been observed in the implementation of, and adherence to, evidence-based procedures.¹²⁻¹⁴ Such unwarranted variation not only undermines healthcare quality and patient safety, but also raises equity concerns, as patients may experience different outcomes depending on the hospital they attend. Importantly, evidence suggests that between-hospital variation appears to be at least partly driven by systemic hospital-level factors. Longitudinal analyses show that hospitals tend to remain consistent in their level of performance over time, whether high or low. Moreover, hospitals that perform poorly or well in one domain often show consistent performance patterns across other areas.^{9, 11, 15}

These quality shortcomings have important consequences. Above all, they lead to avoidable harm for patients and contribute unnecessarily to the overall disease burden. At the same time, they impose a substantial economic cost, reflecting inefficient use of healthcare resources.¹⁶ It is estimated that more than 10% of hospital expenditure in high-income countries is devoted to managing preventable complications of care and patient harm, in addition to broader societal costs related to disability, impaired functioning, and lost productivity, as well as the environmental impact of avoidable care.^{5, 16-19} Furthermore, poor-quality care disproportionately affects more vulnerable population groups, thereby exacerbating health inequalities.¹⁷

Improving the quality of hospital care therefore remains a key priority.²⁰ Growing awareness of quality issues has led healthcare providers and policymakers to seek effective ways to improve care at both institutional and system levels.²⁰⁻²² Over the past decades, a wide range of quality improvement initiatives has been developed, targeting different levels of the healthcare system, from micro- to macro-levels. These include regulatory approaches (e.g. licensing and standards), workforce training, infection prevention measures, external assessments (e.g. accreditation and certification), public reporting and benchmarking, pay-for-quality programmes, inspection and control mechanisms, audit and feedback, safety incident reporting systems, clinical guidelines and decision support tools, care pathways, peer review, and collaborative learning initiatives.^{17, 22-24}

Despite this broad set of tools, governments struggle to find and implement a mix of strategies that best fit their healthcare system context and ensure the quality of care delivered to patients.^{23, 25} One reason is that the effectiveness of these levers varies considerably, depending on programme design, contextual factors, and the extent to which they are implemented as standalone interventions or as part of a coherent and integrated quality framework. Moreover, the available literature rarely addresses the question of the superiority of individual strategies and only provides limited guidance to

policymakers on which approaches are most appropriate in specific settings.²⁵ Policymakers must also balance the need to stimulate intrinsic motivation among healthcare providers with requirements for accountability and transparency, while avoiding unnecessary administrative burden and allowing room for innovation.²⁵

PAY-FOR-PERFORMANCE: FINANCIAL INCENTIVES CONDITIONAL ON QUALITY OF CARE

This report focuses on one specific strategy to improve quality of care: pay-for-performance (P4P), also referred to as pay-for-quality (P4Q), in acute hospital care.

Pay-for-performance (P4P) or pay-for-quality (P4Q, when focusing exclusively on the quality component of performance) refers to schemes in which healthcare providers receive financial incentives – either rewards, penalties or a combination – based on their achievement of specific targets or demonstrated improvements in the quality and/or efficiency of care delivery. Throughout this report, the term P4P is used, in line with the naming of the Belgian programme, although the analysis focuses exclusively on quality of care, for which the term P4Q would be more precise.

Two key features characterise P4P programmes:²⁶

- (1) performance of providers is monitored through quality measurement, typically using structure, process, and/or outcome indicators;
- (2) financial payment is conditional on the achievement or improvement of the measured quality.

The underlying rationale of P4P programmes is to modify incentives for healthcare providers (professionals and organisations) by explicitly rewarding high-quality care or improvements in quality, and in some cases penalising poor performance. These schemes aim to support quality improvement by directing attention to specific areas of care and by strengthening extrinsic motivation. In this respect, P4P differs from other quality improvement strategies, such as clinical practice guidelines, audit and feedback, public reporting, or accreditation, which primarily seek to influence provider behaviour through evidence-based information, peer comparison, or reputational effects rather than financial incentives.

P4P initiatives are part of a broader movement toward value-based purchasing, which seeks to shift healthcare financing from rewarding volume (activity-based payment) to rewarding quality or value (outcomes relative to costs). They go back at least to the 1980s,²⁷ but had their peak in the early 2000s following the US Institute of Medicine's (now the National Academy of Medicine) seminal report *Crossing the Quality Chasm*, which highlighted the lack of incentives for healthcare providers to provide high-quality care as one of the main challenges of the US health system.²⁸ Since then, P4P have been widely adopted as a policy lever in healthcare systems in the US, Europe and beyond in both primary and secondary care.^{21, 26, 29, 30}

P4P, AN APPEALING IDEA WITH UNCERTAIN IMPACT

Despite the initial enthusiasm, likely due to the political appeal and high face validity of the idea, the evidence on the effectiveness of P4P in improving the quality of acute hospital care is less than convincing.^{21, 31-33} Some programmes have shown moderately positive effects, which are at times short-lived, while others have reported limited, no, or even negative effects.

Early reactions to these mixed findings emphasised the need for a nuanced understanding of P4P as a heterogeneous set of interventions with different objectives and characteristics.^{34, 35} Indeed, rather than one specific intervention, P4P is best understood as a complex combination of design choices which all affect the likelihood of P4P achieving its intended aim of improving the quality of care^{21, 26, 33, 36} but the evidence for the impact of these design choices is lacking.

More recent commentary has called for a rethink of P4P³⁷ and a move toward alternative payment models that aim more explicitly at generating 'value' in health spending.³⁸ In line with this, recent international experience suggests a gradual shift away from narrow, condition-specific P4P schemes towards alternative payment models in which quality is embedded in population-based payment



arrangements or episode-based payments across care pathways, often spanning multiple care settings. These models seek to align incentives not only with quality, but also with care coordination, efficiency, equity, and value.³⁹ However, early evidence remains limited and mixed and suggests that, like traditional P4P, such models face important challenges in terms of design, risk adjustment, implementation burden, and their ability to generate meaningful improvements in outcomes.³⁹

1.2 Objective, research questions, scope and methods

OBJECTIVE AND RESEARCH QUESTIONS

The continued political interest in P4P is also evident in Belgium. Although a P4P programme for acute hospitals has been in place since 2018, the substantial increase in its budget in 2024 (from € 7.2 million in 2023 to € 40 million as from 2024 onwards) has prompted renewed discussion among stakeholders and policymakers. The original research proposal therefore called for a revision of the Belgian P4P programme for acute hospitals, including the development of a new set of indicators, as well as adaptations to payment modalities, incentive structures, and budget allocation.

However, given the limited and inconclusive evidence on the effectiveness of P4P in improving hospital care quality, as well as concerns regarding its cost-effectiveness,^{40, 41} it was considered necessary to first undertake a comprehensive assessment of the most recent evidence. This study focuses on both intended and unintended effects, and on identifying which intervention components, design features, and contextual factors are associated with greater or lesser effectiveness. Understanding what works (or does not), under which conditions, and with which design characteristics is essential to draw lessons and recommendations that can inform evidence-based improvements to the Belgian programme.

To meet this objective, the report addresses the following research questions:

- (1) Is P4P an effective and appropriate strategy for improving hospital care quality? Should it be pursued in Belgium?
- (2) If so, under what conditions and with which design features should it be implemented?

SCOPE OF THE STUDY

This report builds on a previous KCE report ([KCE Report 118](#)), published in 2009, which examined the 'Advantages, disadvantages and feasibility of the introduction of 'Pay for Quality' programmes in Belgium'.^{42, 43} That earlier report had a broader scope, covering theoretical literature as well as evidence from primary care, and was conducted in a different context, as no P4P programme had yet been implemented in Belgium.

The scope of the present report is focused on P4P programmes in acute hospital care in high-income countries. It examines their design, implementation, functioning, and impact on quality of care. Only programmes that directly incentivise the quality of hospital care are considered; P4P schemes targeting other dimensions of health system performance, such as efficiency alone, fall outside the scope of this analysis. Broader reforms that introduce alternative payment models with a P4P component, such as quality-adjusted population-based or episode-based payments (see section 1.1), are generally not included. In such reforms, the introduction of a quality adjustment is usually part of a broader and simultaneous change in the underlying payment model, making it difficult to distinguish the effect of the P4P component from the effect of the payment reform as a whole. Moreover, the evaluation of alternative payment models as such is beyond the scope of this report.

The concept of quality of care itself encompasses multiple dimensions and is defined in different ways in the literature. Most definitions include effectiveness, patient safety, and patient-centredness as core components.²⁴ Some definitions also incorporate attributes such as appropriateness, timeliness, efficiency, access, and equity, which may blur the distinction between quality of care and broader health system performance.

In this report, quality of care is defined in line with the most recent Belgian HSPA report, which adopts the Institute of Medicine definition: 'the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge'.^{2, 44} In the HSPA report, quality is further subdivided into five sub-dimensions: effectiveness, safety, appropriateness of care, continuity of care, and people-centredness. P4P interventions may target one or more of these dimensions, depending on their design and the indicators selected.

METHODS

The study follows a mixed-methods approach. The main steps are summarised in Table 1. The methods and more specific research questions are further outlined at the beginning of the respective chapters (where relevant).

Table 1 – Mixed-methods approach

What?	How?
Description of the current Belgian P4P programme and its evolution over time	Review of peer-reviewed and grey literature, including policy papers and programme guidelines
Systematic literature review to assess the effectiveness of P4P programmes in acute hospitals and to identify insights on design, implementation, and evaluation.	Systematic review of peer-reviewed and grey literature in multiple databases (from 2008 onwards) Descriptive summary of results by programme Quantitative synthesis of the evidence using a vote-counting approach Realist-informed narrative synthesis to assess which intervention components, design features or contextual factors impact P4P effectiveness
International comparison of five P4P programmes (objectives, design elements, evaluation strategy, stakeholder involvement, and integration within broader quality policy)	Selection of programmes based on short and long list Development of a template for the programme descriptions Review of peer-reviewed and grey literature, and consultation of international experts Transversal analysis across programmes
Expert panels to discuss findings from systematic literature review and international comparison and assess feasibility and desirability of P4P in Belgium	Consultation of Belgian experts via nominal group technique Stakeholder meeting to discuss findings and recommendations
Scientific validation	Review of the scientific report and supplement by three independent scientific experts

The report is structured as follows. Chapter 2 describes the current Belgian P4P programme, and its evolution over time. Chapter 3 presents the evidence of the systematic literature review on the effectiveness of P4P programmes in acute hospitals as well as on the intervention components, design features, and contextual factors associated with their impact. Chapter 4 provides a transversal analysis of five P4P programmes, covering their objectives, design and functioning, evaluation strategies, stakeholder involvement, and integration within broader quality policies. Detailed programme descriptions are available in [Supplement 1](#) to this report. Chapter 5 presents the results of the expert panels, reflecting the perspectives of Belgian experts and stakeholders on the international evidence and its implications for the feasibility and desirability of P4P in Belgium. Chapter 6 concludes with a discussion of the findings and the limitations of the study.



2 PAY-FOR-PERFORMANCE IN BELGIUM: CURRENT SITUATION AND EVOLUTION

KEY POINTS

- **Limited role of quality in hospital financing.** Hospital financing in Belgium remains largely activity-based, creating strong incentives for volume rather than quality. Quality-related payments, including P4P, represent only a small share of total hospital revenue.
- **From fixed funding to performance-based incentives.** The Belgian P4P programme for acute hospitals was introduced in 2018, replacing earlier lump-sum quality contracts. Participation remained formally voluntary but de facto universal under both systems. Although labelled as P4P, the programme focuses exclusively on quality (P4Q).
- **Current P4P design dominated by structure indicators and relative rewards.**
 - o The 2026 programme mainly rewards accreditation, participation in certificates and data collection (registries, patient experience measurement), alongside a single outcome indicator (hospital-wide mortality).
 - o Payments are allocated through a closed-budget system in which hospitals compete for a fixed envelope and are rewarded based on their own and peers' performance and bed size.
 - o Feedback is provided to hospitals, but results are not publicly disclosed.
- **Key design changes over time.**
 - o The budget increased more than fivefold to € 40 million from 2024 onwards, partly financed through reductions in funding for clinical services. The programme initially operated with a global budget based on a composite score, but shifted to indicator-specific sub-budgets.
 - o The programme evolved from a mix of structure, process, and outcome indicators to a more limited set, with process indicators phased out and increasing dominance of structure indicators alongside a single outcome measure. There was a reduction in number of incentivised indicators between 2018 and 2023.
 - o There were frequent changes in process and outcome measures over time.
 - o Most process and outcome indicators rely on retrospective data, with an important time lag. This means that hospitals are assessed on performance from a period when the indicator was not yet known, limiting their ability to implement improvements.
- **Weak link between improvement and payments.** Improvements in performance are not consistently reflected in financial rewards due to the payment design.

2.1 Introduction

In 2018, a federal pay-for-performance (P4P) programme for acute hospitals was implemented in Belgium. Although labelled as P4P, the programme focuses exclusively on quality metrics and is therefore more accurately described as a pay-for-quality (P4Q) initiative. It replaced earlier quality and patient safety coordination contracts, which had provided fixed, lump-sum funding for quality improvement activities based on hospital bed size, without linking funding to performance. These contracts had been in place for approximately a decade prior to the introduction of P4P.

The P4P programme was implemented in the context of broader hospital financing reforms under Minister De Block. It was developed by a P4Q working group and submitted for advice to the Federal Council for Hospital Facilities (FRZV - CFEH).⁴⁵ Initially modest in scale (€ 6 million in 2018), the

programme budget increased gradually to € 7.2 million in 2023, followed by a substantial expansion to € 40 million from 2024 onwards.

This chapter aims to describe the current P4P programme and to provide an overview of its main developments over time, as well as its position within the broader hospital financing system.

2.2 Role of quality in hospital funding

HOSPITAL FINANCING LARGELY BASED ON ACTIVITY...

General hospitals in Belgium receive revenue from multiple sources. In 2023, total hospital revenue amounted to € 25 billion. The two main sources are the Budget of Financial Means (BFM), accounting for approximately 37% of revenue, and deductions from physician fees and technical services, accounting for about 36%. These are complemented by revenues from pharmaceutical products (22%), RIZIV-INAMI conventions (4%), and several smaller sources.⁴⁶

The BFM is a closed-end national budget divided into several parts and subparts.^{47, 48} It consists of three main parts (A, B, and C) which are further subdivided into a total of 15 components (from A1 to C3). Part A covers capital and investment costs, Part B covers operational costs, and Part C relates to regularisations.^a Each subpart is governed by specific rules and criteria, resulting in a complex and often opaque method for calculating individual hospital budgets.^{47, 48} Although each rule and exception may have a specific rationale, the system as a whole has become highly complex and lacks transparency over time.⁴⁸

Subpart B2 is the largest component, accounting for 37.3% of the BFM for general hospitals in 2025.⁴⁶ It covers the costs of clinical services provided by nursing and other care personnel (excluding physicians), as well as medical equipment. The national budget for subpart B2 is fixed and allocated to individual hospitals using a diagnosis-related group (DRG)-based point system. Hospitals earn points based on the volume of justified activity, which is determined by their case mix and the average length of stay across hospitals over the past three years.^{48, 49} The monetary value of each point depends on the total number of points generated by all hospitals.

This system creates competition for limited resources and provides strong financial incentives for hospitals to increase admissions and reduce LOS in order to remain below the national average per DRG.^{48, 49} At the same time, it offers limited incentives to reduce complications, as complications may lead to reclassification of patients into higher case-mix categories. Combined with fee-for-service remuneration for physician fees and technical services, these mechanisms primarily incentivise increased clinical activity rather than value or quality of care. Well-documented examples of potential overuse include laboratory testing, medical imaging, placing knee or hip replacements too quickly, arthroscopic meniscectomy, and duplicate investigations due to insufficient coordination.^{2, 3, 50-52}

... WITH SOME STRUCTURAL QUALITY ELEMENTS

While activity-based financing may improve efficiency, it has also raised concerns about quality trade-offs and cost-shifting to other care settings.⁴⁹ Hence, alongside activity-based financing, elements of the BFM incorporate structural and procedural incentives intended to support quality of care.

The main part in the BFM that encompasses quality of care initiatives is the B4 component.^b This subpart, which accounts for 16.2% of the BFM in 2025, finances a range of structural and process-related elements that support the quality of hospital care.^{46, 47, 53} These include funding for key organisational roles and teams that are presumed to contribute to quality of care, such as the chief

^a Note that since the 6th State reform in 2014, the federated entities have full responsibility for funding investment costs (part A). While these means are distributed through the BFM, they are financed by the federated entities. The federal level remains responsible for funding operating costs (Part B).

^b Originally, the B4 part was used to compensate hospitals for revenue losses as a result of bed closure.



medical officer, infection control staff, geriatric liaison teams, pain management teams, and transfusion safety systems, as well as support for oncology care programmes and staff training and qualifications. In addition, hospitals receive funding linked to specific quality-related processes, including mandatory (and voluntary) data collection on healthcare-associated infections, participation in specific surveys and surveillance programmes, and the implementation of electronic medical records. Some components also support quality monitoring, evaluation, and participation in pilot projects aimed at improving care delivery and hospital management.

PAY-FOR-PERFORMANCE AND OTHER CONDITIONAL PAYMENTS

Some hospital financing mechanisms within the Belgian system explicitly link payments to the attainment of quality-related targets or improvements in care delivery.

The P4P programme falls within this category. Its initial budget (€ 6 million in 2018, with subsequent indexations) is financed through the B4 subpart of the BFM, while the substantial budget increase between 2023 and 2024 is financed through a reduction in the B2 subpart, which covers clinical services.⁵³ In 2026, the total P4P budget amounts to € 40.8 million (see Figure 1). However, this remains limited relative to overall hospital financing, representing approximately 0.4% of the BFM and around 0.15% of total hospital revenue.

Other examples of conditional financial incentives exist. Since 2014, hospitals have received reduced payments for readmissions occurring within 10 days in the same hospital. In such cases, only a proportion of certain lump-sum payments is reimbursed (e.g. 67% for medical imaging and clinical biology, and 82% for pharmaceutical products and per diem payments), thereby introducing a financial disincentive for avoidable readmissions.⁴⁹

A second example is the RIZIV-INAMI dialysis care convention introduced in 2016, whereby reimbursement levels are linked to the proportion of patients receiving an 'alternative' form of dialysis, with a target threshold (>35%) (see Box 1 for more details).

More recently, in 2026, a pay-for-results (P4R) approach has been introduced within antimicrobial stewardship (AMS) initiatives (see Box 1 for more details). This approach links financial support to demonstrable progress in the implementation of AMS improvement strategies at hospital level. It is designed as a structured improvement trajectory: in an initial phase, funding is conditional on the submission of a well-substantiated improvement plan tailored to the local context; in subsequent phases, funding is progressively linked to the achievement of predefined targets, including both process and outcome indicators, as well as the effective implementation of the improvement strategy.

Box 1 – Examples of hospital financing conditional on performance targets

RIZIV-INAMI dialysis care convention

Since 1 August 2016, a national convention between RIZIV-INAMI and hospitals with licensed dialysis centres has regulated the financing of dialysis care, including both medical fees and lump-sum payments for different dialysis modalities.^{2, 54, 55} Within this framework, hospitals are encouraged to increase the use of 'alternative' dialysis modalities, such as home dialysis, self-care dialysis, or peritoneal dialysis, which are less costly and often more patient-centred than in-hospital haemodialysis.^{2, 54} To this end, hospitals are expected to reach a target proportion of 40% of patients treated with these alternative modalities. Failure to meet this threshold results in a reduction of medical fees and lump-sum payments in the subsequent year, thereby introducing a financial disincentive.⁵⁴

Although not formally labelled as a P4P scheme, this mechanism can be interpreted as a form of performance-based financing, as reimbursement levels are partly conditional on achieving a predefined target related to care delivery. It aims to incentivise a shift towards more efficient and appropriate care modalities and appears to be effective in practice, with most hospitals reaching the target.^{2, 55}

Antimicrobial stewardship

The AIM (Antimicrobial Stewardship, Infection Prevention and Control, and Microbiological Surveillance) project is a national programme running from 2026 to 2030, developed within the framework of the National Action Plan on antimicrobial resistance and financed through subparts B4 and B5 (operational costs of the hospital pharmacy) of the BFM.⁵⁶ Its main objectives are (1) to strengthen antimicrobial stewardship (AMS) and infection prevention and control (IPC) through monitoring, feedback, capacity building, collaborative learning, and the sharing of best practices; (2) to achieve measurable quality improvements through several improvement cycles; and (3) to build experience with result-based financing for future quality improvement policy. While the programme comprises two complementary tracks, both rely on monitoring and peer-to-peer learning. However, only the AMS component is financed through a pay-for-results (P4R) mechanism, whereas the IPC component receives financial support for capacity building without direct performance-based payments.

The AMS P4R approach targets three key national priorities related to antibiotic use, including avoiding unnecessary prescriptions, optimising treatment duration, and promoting de-escalation towards narrower-spectrum agents. Hospitals select one of these priorities and develop a corresponding improvement strategy. The programme is implemented through several structured improvement cycles, in which hospitals first submit a context-specific improvement plan based on baseline measurement and analysis of behavioural barriers, and subsequently define and work towards predefined targets for both process and outcome indicators. Financial support is initially linked to the quality of the proposed plan and, in later stages, increasingly to the achievement of targets and the effective implementation of the strategy.

2.3 Design and functioning of the 2026 P4P programme

The 2026 P4P programme represents a continuation of the design implemented in the previous two years.⁵⁷ Participation is formally voluntary; however, as non-participation implies relinquishing part of the hospital budget, participation is de facto universal among general hospitals.^{58, 59}

The programme operates with a fixed budget of € 40.8 million, which is allocated across three structure indicators and one outcome indicator. The structure indicators primarily focus on participation and compliance rather than gradations in performance. Each indicator has a predefined sub-budget, which is distributed among hospitals based on the number of points they obtain.

STRUCTURE INDICATORS DOMINATE THE PROGRAMME

A first structure indicator relates to hospital accreditation and is associated with a budget of € 8 million. It rewards institutions accredited by an ISQua-recognised body. A valid accreditation on 15 March 2026 yields the maximum of 20 points, while accreditation obtained during a previous P4P cycle (2021-2025) yields 10 points.

A second structure indicator, also with a budget of € 8 million, incentivises participation in clinical registries, surveillance systems, and quality labels. Eligible activities include participation in (1) the Belgian Trauma Registry (registration year 2026); (2) at least two of four non-mandatory surveillance modules on healthcare-associated infections coordinated by Sciensano (registration year 2025); (3) central line-associated bloodstream infection surveillance coordinated by Sciensano (registration year 2025); and (4) the Global Point Prevalence Survey of Antimicrobial Consumption and Resistance (registration year 2025), as well as (5) attainment of the Baby-Friendly Hospital Initiative certificate. Hospitals can obtain up to 20 points, with 4 points awarded per activity.

A third structure indicator concerns patient experience measurement and is likewise allocated € 8 million. Hospitals that participated in 2025 in standardised patient experience surveys conducted by approved organisations (e.g. BSM-management, PAQS, Santhea, VIKZ) receive the maximum score of 20 points, whereas hospitals that did not participate receive 0 points.



Outcome measurement is limited to a single indicator, the Hospital Standardised Mortality Ratio (HSMR), with a budget of € 16.8 million. The HSMR is calculated using administrative data (MZG-RHM, registration year 2024) and adjusted for patient case-mix (including age, sex, and comorbidity burden, with model estimation at the level of DRGs). A maximum of 40 points can be obtained, of which 20 points are based on a continuous score (linearly scaled between the worst (0 points) and best (20 points) HSMR values), and 20 points are based on categorisation, benchmarking hospitals according to whether observed mortality is statistically significantly lower than expected based on the risk-adjustment model ('better than expected', 20 points), not significantly different from expected ('as expected', 16.66 points), or significantly higher than expected ('worse than expected', 13.33 points).

BUDGET ALLOCATION DEPENDS ON RELATIVE PERFORMANCE

The allocation of the budget to individual hospitals depends on both their performance score and their size, measured in justified acute beds, as well as on the performance and size of all other hospitals. The monetary value of a point for each indicator is determined by dividing the sub-budget of the indicator by the total number of points achieved across all hospitals, weighted by bed size:

$$\text{Point value per indicator } i = \frac{\text{Budget for indicator } i}{\sum_{\text{all hospitals } h} \text{Points}_{h,i} \times \text{Size}_h}$$

The monetary value of a point is multiplied by both the number of points obtained by each hospital and its bed size to determine the payment per indicator. The total P4P budget allocated to a hospital is then obtained by summing these payments across all indicators.

$$\text{Budget}_h = \sum_{\text{all indicators } i} \text{Points}_{h,i} \times \text{Size}_h \times \text{Point value}_i$$

The programme follows a closed-budget, tournament-like approach, in which hospitals effectively compete for a share of a fixed envelope. As a result, the financial value of a given score depends on the performance of all participating hospitals. The programme rewards relative performance rather than improvement over time.

FEEDBACK, BUT NOT PUBLIC

Hospitals receive individual feedback reports detailing their scores per indicator, which support internal quality monitoring. However, these reports do not include the performance of other hospitals or national averages, limiting benchmarking opportunities. Although public reporting was envisaged in earlier programme documentation (2019), results are not systematically disclosed.^{c60} Procedures are in place to allow hospitals to review and, where necessary, contest the data underlying their scores.

2.4 Evolution over time

FROM FIXED-FUNDING QUALITY CONTRACTS TO P4P

Prior to the introduction of the P4P programme, hospitals participated in 'quality and patient safety coordination contracts', implemented in two consecutive multiannual programmes (2007–2012 and 2013–2017). These contracts supported hospitals in implementing structured quality improvement initiatives through thematic priorities, including patient safety management, high-risk medication, safe surgery, and transmurality care.^{47, 59} In the second programme, the criteria were aligned with international accreditation requirements, thereby supporting hospitals engaging in accreditation trajectories.⁴⁷

^c Some performance data are publicly reported on a voluntary basis for a selection of indicators (e.g. oncology, PREMs) by Flemish acute hospitals through the Flemish Institute for the Quality of Care (VIKZ), mainly relating to indicators used in previous programme years (see Figure 1).

Hospitals received financial support to implement these measures and for training purposes, and were required to submit annual progress reports.^{47, 59} Funding was largely fixed and linked to hospital size. During the period 2013–2017, the annual budget amounted to approximately € 5.8-5.9 million.⁵⁹

In 2018, these contracts were discontinued for acute hospitals and replaced by the P4P programme, with a budget broadly in line with the previous system. Participation remained formally voluntary but de facto universal under both systems.^{47, 58, 59} To ensure a smooth transition, the P4P programme initially included a fixed component (20% in 2018 and 10% in 2019), before becoming fully performance-based from 2020 onwards, with payments linked to hospitals' scores on structure, process, and outcome indicators.⁵⁹⁻⁶²

KEY DESIGN CHANGES OVER TIME

A strategic P4Q working group, chaired by the FPS Public Health, was established in 2017 to develop the programme through a co-creation process involving the administration, the Minister's policy cell, RIZIV-INAMI, professional organisations, academic experts, expert bodies, and hospital representatives.⁶¹ In addition, domain-specific expert groups involving clinicians and scientists contributed to the development of individual indicators, such as oncology indicators in collaboration with the Belgian Cancer Registry.⁶¹ The resulting indicators and programme proposals were submitted for advice to the Federal Council for Hospital Facilities.⁴⁵ For each indicator, a technical sheet was developed describing its rationale, technical definition, targets, and scoring methodology.

The programme and indicator set were initially revised annually through this governance structure.⁶⁰⁻⁶⁴ However, the strategic P4Q working group was paused in 2021 during the COVID-19 pandemic and was not formally re-established afterwards. Since then, annual revisions have mainly been overseen by the Minister's policy cell and the FPS Public Health, which administers the programme and publishes the indicator specifications and instructions.

Figure 1 presents the evolution of the Belgian P4P programme for acute hospitals between 2018 and 2026, including the total budget (B), the types of indicators used, i.e. structure (S), process (P), and outcome (O), and their relative weight within the performance-based (variable) part of the budget.^{57, 60-68} The percentages shown between square brackets represent the share of each indicator in the annual budget. For indicators spanning multiple years, a range is shown reflecting the minimum and maximum budget share over time. The dates in brackets indicate the data registration period used for the calculation of the indicator. If no year is mentioned, the calculation is based on data from the current or previous year. Given the frequent changes in the programme, the description below focuses on the main trends rather than the detailed specifications of each individual indicator.

In its initial phase, the programme combined structure and process indicators, with outcome indicators added from 2019 onwards. These included both hospital-wide and pathology-specific measures. Process indicators covered areas such as oncology registration (e.g. TNM classification, WHO performance score), antibiotic prophylaxis, and clinical pathways, and accounted for approximately 6% to 31% of the performance-related budget. Over time, these process indicators were phased out and are no longer included in recent programme years. Overall, the number of indicators decreased substantially between 2018 and 2023. Since 2023, the programme relies mainly on hospital-wide indicators, with no process indicators. This shift was reinforced by changes to the PREM indicator, which evolved from a measure combining participation and results to one based solely on participation. As a consequence, structure indicators became dominant, although the weight of outcome indicators has also increased since 2024.



Figure 1 – Evolution of the P4P programme (2018-2026): budget, indicators and budgetary weight

	2018	2019	2020	2021	2022	2023	2024	2025	2026
B	6	6.2	6.2	6.3	6.4	7.2	40	40	40.8
S	Accreditation [20-33%]								
	Data registration and labels [5-33%]								
	Patient safety reporting system [10-13%]								
							PREMs [20%]		
P	PREMs [15-33%]								
	Oncology registration [19-31%]								
	cTNM class (2013-2015)	pTNM class (2015-2016)	WHO performance score (2017)	pTNM class (2015-2018)	cTNM class (2018-2019)				
	Breast cancer QI (2012-2014)		Topography of tumour (2015-2017)						
	Antibiotics prophylaxis [6-12%]								
	Hip, hysterectomy, pacemaker (2012-2014)	Knee (2016)							
	O		Mortality [6-20%]					Mortality [40%-41%]	
SMR hip fracture (2016-17)			SMR hip fracture (2016-18)	HSMR (2019)				HSMR (2023)	HSMR (2024)
90d rectum cancer surgery (2014-2016)			90d colon cancer surgery (2014-2016)						
			SMR stroke (2016-2018)						

Note: B = budget (in million euro), S = Structure indicators; P = Process indicators; O = Outcome indicators. Percentage between square brackets indicate the budgetary weight of the indicator in the performance-based (variable) part of the P4P budget. Dates between brackets indicate the data registration year(s) used to calculate the indicator.

Sources: P4P guidance documents.^{57, 60-68}

Across the entire period, two structure indicators remained consistently present: accreditation, and data registration and quality labels, although the content of the latter changed regularly. The emphasis on accreditation is consistent with the preceding quality and safety contracts (see above).⁵⁹ In contrast, most process and outcome indicators varied from year to year, particularly before 2023, while the programme composition has been more stable since 2024.

An important feature is the time lag between care delivery and performance assessment. Most process and outcome indicators rely on retrospective data, meaning hospitals are assessed based on performance from a period when the indicator was not yet known, limiting the ability to anticipate and plan improvements. This issue was recognised early on, with 2019 guidance indicating the intention to reward improvement and to reintroduce indicators once more recent data become available.⁶⁰ A first attempt to assess improvement was made in 2022 by reusing oncology indicators (cTNM classification) based on data from 2018-2019, i.e. the period in which the indicator was first introduced. This approach was, however, not continued.^{64, 68} Recently, the HSMR indicator in 2026 uses data from 2024, the year in which the indicator was first introduced.

The method of budget allocation has also evolved over time. Initially, the programme operated with a single global budget, distributed across hospitals based on a composite score covering all indicators. From 2024 onwards, each indicator is associated with its own budgetary envelope, which is distributed separately. This change coincided with a substantial increase in the overall budget, from € 7.2 million in 2023 to € 40 million in 2024. As this increase was not accompanied by a proportional expansion in the number of indicators, existing indicators such as accreditation and data registration became associated with substantially larger financial incentives.

Finally, the indicators differ in nature and complexity. Structure indicators are generally simple and transparent, often binary (e.g. participation or accreditation status), though their link to patient outcomes is indirect. In contrast, the current outcome indicator (HSMR) is more complex, less transparent, and broader in scope than earlier, condition-specific outcome measures.

PERFORMANCE IMPROVEMENTS NOT CONSISTENTLY REFLECTED IN FINANCIAL REWARDS

To date, no methodologically robust formal evaluation has assessed the impact of the Belgian P4P programme on quality of care or patient outcomes. Brouwers et al. (2024), however, examined the relationship between hospitals' P4P composite scores and the corresponding budgetary outcomes over the period 2018-2020.⁵⁹

They found that 84% of hospitals achieved a higher P4P score in 2020 compared to 2018, and 55% of hospitals showed a year-on-year increase in their scores. Among these hospitals with consistently improving scores, only about one quarter experienced a corresponding year-on-year increase in incentive payments. This discrepancy reflects the closed-budget nature of the programme. In contrast, incentive payments decreased over time in 29% of hospitals, while performance declined consistently in only two hospitals.

The analysis also suggests that the system favours larger hospitals. While 64% of hospitals saw a decrease in incentive payments per bed, 80% of large hospitals (>800 licensed beds) experienced an increase in total payments, compared to only 7% of smaller hospitals (<400 licensed beds). Finally, high-performing hospitals at baseline experienced a greater decline in incentive payments over time than lower-performing hospitals.



3 PAY-FOR-PERFORMANCE INCENTIVES AND QUALITY OF CARE IN ACUTE HOSPITALS: A SYSTEMATIC REVIEW

KEY POINTS

- A **systematic review of the literature** was conducted to assess the **effect of P4P programmes on the quality of care in acute hospitals**, and to identify insights related to their design, implementation, and evaluation.
- The vast majority of the **90 studies and 31 P4P programmes** included in the systematic literature review were US-based (66 studies, 15 programmes), England being the second largest contributor (10 studies, 3 programmes). Three EU countries (France, Germany, Italy) and 5 other high-income countries accounted for the remaining evidence.
- **Attribution remained a persistent challenge**, as secular trends and concurrent quality improvement initiatives often made it difficult to isolate the independent effect of P4P programmes.
- A **descriptive summary of the impact of quality of care by P4P programme showed very mixed, inconclusive, or mostly modest findings**, providing **little support for P4P as a consistent lever for quality improvement**.
 - **Positive effects were often not sustained over time, when subject to more rigorous analysis, or may partly reflect publication bias favouring positive findings.** More extensively evaluated programmes, such as the Hospital Readmissions Reduction Program (HRRP), the Hospital-Acquired Condition Present on Admission (HAC-POA) Program, and the Premier Hospital Quality Incentive Demonstration (HQID), generally showed mixed to mostly negative results.
 - Even programmes explicitly designed based on theory or past experience, did not consistently achieve better outcomes.
 - A **notable exception was the Best Practice Tariffs (BPT) programme in the English NHS**, for which multiple studies demonstrated improvements in hip fracture care.
- A **quantitative synthesis** was performed **using vote counting** based on the direction of outcome effects and multilevel regression analysis.
 - Meta-analysis was not feasible due to substantial heterogeneity between studies.
 - Vote counting showed that the **proportion of positive outcomes across programmes ranged from 0 to 1.0, with an overall mean of 0.56**. The mean proportion of positive and statistically significant effects was 0.30.
 - A **multilevel regression analysis** accounted for the nested data structure (outcomes within studies, studies within programmes) and examined **associations between outcome effects and incentive design features**, including type of payment incentive, participation requirements, payment frequency, whether outcomes were incentivised, and study design or risk of bias. Analyses were conducted both at the outcome level (positive effect as dependent variable) and study level (proportion of positive effects as dependent variable).
 - **Threshold-based payment was the only design feature associated with a higher likelihood of positive outcomes.**

- A **realist-informed synthesis and programme theory** examined how intervention components, design features, and contextual factors may influence P4P effectiveness across different contexts. The analysis explored context–mechanism–outcome (CMO) relationships, recognising that P4P programmes are embedded in complex systems where context and baseline conditions shape outcomes.
 - o Overall **study outcome and mechanisms reported by study authors** to explain programme effects were extracted and combined with contextual factors (e.g. type of incentive, baseline payment system) into generalised CMO configurations and an **overarching programme theory highlighting pathways with potentially beneficial respectively detrimental influence**.
 - o **Factors shaping the beneficial pathways** included co-creation and involvement of stakeholders, simple and meaningful incentives large enough to support engagement in change, evidence-informed targets demonstrably linked to outcomes and controllable through managerial and clinical activity, and sufficient room for improvement. Besides the incentives themselves, a system ready to act on them supported by adequate organisational infrastructure, and complementary policies (e.g. public reporting), were also important. Under the right circumstances, this could trigger both a willingness and ability for sustained quality improvement.
 - o Poorly developed financial incentives with a complex design, being perceived as too small to trigger real change, or with a high baseline were largely seen as ineffective. Top-down planning often resulted in targets perceived as disconnected from meaningful outcomes, and targets with a high baseline were thought to be subject to ceiling effects for quality improvement. Limited organisational capacity or external constraints could lead to gaming, selective effort, and other unintended consequences. Positive effects could diminish over time due to ceiling effects unless incentives evolved.
 - o A **programme theory alignment score** was developed to compare programme-level outcomes with the programme theory.
 - Reported mechanisms were classified as beneficial or detrimental, aggregated into net study scores, and combined into mean programme-level scores. Programmes were classified in four categories (beneficial pathway only; net positive mixed pathway; net negative mixed pathway; and negative pathway only) and compared with overall study outcome classifications (beneficial, detrimental, mixed, or null).
 - **Programme designs generally aligned poorly with the factors identified as beneficial in the narrative synthesis.** On a theoretical scale of –12 to +12, the highest observed score was 5.7 for the BPT programme. However, poor pathway scores were not consistently associated with poor outcomes, nor were positive outcomes confined to programmes with better pathway scores.
- A summary of subgroup findings explored potential **equity impacts** of P4P programmes.
 - o **Only 15 studies (16%) included equity-relevant subgroup analysis**, limited to place of residence, race/ethnicity, and socio-economic status.
 - o **Findings were mixed**, showing no differential effects, pro-equity effects, or anti-equity effects.
- A narrative synthesis of **cost and cost-effectiveness considerations showed that very few studies reported such outcomes**, none constituting a full cost-effectiveness analysis. **Findings were mixed**, and further study is necessary to clarify whether P4P programmes represent value for money.



3.1 Introduction

Although pay-for-performance (P4P) programmes initially generated considerable enthusiasm, evidence on their effectiveness in improving hospital care quality remains mixed and inconclusive. Questions have also been raised regarding cost-effectiveness and whether the resources invested in P4P represent value for money. In addition, P4P programmes show substantial variation in objectives, design, implementation, and context, and are best understood as complex combinations of design choices that may influence their effectiveness. It is therefore important to assess whether, when, and under which conditions hospital P4P programmes lead to meaningful improvements in quality of care.

Against this background, this chapter presents the findings of the systematic literature review on the effectiveness of P4P programmes in acute hospitals, as well as insights into their design, implementation, and evaluation. The review addressed three research questions:

- RQ1: What is the effect of P4P for acute care hospitals on the quality of care (effectiveness, safety, people-centredness, appropriateness of care, continuity of care), describing both intended and unintended consequences? ^d
- RQ2: Which intervention components, design features or contextual factors are associated with greater or lesser impact of P4P programmes in acute care hospitals?
- RQ3: Are there specific subgroup effects that have implications for equity?

3.2 Methods

We followed a pre-published protocol (PROSPERO: CRD420251178733). We also followed the PRISMA reporting guidelines for systematic reviews, with completed checklist available in Appendix 1 in [Supplement 1](#) to the Scientific Report.

3.2.1 Information sources

We searched the following databases, from 2008 up to when the searches were conducted (October 2025): Ovid MEDLINE, Embase, EconLit, Scopus, and the Cochrane Database of Systematic Reviews. We additionally searched the Bielefeld Academic Search Engine (BASE) database for grey literature.

3.2.2 Search Strategy

We adapted the search strategy used previously by Mathes et al., 2019.³¹ The search strategy included three main blocks of synonyms for, (i) the intervention (e.g., “pay-for-performance”, “pay-for-quality”, “value-based payment”, “results-based financing”); (ii) the acute hospital setting (e.g., “hospital”, “acute care”, “inpatient care”, “emergency department”); and, (iii) for study designs of interest (quantitative studies with a comparator or pre-post intervention trend period). A final block of terms was used to exclude report types which were listed in our exclusion criteria (e.g. editorials). For each database for which it was possible, these keywords were matched up to the relevant standardised keywords specific to that database (e.g. MeSH terms). For full search strategies, see Appendix 2 in [Supplement 1](#) to the Scientific Report.

We included systematic reviews within our initial search and these were used for a backward citation search of the reference lists of any retrieved relevant systematic reviews, which we performed in Scopus (see Appendix 3 in [Supplement 1](#) to the Scientific Report).

^d The research question was slightly refined from the registered protocol, where 'quality of care' replaced 'different quality domains'. This change was purely intended to improve clarity; the quality domains listed in parentheses remain unchanged and serve as a conceptual delineation of what is understood by quality in this review. The scope of eligible studies and outcomes was not affected.

3.2.3 Eligibility criteria

We followed the following in/exclusion criteria when screening studies:

3.2.3.1 Population

INCLUDED

Studies conducted in acute care hospital settings in high-income countries [as defined by the World Bank](#). Acute care hospitals were understood to include general hospitals and university hospitals, and services such as inpatient care, day care, and emergency care, with no exclusions made based on hospital size or national classification differences.

EXCLUDED

Studies focused exclusively on non-acute care settings (e.g. psychiatric hospitals, mental health care, primary care, long-term care) were excluded. Studies conducted exclusively in low- and middle-income countries ([as defined by the World Bank](#)) were also excluded. For studies involving a mix of acute care and other types of hospitals (e.g. psychiatric hospitals, see above), inclusion was permitted if acute care hospitals clearly represented the majority of the study sample or if findings for acute care hospitals could be clearly disaggregated. Similarly, studies involving mixed care settings or countries were only excluded if relevant data for acute care hospitals in high-income countries could not be clearly extracted.

3.2.3.2 Intervention

INCLUDED

Programmes explicitly involving P4P, defined as policies (e.g. laws, rules, financial or administrative mechanisms) introduced by governments, insurers, or provider organisations that aimed to improve the quality of care through financial incentives. These may target structure, process, or outcome indicators and may reward achievement, improvement, or relative performance. Both positive (bonus) and negative (penalty) incentives were eligible.

Target-based payment schemes, where payment is granted only if a minimum threshold is met (e.g. at least 90% of eligible surgical patients receive prophylactic antibiotics within one hour before incision), were included if the targets were based on evidence-based guidelines and explicitly aimed to improve care quality. Despite involving volume-based thresholds, such schemes differ from traditional activity-based payment by explicitly linking payment to recognised quality standards.

EXCLUDED

Patient-directed incentives (e.g. financial rewards for lifestyle change or preventive behaviours) were excluded. Interventions with implicit financial mechanisms that may influence quality but were not explicitly intended to improve it (e.g. hospital base payment systems like fee-for-service, capitation, payments based on diagnosis-related groups, or bundled payments) were excluded.

3.2.3.3 Comparison

INCLUDED

Eligible studies included a comparator/control group to standard practice, i.e. care delivered without the above-described P4P intervention (though other quality improvement strategies such as public reporting or accreditation may be in place), or an alternative P4P/P4Q design with different features, or a before-after comparison.



3.2.3.4 Outcomes

INCLUDED

At least one objective process (or intermediate) outcome measure must have been reported on one of the quality dimensions (effectiveness, safety, people-centredness, appropriateness of care, continuity of care). Studies reporting unintended consequences (e.g. gaming, risk selection, spillover to non-incentivised areas) were also eligible.

EXCLUDED

Subjective or qualitative outcomes (e.g. patient satisfaction, perceptions) were excluded unless assessed using standardised, validated instruments. Studies focused only on cost containment, productivity, or volume increases, without a primary quality objective evaluation reported, were excluded.

3.2.3.5 Study designs

Both randomised and nonrandomised study types were included.

INCLUDED

Primary studies using a comparator group (e.g. hospitals with vs. without incentives) or a pre-post design (temporal comparison before and after intervention, i.e. with more than one time point in each period) were included. Any length of follow-up time was included.

EXCLUDED

Studies without a control group or temporal trend analysis were excluded. Also excluded were narrative reviews, editorials, letters, case reports, qualitative studies, conference abstracts, posters, and presentations.

3.2.3.6 Study types

Both primary studies and systematic reviews were included. While the analysis focused on primary studies exclusively, relevant systematic reviews were used to identify potential additional primary studies. We reviewed and included only the relevant primary studies rather than results from the review itself to avoid risks of double-counting.

3.2.3.7 Language restrictions

Only documents in English were included in the review.

3.2.3.8 Time frame

We included only studies published since 2008, to follow on from the systematic review by Van Herck et al. (2010).⁴²

3.2.4 Study records

3.2.4.1 Data management

Abstracts of all records and full-text versions of included papers were downloaded electronically and managed in EndNote for initial processing before being uploaded to Covidence^e.

3.2.4.2 Selection process

The title/abstract and full-text screening was supported by Covidence. All citations from the database searches were uploaded to this platform, which de-duplicated. Two authors (Stokes and Kristensen) first independently applied the in/exclusion criteria to 10% of the initial search results to assess interrater agreement. This initial alignment process acted to test agreement, but most importantly to identify and discuss specific disagreement cases to further calibrate the remaining screening. All outstanding titles/abstracts were split between two authors (Stokes and Kristensen) for individual review. Any doubts about inclusion were flagged and discussed. Subsequently, all full texts reviewed were independently double-screened with any disagreements discussed until consensus was reached. The study selection process, including reasons for exclusion at the full-text stage, was documented in a PRISMA flowchart.

3.2.5 Data extraction

The extraction process was supported by the AI-supported research tool Elicit^f. The data extraction form was piloted before being applied across studies. All data extracted by Elicit were also manually verified by a member of the research team, with any necessary amendments made. Data were extracted from the full text and, where necessary, from accompanying supplementary material. Data items included:

- Full citation of the source.
- Country, region, and/or insurance fund in which the programme was implemented.
- Name or title of the P4P/P4Q programme.
- Study design (e.g. controlled trial, time series, difference-in-differences, before-after comparison).
- Sample size and characteristics (e.g. number of hospitals, type of included hospitals, study period, number of observations, population subgroups (clinical and/or socioeconomic)).
- Description of the P4P intervention:
 - o Financial incentive: incentive type, incentive size and structure, frequency of incentive payment.
 - o Quality dimensions covered.
 - o Indicators (payment triggers) in the P4P/P4Q programme: name and description, type of indicator (structure, process, outcome), corresponding quality dimension.
 - o Participation: voluntary or mandatory.
- Baseline payment system described using the Jegers et al. (2002) classification.⁶⁹
- Any reported context-mechanism-outcome (CMO) configurations related to the eligible outcome(s) reported. Since CMOs are aggregations across studies, for each individual paper, this

^e Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.

^f Elicit: The AI Research Assistant; Available at <https://elicit.com>



involved extracting any potential mechanisms mentioned by study authors for explaining the results obtained.

- Reported unintended consequences (e.g. gaming, risk selection, spillover to non-incentivised areas).
- Presence of co-interventions (e.g. public reporting, audit and feedback).
- Treatment effects (and measures of unintended effects) for any eligible outcome evaluated in the study (see eligibility criteria) by indicator including whether these were incentivised or non-incentivised. By incentivised, we understood direct correspondence between the payment trigger and the outcome.
- Reported cost-effectiveness or cost-consequences of intervention.
- Differential impacts reported across population sub-groups (using the PROGRESS-PLUS categories – an acronym used to identify characteristics that stratify health opportunities and outcomes used by the Cochrane group, including Place, Race, Occupation, Gender, Religion, Education, Socioeconomic status, Social capital, personal characteristics associated with discrimination, features of relationships and time-dependent relationships and any clinical sub-groups differentiated).

The full data extraction is available in [Supplement 2](#) to the Scientific Report.

3.2.6 Risk of bias in individual studies

We assessed risk of bias for each study using the Cochrane EPOC Risk of Bias Tool^h.

3.2.7 Data synthesis

The data synthesis consisted of (i) a descriptive summary of results by programme; (ii) a quantitative synthesis using multilevel regression analysis; (iii) a narrative synthesis of the evidence, a programme theory built from reported CMO configurations; (iv) a summary of findings by subgroup to emphasise potential equity impacts; and (v) a narrative synthesis of cost(-effectiveness) studies.

3.2.7.1 Summary of results by programme

We began by describing the geographical distribution of programmes. Many studies were aimed at evaluating the same programmes. For this reason, we initially summarised the overall findings descriptively by programme. In doing so, we began by consolidating the programme descriptions from across individual studies evaluating the same programme to ensure that the consolidated descriptions were as detailed as possible. When programme design changed over time, we recorded any design element used across time in the table and described the development over time in the programme descriptions. Next, we summarised the results programme by programme for the programmes evaluated by more than one study. In doing so, we prioritised findings from the most recent studies, studies spanning longer periods, employed more robust designs (e.g. CBA over ITS) and aimed for comprehensive coverage of outcomes. We focused on whether programmes achieved incentivised or non-incentivised outcomes, and on unintended consequences. A few studies evaluated interaction effects between programmes, and these were discussed separately. For programmes evaluated by a single study only, we summarised findings according to whether they were effective in improving incentivised or non-incentivised outcomes or were associated with unintended consequences. Throughout this section, effects described as ‘significant’ refer to statistical significance at the 5% level

^g <https://methods.cochrane.org/equity/projects/evidence-equity/progress-plus>

^h https://epoc.cochrane.org/sites/epoc.cochrane.org/files/uploads/Resources-for-authors2017/suggested_risk_of_bias_criteria_for_epoc_reviews.pdf

unless otherwise stated. Where results are described without qualification, it was statistically significant at this level. Where a cited result did not reach statistical significance, we state this explicitly.

3.2.7.2 *Effect of P4P for acute care hospitals on the quality of care (Quantitative synthesis)*

As different outcome measures are used across studies, a meta-analysis was not possible. Instead, the quantitative synthesis of the evidence followed a vote-counting approach.⁷⁰

DATA FOR THE QUANTITATIVE SYNTHESIS

For each study, we constructed binary variables classifying each outcome evaluated by the study according to whether the outcome was incentivised or not, whether the findings indicated a positive effect of P4P (e.g. lower mortality, better clinical care processes) and whether the findings were statistically significant at the 5% level. For each study, if more than one outcome was reported, we also indicated a main outcome selected according to the following algorithm:

1. Choose the most global measure, e.g. overall effect, rather than subgroup effects.
2. Choose the most adjusted model reported, if models with different combinations of covariates are presented.
3. If there are multiple endpoints (e.g., medium- and long-term effects), select the long-term endpoint.
4. For an ITS, the level change should be selected rather than the slope change.
5. If choice between structure, process and outcome measures, choose outcome.
6. After above, if multiple potential candidates remained available within any given outcome category (effectiveness, safety, people centredness, appropriateness of care, continuity of care), randomly select.

In this part of the analysis, we excluded study outcomes that evaluated unintended effects and studies that evaluated multiple programmes.

We also created binary variables of study-level characteristics, including the type of payment incentive (bonus/reward, penalty/withholding, tournament/relative performance ranking, threshold-based payment), programme participation requirements (mandatory, voluntary, opt-out), payment frequency, whether the outcome was incentivised, and study design (ITS, CBA, Other). As an alternative to the study-design variable, we categorised studies by risk of bias: low (0 high-risk flags), moderate (1–2 high-risk flags), or high (3 or more high-risk flags).

We created several derived variables for analysis. First, for payment frequency, we created both an ungrouped version (consolidating all 'not specified' categories into 'unknown' whilst retaining distinct categories such as daily, monthly, quarterly, annually, ongoing, and varying) and a grouped version with five categories: high frequency (daily, monthly, ongoing), medium frequency (quarterly, semi-annual), low frequency (annually or longer), or variable/unknown. Second, we generated a combined outcome variable indicating whether an effect was both positive and statistically significant for use in sensitivity analyses. Third, for the study-level analyses, we calculated the proportion of outcomes showing positive effects and the proportion showing positive and statistically significant effects for each study.

ANALYTICAL APPROACH

Our analysis employed vote counting based on the direction of effect, which is recognised as an appropriate synthesis method when effect sizes cannot be reliably calculated or pooled across heterogeneous studies.



The Cochrane guidelines suggest that vote-counting analyses should only consider the direction of the effect, not statistical significance, to avoid the situation where underpowered studies not showing an effect are coded as not having a beneficial effect.⁷¹ However, as the majority of studies included in the review are based on large samples of sufficient power, as a robustness test, we repeated the analyses using an outcome requiring statistical significance at the 5% level to be coded as a positive effect.

Some programmes (e.g. Hospital Readmissions Reduction Program) were evaluated by multiple independent studies. This means our data has a three-level hierarchical structure with outcomes (level 1) nested within studies (level 2) nested within programmes (level 3). This created non-independence between observations, requiring clustering at the programme level to be taken into account to ensure valid inference about P4P programmes.⁷² We therefore extended the vote-counting approach by using multilevel regression models to account for the nested structure of our data and to examine associations between incentive design features and outcome effects.

Our primary analysis conducted at the outcome-level used a three-level mixed-effects logistic regression model with positive effect as the dependent variable. The three-level structure accounted for outcomes nested within studies, and studies nested within programmes, with random intercepts at both the study and programme levels. We chose logistic regression because our outcome is binary, which makes the linear probability model potentially inadequate.

The model included the following predictors: dummy variables for bonus/reward payments, penalty/payment withholding, tournament/relative performance ranking, and threshold-based payment an interaction term between bonus and penalty to capture potential complementarity or substitutability effects, payment frequency (grouped into five categories with variable/unknown as the reference), whether participation was voluntary or not, whether the outcome was incentivised, and study design. We calculated average marginal effects from the logistic model to facilitate interpretation as percentage point changes in the probability of a positive effect.

To complement the outcome-level analysis and ensure that each study was weighted equally regardless of the number of reported outcomes, we conducted a parallel study-level analysis. Although there is a concern that this method is too conservative, overestimating standard errors,⁷³ we included this analysis to allow comparison with previous studies that have used a similar approach.^{74, 75} We thus aggregated outcomes to the study level and used the proportion of positive outcomes as the dependent variable. This continuous variable (ranging from 0 to 1) was analysed using a two-level mixed-effects linear regression model with studies (level 1) nested within programmes (level 2). This specification accounts for the non-independence of studies evaluating the same programme through a random intercept at the programme level.

The study-level model included the same predictors as the outcome-level model.

For all OLS models, we used cluster-robust standard errors (clustered at the study level for outcome-level analysis and at the programme level for study-level analysis) to account for within-cluster correlation. For mixed-effects models, the random-effects structure explicitly modelled the correlation within clusters (study or programme), so additional clustering adjustments were not required.

All analyses were conducted in Stata 19.0. Statistical significance was assessed at the 0.10, 0.05, and 0.01 levels.

SENSITIVITY AND ROBUSTNESS ANALYSES

We conducted extensive sensitivity analyses to assess the robustness of our findings across different modelling choices and specifications.

We re-estimated all outcome-level models using positive and statistically significant effect as the dependent variable (rather than positive effect alone) to assess whether requiring statistical significance altered conclusions about incentive design features. In the study level-analyses, we

examined both the proportion of positive effects and the proportion of positive and statistically significant effects to parallel the outcome-level sensitivity analyses.

We restricted the analysis to main outcomes only (excluding secondary outcomes) to determine whether findings differed when focusing on the primary outcomes of the study.

We compared the grouped payment frequency specification (five categories) to an ungrouped specification retaining all original frequency categories to assess whether the level of aggregation affected conclusions.

As an alternative to the study design variable, we tested using the categorical variable classifying studies by their risk of bias level.

We compared our primary three-level logistic model to a two-level mixed-effects model (excluding the programme level) to assess whether the programme-level variance was substantively important.

3.2.7.3 *Intervention components, design features or contextual factors and P4P effectiveness (Programme theory)*

We drew on a realist-informed approach for our narrative synthesis. The realist approach is born from the idea that studies should not only indicate whether an intervention works (on average) or not but should highlight 'what works, how, in which contexts, and for whom'.⁷⁶ In essence, it draws attention to the fact that interventions are embedded in complex social systems where multiple elements interact, that context matters, and different outcomes might emerge depending on the baseline system state.

CODING OF MECHANISMS AND STUDY OUTCOMES

To implement this approach, we began with our extracted mechanisms from the text, the explanations that the study authors themselves gave for explaining their presented results (explicitly or implicitly, the latter only coded where very clearly implied). One researcher then conducted a thematic coding exercise, inductively coding common themes that emerged from these stated mechanisms across studies.⁷⁷ These codes were discussed with co-authors. Similarly, we coded a brief summary of the overall study findings outlined in the abstract, which we categorised as overall 'beneficial' (where favouring P4P, e.g. 'substantial decline in hospital-acquired conditions'), 'detrimental' (where showing negative impacts of P4P, e.g. 'declining process outcomes and patient satisfaction'), 'mixed' (showing some beneficial/detrimental impacts, e.g., 'positive spillovers for non-targeted patients in same specialty, negative spillover to other specialties'), or 'null' (showing no clear direction of effect, e.g. 'little or no impact on quality improvement'). Any study highlighting a mix of beneficial/detrimental effects as well as null were coded as beneficial/detrimental overall. Only studies highlighting overall null conclusions were coded as null overall.

DEVELOPMENT OF THE PROGRAMME THEORY

We prepared descriptive statistics, showing the most commonly recorded mechanisms across all studies, and then by our categorised overall study message (beneficial/detrimental/mixed/null). We then used these descriptive statistics, combined with the other contextual variables we extracted (e.g., country context, baseline payment system, type of financial incentive) to iteratively prepare generalised context-mechanism-outcome (CMO) configurations as a mid-range theory – a theory that bridges the gap between broad, abstract grand theories and specific empirical observations.^{76, 78} Context here refers to the conditions in which the intervention operates, Mechanism refers to the underlying processes triggered by the intervention that generate change, and Outcome refers to the resulting intended/unintended effects. All data informing these CMOs came from re-use of the fields described for extraction above. We first grouped these CMO-related variables into common higher-level themes, and then ordered them according to the logical flow described within studies (e.g. those factors relating to how the incentive was initially designed, then the specifics of the design of the incentive itself, then the impacts that follow from these specific design choices). We presented these



as a narrative, and using a diagram of the programme theory (utilising kumu.io as software for the latter). The diagram highlighted both factors described by study authors as more beneficial to positive outcomes, and those described as potentially more detrimental.

PROGRAMME THEORY ALIGNMENT SCORE

To better compare the programme-level quantitative findings to our programme theory, we created for each study a programme theory alignment score. Since the programme theory is at a different level of abstraction compared to the mechanism coding, there was not a 1:1 lookup between the mechanisms we coded and the pathways described as more beneficial compared to detrimental in the programme theory. Nevertheless, we classified those mechanisms that did have clear overlap with boxes in the final programme theory (e.g. large incentive representing an attribute of 'well-designed financial incentive', or low baseline being the opposite of 'ceiling/diminishing returns' which was described in the detrimental pathway) and assigned one point for each mechanism present to a beneficial pathway score or a detrimental pathway score (12 classifiable mechanisms per pathway). We took the net of these as a total score. Any mechanism we could not clearly classify were not scored (n=11 coded mechanisms). These mechanisms we recorded as more neutral were predominantly contextual factors or mechanisms which could cut either way, such as loss aversion (see [Supplement 2](#) to the Scientific Report for each mechanism recorded by study, together with the relation of each unique mechanism to the final programme theory and the reasoning for classifying the mechanisms as a beneficial / detrimental / not scored decision). We summarised these by programme in a table, taking the mean score for each by aggregating the relevant studies together. For presentation purposes, programmes were classified into one of four categories based on their mean scores: beneficial pathway only (positive net score, no detrimental pathways recorded); net positive mixed pathway (positive net score, detrimental pathways also present); net negative mixed pathway (negative or zero net score, beneficial pathways present); and negative pathway only (negative or zero net score, no beneficial pathways recorded). Studies evaluating multiple programmes were excluded.

Importantly, a zero score might indicate missing mechanisms discussion or under-reporting (where there was nothing to score), or equally a cancelling out of beneficial and detrimental pathways. Similarly, a positive or negative net score should be seen as indicative of the mechanisms discussed only, subject to reporting bias, rather than a comprehensive overview of each programme.

Finally, we combined the pathway classification and the outcome breakdown into a classification matrix, cross-tabulating the four pathway categories against three observed outcome categories: beneficial only (all evaluated studies reporting beneficial effects), detrimental only (all studies reporting detrimental effects), and mixed/null (studies reporting a combination of outcomes, or no outcome classification recorded). Within each cell, programmes were ordered by mean net score from lowest to highest (left to right), and bubble size reflected the number of studies evaluating that programme.

3.2.7.4 Subgroup effects and equity

To summarise potential equity effects, we narratively summarised those studies reporting any of the extracted subgroup effects (by PROGRESS-PLUS categories).

3.2.7.5 Cost or cost-effectiveness considerations

We identified very few studies addressing cost or cost-effectiveness considerations, so we also summarised these narratively and in tabular form.

3.3 Results

3.3.1 Overview of selected studies

3.3.1.1 Search results

Details of the search results are presented in Figure 2. The systematic search was conducted across five electronic databases yielding a total of 6 380 records. An additional 1 031 references were identified through citation searching of identified reviews (n=999) and the grey literature search (n=32).

Following the removal of 2 679 duplicate records, 4 732 unique records underwent title and abstract screening. Screening of the initial 10% alignment sample led to a Cohen's kappa of 0.66, commonly interpreted as moderate to substantial agreement. All initial disagreements were discussed between co-authors and generalised learning incorporated into the remaining screening. Of these, 4 468 were excluded as not meeting the inclusion criteria, leaving 264 records for full-text retrieval. All 264 full-text articles were successfully obtained and assessed for eligibility.

Upon full-text review, 174 studies were excluded (see Appendix 4 in [Supplement 1](#) to the Scientific Report). Exclusions included systematic reviews that were used for additional citation searching (see Appendix 3 in [Supplement 1](#) to the Scientific Report). A total of 90 studies met all eligibility criteria and were included in the final synthesis.

3.3.1.2 Quality appraisal

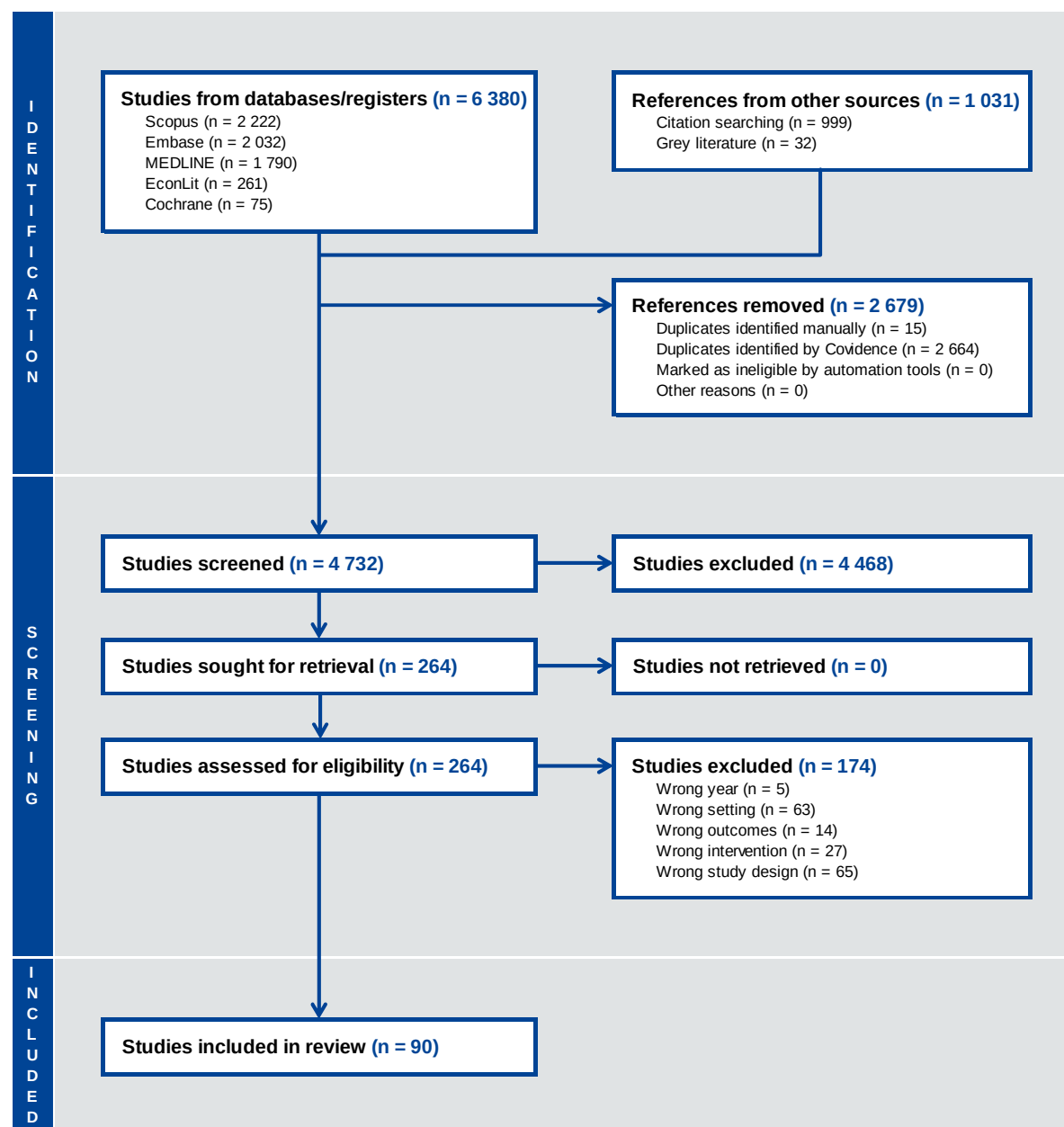
We appraised 36 studies using the Cochrane EPOC interrupted time series domains. The table detailing results by study is available in Appendix 5.1 in [Supplement 1](#) to the Scientific Report. The majority of domains were judged as low risk of bias for most studies. Notable exceptions were the 'intervention independent of other changes' domain, where many studies recognised other policies or national secular trends that might have driven results (83% judged high risk). There was also systematic underreporting of missing data, leading to majority (64%) unclear risk recorded for that domain. Over a third (42%) of these studies were also generally of lesser quality, with a high risk categorisation recorded in the 'other risks of bias' domain.

We appraised 54 studies using the Cochrane EPOC domains for studies with a separate control group. The table detailing results by study is available in Appendix 5.2 in [Supplement 1](#) to the Scientific Report. Since there were no randomised trials included, following the EPOC guidelines, all studies were automatically recorded as high risk in the 'random sequence generation' and 'allocation concealment' domains. Apart from these domains, the majority of domains were again judged as low risk of bias for most studies, with many high-quality quasi-experimental (controlled before-after) studies included. Like the ITS studies, there was a systematic underreporting of missing data leading to majority (80%) recorded as unclear risk in the 'incomplete outcome data' domain. A similar proportion (31%) of studies also used a weaker design and were judged as relatively high risk under the 'other risks of bias' domain.

We did not exclude any study from the subsequent analyses due to risk of bias.



Figure 2 – PRISMA flow diagram



3.3.2 Summary of results by programme

Across the 90 studies, 31 unique P4P programmes were identified (Table 2). The vast majority of studies (66) are conducted in a US setting, with England being the second largest contributor with 10 studies. The remaining 14 studies come from 8 different countries, with Australia and Canada each contributing 3, Japan and Taiwan each contributing 2, and France, Italy, Germany, Korea and each contributing one study to the evidence base. Four studies explore programme interaction effects and these are all US based; one additional examined two US-based programmes but did not assess interaction effects. The programmes analysed by multiple studies are all from either the US Medicare setting, the English National Health Service, or Queensland, Australia.

Incentive design varies considerably between programmes. For some programmes, the design also varies over time, and Table 2 reports design elements ever used by a programme, even if not used



throughout the full period. We elaborate on this within programme variation in the programme descriptions below.

Bonus-only designs are most common: 17 programmes use bonuses without penalties. Five programmes use penalty-only mechanisms, withholding payment or reducing reimbursement for substandard performance without offering rewards for high performance. Nine programmes use a combination of bonuses and penalties.

Tournament-based designs, where payment depends on rank-order performance relative to peers, are used in eight programmes, of which seven used tournament incentives exclusively, and one (HVBP) combines tournament and thresholds.

Threshold-based designs, where providers meeting a specified performance level receive payments, are used in 12 programmes.

The frequency of incentive payments varies substantially. Annual or longer payment intervals are most common, applied in 15 programmes. Six programmes provide ongoing incentives, typically through case-level price adjustments (such as Best Practice Tariffs), where the incentive is embedded in each reimbursement rather than paid as a discrete sum.

Programmes were roughly evenly divided between mandatory participation (17 programmes, 55%) and voluntary participation (14 programmes, 45%). Six out of 14 voluntary programmes are bonus-only and threshold-based, while programmes employing penalties are most frequently mandatory.



Table 2 – Programme characteristics

Programme	Country	Bonus	Penalty	Tournament	Threshold based	Frequency	Participation	Number of Studies
<i>National HAC financial penalty policy</i>	Australia	No	Yes	No	No	Not specified	Mandatory	1 ⁷⁹
<i>Queensland Quality Improvement Payments (QIP)</i>	Australia	Yes	No	No	Yes	Semi-annual	Mandatory	2 ^{80, 81}
<i>Pay for Results Program</i>	Canada	Yes	No	No	Yes	Annually or longer	Voluntary	1 ⁸²
<i>Ontario Quality Improvement Plan (OQIP)</i>	Canada	Yes	No	No	No	Annually or longer	Voluntary	1 ⁸³
<i>Unnamed P4P Programme</i>	Canada	Yes	No	No	No	Ongoing	Voluntary	1 ⁸⁴
<i>Advancing Quality (AQ)</i>	England	Yes	Yes	Yes	No	Varying	Mandatory	5 ⁸⁵⁻⁸⁹
<i>Best Practice Tariff (BPT)</i>	England	Yes	No	No	Yes	Ongoing	Mandatory	3 ⁹⁰⁻⁹²
<i>Prescribed Specialised Services Commissioning for Quality and Innovation (PSS CQUIN)</i>	England	No	Yes	No	Yes	Varying	Voluntary	2 ^{93, 94}
<i>Incitation financière à l'amélioration de la qualité (IFAQ) – pilot phase</i>	France	Yes	No	Yes	No	Annually or longer	Voluntary	1 ⁹⁵
<i>Ophthalmology P4P</i>	Germany	Yes	Yes	No	Yes	Ongoing	Voluntary	1 ⁹⁶
<i>Unnamed P4P Programme</i>	Italy	Yes	Yes	Yes	No	Annually or longer	Mandatory	1 ⁹⁷
<i>Additional Charge for Emergency Reduction and Internal Fixation or Hemiarthroplasty for Hip Fracture (ACERIFH)</i>	Japan	Yes	No	No	No	Not specified	Mandatory	1 ⁹⁸
<i>Unnamed P4P Programme</i>	Japan	Yes	No	No	No	Ongoing	Voluntary	1 ⁹⁹
<i>Korean Value Incentive Program (VIP)</i>	Korea	Yes	Yes	No	Yes	Ongoing	Mandatory	1 ¹⁰⁰
<i>Breast Cancer Pay-for-Performance Program (BC-P4P)</i>	Taiwan	Yes	Yes	No	No	Annually or longer	Voluntary	1 ¹⁰¹
<i>Emergency Department Quality Improvement Program (EDQIP)</i>	Taiwan	Yes	No	No	Yes	Not specified	Voluntary	1 ¹⁰²

Programme	Country	Bonus	Penalty	Tournament	Threshold based	Frequency	Participation	Number of Studies
<i>Delivery System Reform Incentive Payment (DSRIP)</i>	United States	Yes	No	No	No	Not specified	Voluntary	1 ¹⁰³
<i>Hospital-Acquired Condition Present on Admission (HAC-POA)</i>	United States	No	Yes	No	No	Ongoing	Mandatory	12 ¹⁰⁴⁻¹¹⁵
<i>Hospital-Acquired Condition Reduction Program (HACRP)</i>	United States	No	Yes	Yes	No	Annually or longer	Mandatory	5 ^{*116-120}
<i>Housestaff Incentive Program</i>	United States	Yes	No	No	Yes	Annually or longer	Voluntary	1 ¹²¹
<i>Premier Hospital Quality Incentive Demonstration (HQID)</i>	United States	Yes	Yes	Yes	No	Annually or longer	Voluntary	11 ^{*122-132}
<i>Hospital Readmissions Reduction Program (HRRP)</i>	United States	No	Yes	No	Yes	Annually or longer	Mandatory	23 ^{*119, 133-154}
<i>Hospital Value-Based Purchasing (HVBP)</i>	United States	Yes	Yes	Yes	Yes	Annually or longer	Mandatory	11 ^{*118, 120, 132, 154-161}
<i>ICU P4P</i>	United States	Yes	No	No	No	Annually or longer	Mandatory	1 ¹⁶²
<i>Maryland Hospital-Acquired Conditions Program</i>	United States	Yes	Yes	Yes	No	Annually or longer	Mandatory	1 ^{*163}
<i>Maryland Quality-Based Reimbursement Program</i>	United States	Yes	Yes	Yes	No	Annually or longer	Mandatory	1 ^{*163}
<i>MassHealth P4P</i>	United States	Yes	No	No	No	Not specified	Mandatory	1 ¹⁶⁴
<i>Performance Award</i>	United States	Yes	No	No	Yes	Annually or longer	Mandatory	1 ¹⁶⁵
<i>Radiology Report Signature Times incentive</i>	United States	Yes	No	No	No	Semi-annual	Voluntary	1 ¹⁶⁶
<i>Thoracic surgery value-based reimbursement</i>	United States	Yes	No	No	Yes	Annually or longer	Voluntary	1 ¹⁶⁷
<i>Unnamed P4P Programme</i>	United States	Yes	No	No	No	Quarterly	Mandatory	1 ¹⁶⁸

Note: * Studies covering multiple programmes are counted under each programme



3.3.2.1 Hospital Readmissions Reduction Program (HRRP)

The Hospital Readmissions Reduction Program (HRRP) is the most extensively evaluated P4P programme, with 23 evaluations included in this review. Introduced by CMS / Medicare in 2012, this mandatory programme penalises hospitals with higher than expected 30-day readmission rates for targeted conditions. The targeted conditions were initially Acute Myocardial Infarction (AMI), heart failure, and pneumonia. The list was later expanded to include chronic obstructive pulmonary disease (COPD), coronary artery bypass graft surgery (CABG), and hip/knee replacement.

An early study by Zuckerman et al., studying the impact until 2015, found readmission rates for targeted conditions declined more sharply than readmission rates for non-targeted conditions, and that decline could not be explained by changes in the use of observation stays (a potential gaming strategy, because such stays do not count as readmissions in the HRRP).¹⁴⁶ This conclusion was supported by Desai et al. who, studying the same period, found that, after the announcement of the HRRP, readmission rates decreased more rapidly for targeted conditions compared to non-targeted conditions in penalised hospitals, compared to both targeted and non-targeted readmissions in non-penalised hospitals, while readmission rates plateaued after the actual implementation of the programme.¹⁴¹

A later study by Gupta but still only covering the period until 2014, found that hospitals' responses to penalties accounted for two-thirds of the decrease in readmissions over this period.¹³⁶ He estimated reductions of around 3.5 percentage points for AMI readmissions (from an 18% baseline) and 2.8 percentage points for pneumonia (from 17%). The study also found an associated decrease in heart attack mortality, and that improvement in quality accounted for half of the decrease in readmissions, suggesting that the remaining decrease could be explained by selective admission of returning patients.

Ziedan and Kaestner questioned the validity of the empirical approaches in earlier studies that found the HRRP to be effective in reducing readmission rates, and instead demonstrated that, using their preferred method (a regression-kink design), the HRRP was not statistically significantly reducing readmission rates.¹³⁴

Gupta et al. studying the impact of HRRP on readmission rates and mortality for heart failure patients until 2014, found that the programme was associated with a decrease in readmission rates, but also found evidence of an associated increase in mortality rates (a potential unintended consequence).¹⁴⁷ This finding is in contrast to Ramaswamy et al., who investigated the impact on both readmission and mortality rates for CABG patients until 2015, finding that a preexisting trend of decreasing readmission rates continued under the HRRP at a slightly steeper rate, although the difference was not demonstrated to be statistically significant and without any evidence of an associated change in mortality.¹⁴³

Studies also examined whether the HRRP generated spillover effects to conditions and outcomes not directly targeted by the programme. Ibrahim et al. found evidence of both anticipatory and positive spillover effects on readmission rates for surgical procedures that were not yet incentivised at the time of the study.¹⁵⁰ Deore et al. found that the HRRP was associated with an increase in the share of patients admitted to a short- or long-term care facility after admission, both for targeted and non-targeted conditions and interpreted this increase as a signal of improved continuity of care.¹⁵¹

Several studies pursued the mechanisms behind the observed reductions in readmission rates. A study by Sabbatini et al. suggested that more than half of the observed reduction in readmission rates following HRRP could be attributable to the reclassification of inpatient admissions as observation stays.¹³³ Popescu et al. found that referrals to skilled nursing facilities increased for pneumonia patients after HRRP, but use of home health care decreased and there was no change in referral patterns for patients with AMI or heart failure.¹⁴⁸ On the other hand, Nathan et al. found no evidence of gaming the 30-day threshold.¹³⁹ Mellor et al. found no evidence of patient selection or gaming in response to HRRP penalties.¹⁵² Hasan et al. similarly found no significant increase in readmissions to

non-index hospitals, suggesting hospitals did not systematically redirect patients to avoid penalty attribution.¹⁵³

In summary, most studies found the HRRP to be associated with reductions in 30-day readmission rates for targeted conditions, with some evidence of positive spillovers to non-targeted procedures, although one study challenged this conclusion on methodological grounds. It is also contested whether reductions can be attributed to genuine improvements in quality or to changes in coding and admission practices following the programme. The mortality evidence is contradictory suggesting the programme was associated with a reduction in mortality for AMI but an increase in mortality for heart failure patients.

3.3.2.2 *Hospital-Acquired Condition Present on Admission (HAC-POA) Program and Hospital-Acquired Condition Reduction Program (HACRP)*

Hospital Acquired Conditions (HACs) are safety-events that can be attributed to inappropriate care or care of inadequate quality. Since 2008, Medicare has implemented incentives to discourage such events. A policy targeting HACs requires the ability to distinguish between diagnoses, e.g. pressure ulcers or infections, that were and were not present upon admission. Prior to 2008, this was not possible in the payment system used by US Medicare. Although pressure ulcers and other safety events were coded, the coding did not indicate whether these events were present on admission or not, and therefore whether they were related to poor quality care. Because such co-morbidities would often be associated with increased DRG payments, payers became concerned that Medicare could be paying hospitals for what could essentially be considered the expression of low-quality care. The earliest programme – the Hospital-Acquired Condition Present on Admission (HAC-POA) – introduced a present-on-admission variable that allowed for that distinction. The policy then eliminated additional payment for specific complications thought to be preventable on a case-by-case basis.

Several studies demonstrated positive effects of the programme for one or more conditions. Gidwani and Bhattacharya found lower incidence of hospital-acquired pulmonary embolism or deep vein thrombosis after hip or knee replacement surgery during the first years of the programme.¹⁰⁷ Thirukumaran et al. found reduced venous thromboembolism after hip and knee replacements until 2013.¹¹⁴ Kim et al., extending the study period until 2017, found evidence of a modest decrease in surgical site infections, LOS (length of stay) and hospital costs for Medicare patients, although these authors found no effect on deep venous thrombosis or mortality.¹⁰⁵

Thirukumaran et al. found reductions in a summary any-or-none indicator and specifically for catheter-associated urinary tract infections in hospitals with a high share of Medicare patients, but no effect on falls.¹¹⁵ Waters et al. studying the impact until 2016 found reductions in hospital-acquired central-line infections and catheter-associated infections, but no effect on pressure ulcers or falls.¹⁰⁹

On the other hand, multiple studies found no effect of the programme. Schuller et al. found no effect on catheter-associated urinary tract infections (CAUTIs) during the first year of the programme.¹¹⁰ A study by Lee et al. found no effect on catheter-associated bloodstream infections or catheter-associated urinary tract infections until 2011.¹¹¹ Kwong et al. (2017) found that until 2013, surgical site infections after orthopaedic procedures declined after the HAC-POA programme was introduced, but not above existing trends.¹⁰⁸ Vaz et al. found no effect on surgical site infections until 2012.¹⁰⁴

A study by Morgan et al. examined whether the HAC-POA programme encouraged gaming in the form of increased diagnostic testing for CAUTI on admission but found no evidence that the programme was associated with increased testing on admission.¹⁰⁶

From 2015, the Hospital-Acquired Condition Reduction Program (HACRP) introduced further mandatory penalties for hospitals with higher-than-expected HAC rates. Hospitals in the worst-performing quartile on measures of hospital-acquired conditions were penalised with a 1% reduction across all Medicare payments.



Rude et al. did not find the HACRP programme to be associated with a decrease in HAC rates in patients undergoing major urologic surgery.¹¹⁷ However, Arntson et al. found that the HACRP programme was associated with statistically significant decreases in HACs targeted by the programme, although this was not associated with greater reductions in mortality rates.¹¹⁶

In summary, the evidence on the HAC-POA programme points to modest improvements in some targeted outcomes, but no consistent effect across all incentivised conditions. The more recent HACRP showed similarly mixed results, with one study finding reductions in targeted HACs and another finding no effect, and neither programme demonstrating associated reductions in mortality.

3.3.2.3 Premier Hospital Quality Incentive Demonstration (HQID)

The Premier Hospital Quality Incentive Demonstration (HQID) was launched in October 2003 by CMS. The programme ran for six years, concluding in September 2009, and played a critical role in informing future incentive design, specifically the Medicare Hospital Value-Based Purchasing (HVBP) programme. It was also the inspiration for the English Advancing Quality programme. HQID was voluntary and available only to hospitals participating in the Premier quality database. It targeted patients with AMI, CHF, pneumonia, CABG and hip and knee replacement, for which quality was measured by a combination of clinical process and outcome measures. From the beginning, the scheme was designed as a tournament-style bonus scheme, with 2% bonus payments per DRG payment for patients with a targeted condition to hospitals performing in the top decile, and 1% bonus payments to hospitals in the second decile. From the third year, the programme mandated 1% and 2% penalties to hospitals in the two last performance deciles.

While evidence suggested that incentivised hospitals improved performance on a composite performance index more than non-incentivised hospitals during the first three years of the programme by year four, the difference began to diminish and was no longer statistically significant.¹³¹ This finding was confirmed by Jha et al., who demonstrated that there was no difference in 30-day mortality between incentivised and non-incentivised hospitals in the long term.¹²⁹ Ryan et al. showed that a 2006 change in HQID incentive design to incentivise performance improvement (rather than simply rewarding high performers) was not associated with greater improvements in composite measures of process quality among low performers.¹²⁵ The analysis was extended by Shih et al. who investigated whether the incentive change improved mortality rates after surgery and also found no difference in mortality rates between incentivised and control hospitals.¹²⁴

A study by Epstein et al. examined whether HQID was associated with reduced access to CABG surgery for high-risk patients (a potential gaming response) but found no evidence to support this concern.¹²⁸

Thus, the several studies that have investigated the impact of the HQID programme, paint an overall picture of limited effectiveness of the programme.

3.3.2.4 Hospital Value Based Purchasing (HVBP)

The Hospital Value-Based Purchasing (HVBP) programme is a mandatory pay-for-performance initiative implemented by CMS from fiscal year 2013. The programme is designed to be budget-neutral. A percentage (increasing from 1% to 2%) of each hospital's base operating DRG-payments is withheld for each fiscal year. The pooled amount is then redistributed to hospitals as incentive payments based on hospitals' quality performance as defined by a clinical process composite, a patient experience composite, and a set of outcome measures. Hospitals may receive back less than, equal to, or more than the amount withheld, depending on their Total Performance Score relative to other hospitals.

Despite being designed on the basis of lessons from the HQID, the HVBP has also not been found to be effective. Ryan et al. found no evidence of improvements in clinical process measures of quality or patient experience in hospitals exposed to HVBP compared to control hospitals until 2015.^{156, 161} The later study by Ryan et al. also examined the impact of HVBP on 30-day mortality, finding no effect.¹⁶¹

This finding was further supported by Figueroa et al. and Banerjee et al. who also found no effect of the programme on mortality covering the period until 2016.^{155, 160} Finally, a study by Papanicolas et al. found no impact of the programme on patient experience.¹⁵⁹

3.3.2.5 *Studies of multiple US Medicare programmes*

Alrawashdeh et al., using data from 2013-2019, studied the combined effect of including hospital acquired *Clostridioides difficile* infection rates in the HVBP and the HACRP from October 2016 and found that the inclusion was associated with an immediate level and subsequent slope decline in *Clostridioides difficile* infection rates.¹²⁰

Bonfrer et al., using data from 2003-2013, demonstrated that there was no difference in process measures or mortality between hospitals that had been operating under P4P since the HQID was introduced, compared to similar hospitals that had only been incentivised during the HVBP programme.¹³²

Hsu et al. found no differential effect of the HACRP and HVBP programmes on healthcare-associated infection rates between safety-net and non-safety-net hospitals, with neither hospital type showing improvements and pre-existing disparities persisting throughout the study period from 2013-2018.¹¹⁸

Waters et al., examining the combined impact of the HRRP, HVBP, and HACRP programmes on both targeted and non-targeted quality and safety indicators, found no consistent improvements between 2007 and 2016.¹⁵⁴

3.3.2.6 *Advancing Quality (AQ)*

The Advancing Quality (AQ) programme was a hospital-based P4P programme introduced in the North West region of the English National Health Service (NHS) from October 2008. The programme was the first large-scale hospital P4P initiative introduced in England and was explicitly modelled on the US Premier HQID, targeting patients with AMI, CHF, pneumonia, CABG and hip and knee replacement. All 24 acute hospital trusts providing emergency care in the North West region participated in AQ. The programme began as a tournament, where, during the first year, hospitals in the top quartile on a composite quality metric received a 4% bonus payment on relevant activity, and those in the second quartile received a 2% bonus. For the next six months, a mixed incentive structure was used, rewarding attainment, improvement, and relative performance. Finally, from year 3 onwards, AQ was incorporated into the national Commissioning for Quality and Innovation (CQUIN) payment framework. This shifted the incentive from a tournament-based bonus system to a penalty-based incentive. Under CQUIN, a fixed proportion of each hospital's expected income was withheld and paid out only if hospitals achieved locally negotiated performance thresholds.

An early study found the AQ programme to be effective in improving in-hospital mortality by approximately 1 percentage point (from a baseline of 22% in the North West) across the incentivised conditions during the first 18 months, although only the impact on pneumonia was statistically significant when the conditions were analysed separately.⁸⁶ Meacock et al. found that AQ incentives were a cost-effective way of improving care.⁸⁵ However, a later study found that, although mortality continued to decrease, the P4P impact on mortality was not sustained in the long term, as mortality reductions were greater in the control group hospitals.⁸⁸

Britteon et al. investigated whether the AQ programme generated spillover effects to non-targeted patient groups.⁸⁷ They found that mortality in incentivised hospitals increased for emergency patients who were not targeted by the programme and were treated by physicians with no exposure to the incentives. On the other hand, mortality improved slightly for non-targeted patients treated by physicians exposed to the incentives, suggesting that hospitals responded to incentives by diverting resources away from non-targeted patients, but that exposed physicians applied the incentivised improvements in care delivery more broadly.



3.3.2.7 *Best Practice Tariffs (BPTs)*

Best Practice Tariffs (BPTs) is a national P4P mechanism within the English NHS. BPTs can be understood as an adjustment of the existing activity-based Payment by Results payment system to incentivise the delivery of high-quality care, by paying more for care delivered according to best practice. The first BPTs were introduced in April 2010 (financial year 2010/11), focusing on four clinical areas, but have expanded considerably since then. The BPT framework does not represent a single approach to incentive design, rather, BPTs reward quality in different ways. Two common ways are top-up payments, conditional on meeting a set of clinical process standards, and differential prices that incentivise shifting activity for example from the inpatient to the outpatient setting.

Three studies evaluated the BPT programme related to hip fracture care, which in summary demonstrated effectiveness in achieving incentivised outcomes.

Metcalfe et al. found improvements in time to operation and LOS after BPT for hip fracture was introduced.⁹⁰ Similarly, an increasing trend in 30-day readmissions was halted, and mortality was reduced in comparison to a Scottish control group. These findings were confirmed by Zogg et al., who compared outcomes in England to outcomes in a US Medicare population.⁹² The mortality reductions estimated by these two studies were of a similar magnitude: approximately 1.6 percentage points at 30 days (baseline ~9–10%) and around 1.7–2.0 pp at 60–90 days, based on difference-in-differences comparisons with Scottish (Metcalfe et al.) and US Medicare (Zogg et al.) control populations.

Grašič et al. demonstrated how these outcome improvements were achieved by documenting the improvements in the care processes incentivised by the scheme.⁹¹

3.3.2.8 *Prescribed Specialised Services Commissioning for Quality and Innovation (PSS CQUIN)*

The Prescribed Specialised Services Commissioning for Quality and Innovation (PSS CQUIN) programme is a subset of the wider CQUIN pay-for-performance programme mentioned above, focusing on prescribed specialised services. These are services for patients with rare and complex conditions. Even more so than BPTs, PSS CQUIN should not be understood as a single approach to incentive design, but rather an overall framework, which links a proportion of hospitals' income to the achievement of quality improvement goals negotiated between commissioners and hospitals, while the incentive mechanisms for reclaiming the funds vary from payments for taking up a specific programme, supporting the necessary structures, to threshold-based payments rewarding the achievement of specific process or outcome goals.¹⁶⁹

Feng et al. that evaluated a subset of ten schemes effective between 2016/2017 and 2018/2019 found that only one of these schemes (in the area of trauma care) had a statistically significant improvement on the targeted outcome (discharge from Adult Critical Care within four hours of being deemed clinically ready).⁹⁴

When testing for differential effects across socioeconomic groups for two PSS CQUIN schemes, Núñez-Elvira et al. found widened inequalities for the trauma care scheme, while there was still no effect identified across either subgroup for a scheme aiming at reducing in-hospital waiting time and LOS for patients requiring urgent coronary bypass grafting.⁹³

3.3.2.9 *Queensland Quality Improvement Payment (QIP)*

The Quality Improvement Payment (QIP) Stroke Unit Care programme was implemented in Queensland, Australia, in 2012. The programme, characterised by being clinician-led and co-designed, aimed to improve stroke unit access in public hospitals in Queensland. Hospitals were paid biannually contingent on meeting performance targets for stroke unit access.

The programme was found to increase access to stroke unit care, and reduced 6-month stroke mortality.⁸⁰ A later study demonstrated that the programme was not associated with increased hospital costs, LOS, or readmission rates for patients with stroke.⁸¹

3.3.2.10 Other programmes

Among the remaining 22 programmes, based on the categorisation of study outcomes described in section 3.2.7.3, eleven were found to be effective, five demonstrated mixed findings, and six were not found to be effective.

A national policy in Australia penalised public hospitals for hospital-acquired complications (HACs) by deducting the additional cost of HACs from hospital funding. Slawomirski et al. found that the policy was associated with a decline in HAC prevalence, although the authors note that the impact of the financial penalty could not be easily isolated from the impact of concurrent policies aiming to improve patient safety in acute hospitals.⁷⁹

The removal of the Korean Value Incentive Programme was found to be associated with an increase in Caesarean-section rates, an outcome previously incentivised by the programme.¹⁰⁰

The Breast Cancer P4P Programme (BC-P4P) in Taiwan provides a bundled per-patient payment based on guideline-recommended treatment and penalises treatment not delivered in accordance with guidelines. In addition, participating hospitals are rewarded for meeting stage-specific survival rate goals. A study found that enrolled patients received higher-quality care and had better outcomes than non-enrolled patients, although selection bias cannot be ruled out.¹⁰¹

The Emergency Department Quality Improvement Programme (EDQIP) in Taiwan introduced financial incentives to improve emergency department performance such as timely thrombolysis, emergency department (ED) overcrowding, and critical care transfers. Chu et al. found that, for patients with stroke, the program was associated with reduced mortality rates, shorter LOS and fewer days in ICU.¹⁰²

The Performance Award Program in Baylor Health Care System (BHCS), in Dallas–Fort Worth, Texas (US), rewarded individual employees for performance on fiscal operating margin, clinical quality, and patient satisfaction. Herrin et al. found improved performance at hospitals exposed to incentives following the introduction of the programme.¹⁶⁵

The Radiology Report Signature Times Incentive (US) enhanced the introduction of new technology in the radiology department by offering a final incentive for signing off on the radiology report, demonstrating improved signing performance even after the incentives were discontinued.¹⁶⁶

A value-based reimbursement incentive for lymph-node staging from the Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative (US) was associated with an initial brief drop in preoperative mediastinal staging rates but then significantly increased. Nodal sampling was not affected.¹⁶⁷

The Maryland Quality-Based Reimbursement Program (US) used bonuses and penalties to incentivise process-based measures of clinical performance, while the Maryland Hospital-Acquired Conditions Program (described in the same paper) incentivises hospitals to reduce HACs, again using a combination of financial bonuses and penalties. Both programmes were associated with significant improvements in performance.¹⁶³

The Housestaff Incentive Program for internal medicine residents at the University of California at San Francisco Medical Center (UCSFMC, US) promised a financial bonus per resident for having 75% of discharge summaries completed on the day of discharge, while relying on collective, rather than individual performance, to determine if all or no residents would receive the bonus. The programme was associated with substantial decreases in the average time to complete discharge summaries, and the quality of the summaries improved.¹²¹

Finally, a P4P programme within the Veterans Health Administration (US) for acute coronary syndrome (ACS), heart failure (HF), and pneumonia (PNU) demonstrated improvements in performance on six of seven process quality measures, which were sustained after the removal of the incentive.¹⁶⁸



Five studies showed mixed effects.

The New Jersey Medicaid Delivery System Reform Incentive Payment (NJ DSRIP, US) was a P4P disease management program focused on outcomes for targeted chronic conditions. The programme was found to reduce asthma admissions and emergency department visits, pneumonia readmissions, and improve alcohol and drug treatment, but showed no effect on readmission rates or follow-up after mental illness hospitalisation.¹⁰³

An emergency department (ED) P4P in Ontario, Canada was associated with modest reductions for several but not all measures of ED LOS with no effect on measures of ED quality, including readmission and mortality rates.⁸²

An ICU P4P for Spontaneous Breathing Trials at the University of Pittsburgh Medical Center (UPMC, US) was a physician-targeted financial incentive for daily completion of spontaneous breathing trials in mechanically ventilated patients. While the programme was associated with an increase in trial completion rates in low-performing intensive care units, the results could in part be explained by higher exclusion rates, casting doubt on the programme's clinical value.¹⁶²

An ICU incentive scheme in Japan, rewarding early rehabilitation, increased the proportion of ICU patients who received early rehabilitation, but had no impact on the ICU LOS, hospital LOS, or the proportion of patients discharged to home.⁹⁹

A P4P programme in Lombardy, Italy introduced financial bonuses and penalties based on ranking on a weighted mean of hospital performance on high-level outcomes, including mortality, readmissions and the number of discharges against medical advice. The scheme was associated with improvements in the number of readmissions, transfers and returns to the surgery room, with no effect on the number of voluntary discharges or mortality.⁹⁷

Finally, six studies from five different countries (Canada, Germany, Japan, US, France) demonstrated no effect of the programmes they evaluated.^{83, 84, 95, 96, 98, 164}

3.3.3 *Association between study outcomes, programme and study characteristics*

3.3.3.1 *Descriptive analysis*

The quantitative analysis excluded studies of unintended effects (n=7), multi-programme studies (n=5) one study with no reported point estimates, and one whose analytical sample was fully nested within another included study. The final sample included 627 outcomes from 76 studies evaluating 29 P4P programmes (Table 3). The number of outcome-study observations per programme ranged from 2 to 128, and the number of studies per programme ranged from 1 to 18.

Twenty programmes were evaluated by a single study, three programmes by two studies, four programmes by 3-9 studies, and two programmes by 10 or more studies.

Across all programmes, the proportion of outcomes showing positive effects ranged from 0 to 1.0 with an overall mean of 0.56 (0.60 at the study level), while the average proportion of outcomes showing positive and statistically significant effects was 0.30 (0.43 at the study level).

Six programmes showed positive (and statistically significant) effects for all reported outcomes, while three programmes showed no positive effects (four, when also considering statistical significance). These were all evaluated by one study only. For 11 of the remaining 20 programmes, the proportion of positive outcomes exceeded 0.5, whereas for 9 programmes it was 0.5 or less. Considering statistically significant positive outcomes, two of the remaining programmes had positive proportions above 0.5, while for the rest, the proportion of positive outcomes was 0.5 or less.

Programme performance varied by evaluation intensity. Overall, 59% (17 of 29) of programmes showed positive effects for more than half of their reported outcomes, compared to 28% (8 of 29) programmes with more than half of outcomes positive and statistically significant. Among programmes evaluated by a single study, 12 of 20 showed a positive effect for more than half of reported outcomes,

with 7 of 20 programmes being statistically significant. Among programmes evaluated by at least 2 studies, 5 of 9 showed positive effects for >0.5 proportion of outcomes, compared to 1 of 9 for positive and statistically significant effects. Among programmes evaluated by at least 3 studies, 4 of 6 showed positive effects for >0.5 proportion of outcomes, while only 1 of 6 had a proportion of positive and statistically significant outcomes for >0.5.

Table 3 – Proportion of positive outcomes by programme

Programme	Proportion positive outcomes	Proportion positive & significant outcomes	Number of outcomes	Number of studies
<i>Emergency Department Quality Improvement Program (EDQIP)</i>	1.000	1.000	4	1
<i>Australian HAC financial penalty policy</i>	1.000	1.000	4	1
<i>Breast Cancer Pay-for-Performance Program (BC-P4P)</i>	1.000	1.000	2	1
<i>Housestaff Incentive Program</i>	1.000	1.000	2	1
<i>Korean Value Incentive Program (VIP)</i>	1.000	1.000	2	1
<i>Radiology Report Signature Times incentive</i>	1.000	1.000	2	1
<i>Pay for Results Program</i>	0.941	0.647	17	1
<i>Best Practice Tariff (BPT)</i>	0.875	0.708	24	3
<i>Incitation financière à l'amélioration de la qualité (IFAQ) - pilot phase</i>	0.750	0.125	8	1
<i>Hospital Readmissions Reduction Program (HRRP)</i>	0.703	0.430	128	18
<i>Ophthalmology P4P</i>	0.692	0.231	13	1
<i>Hospital-Acquired Condition Reduction Program (HACRP)</i>	0.667	0.500	6	2
<i>ICU P4P</i>	0.667	0.417	12	1
<i>Unnamed P4P Programme (Japan)</i>	0.625	0.250	8	1
<i>Delivery System Reform Incentive Payment (DSRIP)</i>	0.600	0.200	5	1
<i>Hospital-Acquired Condition Present on Admission (HAC-POA)</i>	0.567	0.250	60	11
<i>Advancing Quality (AQ)</i>	0.563	0.296	71	4
<i>Performance Award</i>	0.500	0.500	6	1
<i>Queensland Quality Improvement Payments (QIP)</i>	0.500	0.400	10	2
<i>Thoracic surgery value-based reimbursement</i>	0.500	0.500	2	1
<i>Unnamed P4P Programme (United States)</i>	0.500	0.429	14	1
<i>Hospital Value-Based Purchasing (HVBP)</i>	0.447	0.105	38	7



Programme	Proportion positive outcomes	Proportion positive & significant outcomes	Number of outcomes	Number of studies
<i>Prescribed Specialised Services Commissioning for Quality and Innovation (PSS CQUIN)</i>	0.439	0.112	98	2
<i>Unnamed P4P Programme (Italy)</i>	0.400	0.400	10	1
<i>Premier Hospital Quality Incentive Demonstration (HQID)</i>	0.295	0.115	61	7
<i>Additional Charge for Emergency Reduction and Internal Fixation or Hemiarthroplasty for Hip Fracture (ACERIFH)</i>	0.143	0.000	14	1
<i>MassHealth P4P</i>	0.000	0.000	2	1
<i>Ontario Quality Improvement Plan (OQIP)</i>	0.000	0.000	2	1
<i>Unnamed P4P Programme (Canada)</i>	0.000	0.000	2	1
Summary statistics				
<i>Outcome level mean</i>	0.56	0.30	627	76
<i>Study level mean</i>	0.60	0.43	627	76
Share of outcomes with positive effect: programmes with proportion > 0.5				
<i>All programmes (n=29)</i>	0.59	0.28		
<i>Single study programmes (n=20)</i>	0.60	0.35		
<i>Programmes with ≥2 studies (n=9)</i>	0.55	0.11		
<i>Programmes with ≥3 studies (n=6)</i>	0.67	0.17		

3.3.3.2 Regression analysis

Our preferred model used outcome-level data, estimating a three-level mixed-effects logistic regression model using the grouped frequency specification (Table 4, Column 1). The model included random intercepts at both the study and programme levels to account for clustering at these levels.

Threshold-based payment was the only incentive design feature which showed a statistically significant association with positive outcomes at conventional levels ($p < 0.05$). The average marginal effect for threshold-based payment was 0.1921 (SE: 0.09568, $p = 0.045$), indicating a 19 percentage point higher probability of a positive outcome for programmes that include threshold-based payment compared to programmes without this design feature.

No other incentive design features showed statistically significant associations with positive outcomes. The coefficients on the design features bonus/reward payments, penalty/withholding and their interaction were all negative, but far from showing statistical significance at conventional levels. The coefficient on tournament/ranking and whether the specific outcome was incentivised both showed a positive association with programme effectiveness, but these associations were also not statistically significant.

Payment frequency categories (high, low, and medium frequency, relative to variable/unknown) showed no significant associations with outcomes, and neither did voluntary participation. The study design also showed no significant association with the probability of a positive outcome compared to the reference category, ITS.

Table 4 – Associations between incentive and study design and positive P4P effects

	Outcome-Level	Study-Level
Financial Incentive		
Bonus/reward payment	-0.1085 (0.1408)	-0.2254 (0.3419)
Penalty/withholding	-0.1251 (0.1333)	-0.1099 (0.3175)
Bonus/reward payment # penalty/withholding	-0.0473 (0.2362)	0.0382 (0.3412)
Tournament/ranking	0.0404 (0.1700)	-0.0865 (0.1628)
Threshold-based payment	0.1921 (0.0956)**	0.1756 (0.0978)*
Payment frequency (Reference: Variable/Unknown)		
High frequency	0.0439 (0.1241)	-0.0391 (0.1213)
Medium frequency	-0.0727 (0.1953)	-0.0250 (0.1926)
Low frequency	-0.0475 (0.1083)	-0.1357 (0.1182)
Participation (Reference: Mandatory)		
Voluntary	0.0062 (0.1044)	0.0520 (0.0972)
Outcome (Reference: Non-incentivised)		
Outcome incentivised	0.0817 (0.0537)	-0.0713 (0.1143)
Study design (Reference: Interrupted time series)		
Controlled before-after	-0.1028 (0.0779)	-0.0923 (0.0854)
Other	-0.0132 (0.1169)	0.0381 (0.1362)
Observations	627	76

Notes: Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Outcome-level analysis: marginal effects on probability of positive effect (percentage points).

Study-level: effects on proportion of positive outcomes per study.

Both models include random intercepts (study and programme for outcome-level; programme for study-level).

The study-level analysis examined the proportion of positive outcomes per study using a two-level mixed-effects linear model with studies nested within programmes (Table 4, Column 2). This analysis included 76 studies. The model specification was the same as the outcome-level analysis, but the variable 'outcome incentivised' was redefined as an indicator for whether any outcome in the study was incentivised.

Consistent with the outcome-level analysis, threshold-based payment showed a statistically significant association with the proportion of positive outcomes, also suggesting that threshold-based payments were associated with 18 percentage points higher chance of positive outcomes than studies without this design feature (SE: 0.0978 $p = 0.073$).

No other incentive or study design features showed statistically significant associations with the proportion of positive outcomes in the study-level analysis.

In our preferred outcome-level model, the programme-level variance was 0.191 (SE: 0.374), corresponding to an intraclass correlation (ICC) of 0.045 (SE: 0.085, 95% CI: 0.001 to 0.705), while the study-level variance was 0.812 (SE: 0.413), corresponding to an ICC of 0.234 (SE: 0.073, 95% CI: 0.121 to 0.403). These estimates indicate that approximately 23% of the variance in outcomes was attributable to differences between studies (conditional on programme), and approximately 4.5% to



differences between programmes. The small programme-level variance likely reflects that design features are constant within programmes and thus absorbed into the incentive variables, combined with the limited number of programme clusters ($n=29$). However, the three-level structure was retained for conceptual reasons, given the non-independence of studies evaluating the same programme.

3.3.3.3 Sensitivity analysis

We tested sensitivity to the definition of the dependent variable by re-estimating models using a more restrictive outcome: whether effects were both positive and statistically significant (Appendix 6.1 in [Supplement 1](#) to the Scientific Report). This affected the results of both the outcome- and study-level analyses, with threshold-based payment schemes no longer being statistically significantly associated with positive outcomes in either analysis. However, in the outcome-level analysis, incentivised outcomes were associated with a 10 percentage points (SE: 0.043, $p=0.016$) increase in the probability of showing positive and statistically significant effects. In the study-level analysis, no design or study feature was statistically significantly associated with this outcome.

In an alternative formulation, we restricted the outcome-level analysis to one main outcome per study only (Appendix 6.2 in [Supplement 1](#) to the Scientific Report). In this model, threshold-based payment showed a stronger association with positive outcomes, but slightly less statistically significant (33 percentage points (SE: 0.1788, $p=0.065$)). Other variables remained non-significant.

Our findings were affected by the level of payment frequency categorisation. We compared results using ungrouped frequency categories (seven categories: annually or longer, monthly, ongoing, quarterly, semi-annual, unknown, varying) versus grouped categories (four categories: high frequency, low frequency, medium frequency, variable/unknown) to assess whether the level of aggregation affected findings (Appendix 6.3 in [Supplement 1](#) to the Scientific Report). In this model, the marginal effect on threshold-based payment was not statistically significant at the 10% level (0.1405, SE: 0.0952, $p=0.140$), and no other variables became statistically significantly associated with a positive outcome of P4P.

We tested using the risk of bias classification as an alternative to study design, and found no statistically significant association between positive outcomes and risk of bias classification (Appendix 6.4 in [Supplement 1](#) to the Scientific Report). In this model, threshold-based payment was not statistically significant at a 10% level (0.1334, SE: 0.0865, $p=0.123$).

As a final robustness analysis, we compared results from our preferred 3-level outcome model to a two-level model (outcomes nested in studies (Appendix 6.5 in [Supplement 1](#) to the Scientific Report)). The marginal effects for threshold-based payment were similar in two-level and three-level models, confirming robustness to this modelling choice (coefficient 0.1766, SE=0.0853, $p=0.038$).

3.3.4 Intervention components, design features or contextual factors and P4P effectiveness (Programme theory)

COMMON MECHANISMS EXPLAINING P4P FINDINGS

Across the 90 studies, we coded 35 unique mechanisms pointed to by study authors to help explain their findings. The ten most common explanations of P4P findings overall are shown in Table 5 (all mechanisms by prevalence are outlined in Appendix 7 in [Supplement 1](#) to the Scientific Report). A third of studies notably pointed to other existing initiatives, such as pre-existing national safety improvement programmes, to explain their findings, across beneficial, detrimental, mixed and null findings. This highlights that P4P programmes are often part of wider quality improvement initiatives, with implications particularly for the interpretation of studies attempting to evaluate them without a control group. Similarly, many studies (19%) highlighted reputational or other non-financial pressures, which might explain part of P4P findings when implemented alongside, or indeed a lack of effectiveness findings when impacting the control group also. Complementary systems and policies,

such as technology adoption, were also commonly (17%) called attention to for explaining beneficial results.

Other common mechanisms outlined included incentives being considered too small (relative to implementation cost, 31%) and external constraints (27%) beyond the control of those implementing (e.g., outcomes depending on social determinants of health, patient actions, and services outside of the targeted sector). Institutional change (24%), such as encouraging shifts in organisational norms or care pathways, and organisational capacity to adapt and deliver change (18%), such as ability to invest in resources to achieve change, both featured commonly too.

Table 5 – Top ten mechanisms coded for explaining P4P programme results across studies

Rank	Mechanism	Number of studies	Share of studies
1	Existing initiative/secular trends	30	33%
2	Incentive too small	28	31%
3	External constraints	24	27%
4	Institutional change	22	24%
5	High baseline/ceiling effects	17	19%
6	Reputational/non-financial pressure	17	19%
7	Organisational capacity	16	18%
8	Potential gaming	16	18%
9	Complementary systems/policies	15	17%
10	Targeting evidence-based process	14	16%

MECHANISMS ASSOCIATED WITH BENEFICIAL P4P FINDINGS

Overall, 47 studies (52%) were coded as beneficial headline findings of their P4P evaluation (method of classification as described in section 2.7.3), seven (8%) were coded as mixed findings overall, while only six (7%) reported headline findings coded as detrimental. A third (n=30, 33%) of studies reported overall null effects. The full set of mechanism coding by outcome category is available in Appendix 8 in [Supplement 1](#) to the Scientific Report.

The mechanisms in Table 6 were found in more than 20% of beneficial studies but less than 10% of null/detrimental studies, generally suggesting they were seen as key enablers of P4P success. These included the incentives triggering overall institutional changes rather than one-off or surface-level responses; targeting highly evidence-based processes with incentives, where there was little disagreement that they were valuable targets, amenable to change, and with clear links to relevant outcomes; having complementary systems and policies (co-interventions) in place alongside, such as other policies drawing attention to patient safety, or databases and technology changes alongside implementation (e.g. coordination through electronic information exchange, registries, and patient engagement tools¹⁶¹); having organisational capacity and resources to shift practices in response to and aligned with incentives (e.g. some hospital wards having more flexibility to respond than others⁹⁷); and incentive targets coming from a low baseline, with clear room for improvement.



Table 6 – Mechanisms found more commonly in studies reporting overall beneficial outcomes for P4P

Mechanism	Beneficial	Null	Detrimental
Institutional change	45%	3%	0%
Targeting evidence-based process	30%	0%	0%
Complementary systems/policies	30%	0%	0%
Organisational capacity	30%	3%	0%
Low baseline (room for improvement)	23%	3%	0%

MECHANISMS ASSOCIATED WITH NULL OR DETRIMENTAL P4P FINDINGS

Conversely, the mechanisms in Table 7 were found in more than 20% of null/detrimental studies but less than 10% of beneficial studies, suggesting they might generally represent barriers to P4P effectiveness. These included the size of incentive being perceived as too small to trigger real change; a perception of external constraints, that achieving the targets was outside of the control of those being incentivised, or particularly for those hospitals facing cost or demographic pressures which effectively made the targets feel out of their control; financial incentives with a complex design were largely seen as ineffective, uninterpretable, and potentially preventing strategic planning and true institutional change to implement effective solutions; those targets with a high baseline were thought to be subject to ceiling effects for quality improvement; many studies also highlighted a disconnection between some process targets and outcomes, where even if the incentives worked as planned, the intended outcomes would not necessarily manifest; and potential gaming was highlighted, where the measure could be easily gamed (e.g. via changes in coding rather than real changes in practice), or on many occasions described as a result of alternative attempts to respond to incentives where there were felt to be too many external constraints or other factors preventing real positive changes (e.g. where seen as poor targets disconnected from outcomes).

Table 7 – Mechanisms found more commonly in studies reporting overall detrimental or null outcomes for P4P

Mechanism	Beneficial	Null	Detrimental
Incentive too small	4%	73%	67%
External constraints	9%	43%	83%
Complex design	0%	37%	0%
High baseline/ceiling	6%	37%	33%
Process-outcome disconnect	4%	23%	33%
Potential gaming	6%	27%	33%

Focusing only on the subset of studies where a threshold-based payment was recorded, these were slightly more likely to be recorded as an overall beneficial headline study finding. 39 studies were coded with a threshold-based payment, and 21 of these (54%) with an overall beneficial headline study finding (versus 52% of the total study sample). Focusing on the studies with a threshold-based payment as well as an overall beneficial headline finding (n=21, see Table 8), the top mechanisms looked very similar to those in Table 6 (slightly reordered).

Table 8 – Mechanisms found more commonly in studies coded as using threshold-based payments and where overall headline finding beneficial

Mechanism	Number of studies	Share of studies
Institutional change	9/21	43%
Complementary systems/policies	8/21	38%
Targeting evidence-based process	7/21	33%
Existing initiative/secular trends	6/21	29%
Low baseline (room for improvement)	6/21	29%
Organisational capacity	6/21	29%

Note: Table restricted to those mechanisms present in over a quarter of these studies for brevity

PROGRAMME THEORY DEVELOPMENT AND INTERPRETATION

The programme theory abstracted across studies is outlined in Figure 3, where the mechanisms described above and other CMO-related variables we extracted are grouped into common higher-level themes (no longer at the single study or programme level). These higher-level themes were then ordered according to the logical flow described within studies (e.g. those factors relating to how the incentive was initially designed, then the specifics of the design of the incentive itself, then the impacts that follow from these specific design choices). See [Supplement 2](#) to the Scientific Report for the study-by study breakdown of extracted mechanisms (above) together with the reasoning on overlap (row 94) with the Programme theory (below).

Both incentive design and baseline system conditions emerged as important contextual factors shaping P4P effectiveness. Inclusive stakeholder involvement and clinical engagement in programme development, combined with a solid evidence base, were frequently described as key to designing effective incentives. In contrast, top-down planning often resulted in targets perceived as disconnected from meaningful outcomes. Well-designed incentives were commonly characterised as being large enough to motivate behavioural change, simple and focused rather than diffuse, targeting well-evidenced and highly controllable (for the hospital alone) processes or outcomes, applied where there was a low current baseline with clear room for improvement, and framed as penalties rather than bonuses. Crucially, the literature drew attention to very few processes that met these criteria (an exception, for example, was stroke unit care as an ideal target with clear bundle of care incorporating adherence to multiple evidence-based elements with a large effect on outcomes⁸⁰). While a policymaker might identify an outcome they wanted to improve, rarely was there a clear evidence-base for a manipulable process on that precise pathway, particularly one which was not already being done (i.e. low current baseline).

Ability as well as willingness to change were identified as important mechanisms. Financial incentives primarily influenced willingness, operating through loss aversion and shifting organisational priorities. Ability to change, however, more often depended on existing system capacity, time, monetary and workforce as well as agency to implement alternative strategies. Incentives were frequently seen as insufficient in the absence of complementary systems and policies, with audit and feedback mechanisms, clinical networks, and public reporting commonly cited as essential accompaniments.

Even where programmes initially demonstrated beneficial effects, these could diminish over time through ceiling effects and diminishing returns as performance approached upper limits. Gaming emerged as a response in two distinct circumstances: as a potential opportunistic strategy where monitoring was weak (with perhaps then links to complementary policies in place), but also as a rational alternative where genuine improvement was not achievable due to system constraints or poor incentive design. Targeted incentives could also generate spillover effects in both directions, with some studies reporting positive spillovers through shared learning and broader care pathway improvements, while others documented negative spillovers through effort diversion away from non-



incentivised areas. Studies reporting mixed outcomes frequently emphasised this trade-off between targeted gains and unintended consequences elsewhere.

Penalties represented a double-edged sword. While likely to prompt additional organisational attention through loss aversion, penalty-based programmes risked compounding difficulties for organisations already under strain, potentially limiting their capacity to respond effectively. Finally, attribution presented a persistent challenge across the literature. Studies commonly pointed to secular trends and pre-existing quality improvement initiatives occurring alongside P4P implementation, making it extremely difficult to isolate the independent effect of the financial incentive.

Figure 3 – Programme theory relating P4P context-mechanisms-outcomes



Note: Grey indicates pathways that were generally outlined as more beneficial, yellow indicates those generally outlined as more detrimental.

PROGRAMME THEORY ALIGNMENT SCORES: CLASSIFYING THE DESCRIBED MECHANISMS

Calculating the programme-level alignment scores (see Table 9, study level scores available in [Supplement 2](#) to the Scientific Report), no programme showed a very high score, with the English Best Practice Tariffs (BPT) scoring the maximum observed programme-level score (5.7 out of a maximum of 12 points). Similarly, the minimum programme-level score observed was minus four (out of a minimum of -12).

Many programmes reflected their mixed evidence-base, for instance the US Hospital Readmissions Reduction Program (HRRP) across 22 studies showed a neutral net score of 0.27, an artefact of averaging across a highly heterogeneous set of evaluations. The beneficial-outcome studies for HRRP are numerous (14 of 22), but individually have fairly low alignment scores. Meanwhile, the smaller cluster of detrimental/mixed/null-outcome studies have quite concentrated detrimental pathway scores, which drags the programme-level mean net score down close to zero. The relatively high standard deviations across the programmes with more than one study further emphasise the heterogeneity.

It should be noted that the outcome classification (Column 4 in Table 9) is based on a brief summary of the overall study findings reported in the abstract (see section 3.2.7.3). In a small number of cases, this classification diverges from the scoring derived from the quantitative analysis (Table 3).

ASSOCIATION BETWEEN PATHWAY ALIGNMENT SCORES AND PROGRAMME OUTCOMES

Figure 4 displays the combined design pathway and outcome information in a matrix format, where bubble sizes show the volume of evidence. Columns are organised by the observed outcome classification across studies evaluating each programme: the left column contains programmes where all evaluated studies reported detrimental effects only, the middle column groups programmes with mixed outcomes, and the right column contains programmes where all evaluated studies reported beneficial effects only. Rows are organised by the design pathway classification and the mean net pathway score: from programmes classified as negative pathway only (negative or zero net score with no beneficial pathways recorded) at the top to programmes classified as beneficial pathway only (positive net score with no detrimental pathways recorded) at the bottom. As an example, the HRRP is situated in the mixed outcome, net positive mixed pathway field, reflecting the positive (albeit modest) net score arising from a mixed pathway combination. Although the beneficial pathway-beneficial outcome corner contains eight programmes, only three programmes (US Performance award, Unnamed Italian programme and English BPT) have net scores above 2, and only one programme (English BPT) is evaluated by more than one study.



Table 9 – Programme theory alignment score – Programme level

Programme				Beneficial Pathway Score			Detrimental Pathway Score			Net Score			
Programme Name	Country	N studies	Outcome breakdown	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	SD
Best Practice Tariff (BPT)	England	3	beneficial: 3	5.67	4	7	0	0	0	5.67	4	7	1.53
Performance Award Program	United States	1	beneficial: 1	5	5	5	0	0	0	5	5	5	—
Unnamed P4P Programme (Italy)	Italy	1	beneficial: 1	4	4	4	0	0	0	4	4	4	—
Advancing Quality (AQ)	England	5	beneficial: 4, mixed: 1	2.6	1	5	0.4	0	1	2.2	0	5	2.28
Unnamed P4P Programme (United States)	United States	1	beneficial: 1	2	2	2	0	0	0	2	2	2	—
BC-P4P (Breast Cancer Pay-for-Performance Program)	Taiwan	1	beneficial: 1	3	3	3	1	1	1	2	2	2	—
Emergency Department Quality Improvement Program (EDQIP)	Taiwan	1	beneficial: 1	2	2	2	0	0	0	2	2	2	—
Quality Improvement Payments (QIP)	Australia	2	beneficial: 1, null: 1	2.5	0	5	0.5	0	1	2	0	4	2.83
Radiology Report Signature Times incentive	United States	1	beneficial: 1	2	2	2	0	0	0	2	2	2	—
National HAC financial penalty policy	Australia	1	beneficial: 1	2	2	2	0	0	0	2	2	2	—
Housestaff Incentive Program	United States	1	beneficial: 1	1	1	1	0	0	0	1	1	1	—
Hospital Readmissions Reduction Program (HRRP)	United States	22	beneficial: 14, null: 4, mixed: 3, detrimental: 1	1.36	0	4	1.09	0	3	0.27	-3	4	2.25
Hospital-Acquired Condition Present on Admission (HAC-POA) Program	United States	12	beneficial: 8, null: 4	1.42	0	4	1.25	0	4	0.17	-4	3	2.25
Pay for Results Program	Canada	1	beneficial: 1	2	2	2	2	2	2	0	0	0	—
Quality Improvement Plan (OQIP)	Canada	1	beneficial: 1	1	1	1	1	1	1	0	0	0	—

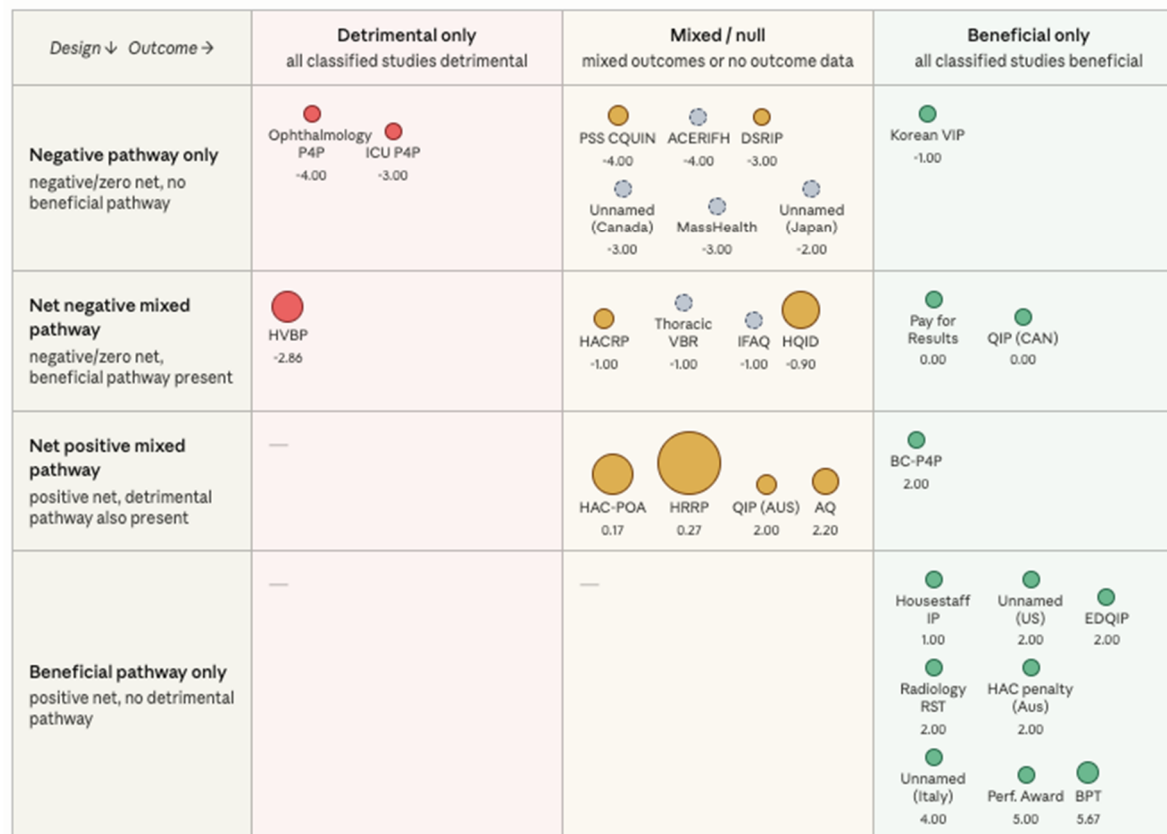
Programme				Beneficial Pathway Score			Detrimental Pathway Score			Net Score			
Programme Name	Country	N studies	Outcome breakdown	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	SD
Premier Hospital Quality Incentive Demonstration (HQID)	United States	10	beneficial: 3, null: 6, mixed: 1	1	0	4	1.9	0	4	-0.9	-4	3	2.51
Thoracic surgery value-based reimbursement initiative	United States	1	null: 1	1	1	1	2	2	2	-1	-1	-1	—
Korean Value Incentive Program	Korea	1	beneficial: 1	0	0	0	1	1	1	-1	-1	-1	—
Hospital-Acquired Condition Reduction Program (HACRP)	United States	2	beneficial: 1, null: 1	0.5	0	1	1.5	1	2	-1	-1	-1	0
IFAQ (<i>Incitation financière à l'amélioration de la qualité</i>) – Financial Incentive to Quality Improvement	France	1	null: 1	1	1	1	2	2	2	-1	-1	-1	—
Unnamed P4P Programme (Japan)	Japan	1	null: 1	0	0	0	2	2	2	-2	-2	-2	—
Hospital Value-Based Purchasing (HVBP)	United States	7	null: 5, detrimental: 2	0.14	0	1	3	2	4	-2.86	-3	-2	0.38
Medicaid Delivery System Reform Incentive Payment (DSRIP) program	United States	1	mixed: 1	0	0	0	3	3	3	-3	-3	-3	—
Unnamed P4P Programme (Canada)	Canada	1	null: 1	0	0	0	3	3	3	-3	-3	-3	—
MassHealth P4P program	United States	1	null: 1	0	0	0	3	3	3	-3	-3	-3	—
ICU P4P	United States	1	detrimental: 1	0	0	0	3	3	3	-3	-3	-3	—
Ophthalmology P4P	Germany	1	detrimental: 1	0	0	0	4	4	4	-4	-4	-4	—
Prescribed Specialised Services Commissioning for Quality and Innovation (PSS CQUIN)	England	2	null: 1, mixed: 1	0	0	0	4	1	7	-4	-7	-1	4.24
Additional Charge for Emergency Reduction and Internal Fixation or Hemiarthroplasty for Hip Fracture (ACERIFH)	Japan	1	null: 1	0	0	0	4	4	4	-4	-4	-4	—

Note: Programmes sorted by mean Net Score (descending). Multi-programme studies excluded (n=5 studies). Min/Max reflect range across papers evaluating the same programme. Net Score SD shown for programmes with >1 paper. Colour coding: green = beneficial pathway only (positive net score, no detrimental pathways recorded); amber = net positive mixed pathway (positive net score, detrimental pathways also present); orange = net negative mixed pathway (negative or zero net score, beneficial pathways present); red = negative pathway only (negative or zero net score, no beneficial pathways recorded).

Figure 4 – Cross tabulation of observed outcomes and design pathways by programme

Observed outcomes: ● Beneficial only ● Mixed ● Detrimental only ● Null

— bubble size = n studies | left to right = lowest to highest net score



3.3.5 Subgroup effects and equity

Nearly half (n=42, 46%) of the studies presented some clinical or equity-relevant subgroup effects. The majority of these studies examined clinical subgroups (e.g. by target condition, especially AMI, heart failure, pneumonia) rather than equity-relevant population characteristics (only n=15, 16% of total studies for equity-relevant characteristics).

Focusing on the PROGRESS-PLUS equity dimensions, three dimensions were studied: place of residence, race/ethnicity, and socioeconomic status. Findings were mixed across all three dimensions, as summarised in Table 10. The summary by individual study reporting on each of these dimensions can be found in Appendix 9 in [Supplement 1](#) to the Scientific Report.

Studies examining place of residence differentiated results by rural/urban, predominantly finding no differential effects. One study, evaluating the QIP programme in Australia, identified pro-equity effects for rural residents. The findings by race/ethnicity were all from programmes in the US and showed very mixed results, even within the same programme being evaluated. For instance, the HRRP programme was evaluated by two separate studies showing both a pro-equity effect (larger mortality reduction for Black patients with AMI compared to White patients, with no differential effects for heart failure or pneumonia) and an anti-equity effect (larger increases in 30-day mortality among Hispanics).

Table 10 – Summary of study effects examining equity-relevant subgroups

Dimension	Number of studies	Pro-equity	Anti-equity	No differential	Mixed
<i>Place of residence</i>	4	1	0	3	0
<i>Race/ethnicity</i>	6	2	2	2	0
<i>Socioeconomic status</i>	6	3	1	1	1
Total	15*	6	3	6	1

Note: * Gupta et al. (2018) examined both place of residence and race/ethnicity, so 15 unique studies across 16 dimension-study combinations.¹⁴⁷

Finally, the studies examining socioeconomic status leaned slightly towards pro-equity impacts, with potentially larger quality effects for safety net hospitals – hospitals that provide care to people regardless of ability to pay (these were a study evaluating the Premier HQID and two studies evaluating the HRRP in the US). Although, one study also identified larger quality increases for high-socioeconomic status, anti-equity, in an evaluation of the PSS CQUIN programme (though, they also found no differential impact for another scheme). Another study of the HRRP found mixed results, larger effects for safety-net hospitals for heart failure, but smaller effects for pneumonia. This suggests the relationship is not straightforward. This evidence on equity impacts by socioeconomic status is particularly mixed and requires careful interpretation when considering the financial considerations from the programme theory. While several studies found larger quality improvements in safety-net hospitals, this may reflect greater room for improvement from lower baselines rather than equitable programme design. These findings do not address concerns raised in programme theory about whether penalty-based incentives place disproportionate financial strain on already under-resourced hospitals, potentially threatening their long-term viability even when short-term quality gains are achieved.

Notably, other important equity dimensions emphasised by the PROGRESS-PLUS framework, such as occupation, gender/sex, religion, education, disability, sexual orientation, age were not considered in the current evidence-base.

3.3.6 Cost or cost-effectiveness considerations

Very few (n=6, 7%) of the studies reported any cost or cost-effectiveness considerations. These are summarised in Table 11. Findings were mixed overall, and none would be classified as a conventional (full) cost-effectiveness analysis. The Advancing Quality programme was evaluated in two studies and found to be highly cost-effective over 18-months (although, these studies shared the same authors and time period, so more strictly a single analysis updated for a report and then journal publication, and this analysis did not include a later study that questioned the long-term effect of AQ). The uncertainty analysis is also unclear and there are few details to fully judge the overall quality of the cost analysis. One study used a cost-consequence framework, an evaluation of the PSS CQUIN programme in England, and found an unfavourable cost-consequence balance over the first two years. Three further studies examined costs separately as outcomes, sometimes alongside other outcomes but not in parallel in a full cost-effectiveness or cost-consequence framework. An evaluation of the Quality Improvement Payments programme in Australia found no statistically significant difference in hospital admission costs, but the authors concluded informally that value improved alongside. An evaluation of the HAC-POA Program in the US identified an 8.1% (95% CI, -10.2% to -6.1%) reduction in hospital costs over eight years (measured for an inpatient stay with cost-to-charge ratios at the hospital level, not accounting for programme costs). Finally, an evaluation of the Premier HQID programme in the US identified no impact on Medicare costs (sum of each patient's Medicare hospital costs over a 60-day postadmission interval with costs attributed to the hospital in which the patient was initially admitted, not accounting for programme costs). Seven additional studies mentioned costs



much more informally, sometimes with back-of-the-envelope calculations (these are summarised in Appendix 10 in Supplement 1 to the Scientific Report).

Table 11 – Cost-effectiveness summary of findings

Study	Programme	Type	Finding
<i>Meacock et al. (2014)</i> ^{85*}	Advancing Quality (AQ)	Full CEA	Cost-effective: ~5,200 QALYs gained, health gain (£105m, based on the assumption that one QALY is worth £20 000; an estimated £4.4m based on bed-days saved) exceeded programme costs (£13m)
<i>Feng et al. (2024)</i> ⁹⁴	PSS CQUIN	Cost-consequence	Costs of the programme (£10.8m) exceeded efficiency savings (£1.6m, based on bed-days saved) for the trauma programme. No improvements detected for other programmes (so not analysed)
<i>Grimley et al. (2025)</i> ⁸¹	Quality Improvement Payments (QIP)	Cost analysis	Hospital admission costs after the implementation of the programme were not significantly different with the period before the implementation. Authors concluded improved value for expenditure
<i>Kim et al. (2021)</i> ¹⁰⁵	HAC-POA Program	Cost analysis	8.1% reduction in hospital costs, ~\$170m annual savings for hospitals (only including costs of the inpatient stay estimated from cost-to-charge ratios, not accounting for programme costs)
<i>Ryan (2009)</i> ¹²⁷	Premier HQID	Cost analysis	No impact on Medicare costs identified (sum of each patient's Medicare hospital costs over a 60-day post-admission interval with costs attributed to the hospital in which the patient was initially admitted, not accounting for programme costs)

Note: *Meacock et al., 2014 and McDonald et al., 2015 report findings from the same cost-effectiveness analysis.^{85, 89} We only report here findings from Meacock et al., 2014 as the journal peer-reviewed version.

3.4 DISCUSSION

SUMMARY OF FINDINGS

We conducted a systematic literature review of studies evaluating the impacts of hospital P4P programmes on quality dimensions. The identified evidence base was heavily skewed towards a small number of healthcare settings. Of the 90 studies included, 66 were conducted in the United States, with England contributing a further 10; the remaining 14 studies are distributed across eight countries, none contributing more than three studies each.

The overall picture is not suggestive of P4P being a consistent lever for quality improvement. Short-term positive effects are not consistently sustained in the long run, or when subject to methodologically more rigorous analysis, and where programmes have been deliberately designed based on theory or past experience, this does not appear to have resulted in consistently better designs or outcomes.

Our quantitative analysis found that the implementation of a threshold-based payment in P4P programmes was the only clear individual differentiator predicting likelihood of more positive outcomes when other factors were controlled for.

To examine in more depth whether and under what conditions P4P might be more or less effective, we supplemented these findings with a realist-inspired narrative synthesis analysing the relationships between context, mechanisms and outcomes across the same studies. The involvement of stakeholders in designing the incentives and having very clearly evidence-informed controllable

targets demonstrably linked to outcomes were emphasised as key context. Besides the incentives themselves, a system ready to act on them was also described as essential, as well as complementary systems and policies alongside (either together or already in place). In the right circumstances, this could trigger both a willingness and ability for positive institutional change. On the other hand, poorly designed incentives, and/or where external constraints threatened the ability of the institution to adapt to them, unintended gaming and selective effort responses were described. Even with initially successful incentives, ceilings were likely to be met unless incentives adapted, meaning even positive effects could be short-lived. Both positive and negative spillovers were also possible beyond the targeted outcomes and were sometimes simultaneously described (e.g. within versus outside the targeted specialty).

Linking the quantitative and narrative syntheses, the subset of studies coded as implementing a threshold-based payment did not differ substantially in the mechanisms outlined when compared to all studies reporting beneficial headline findings. Our analysis cannot conclude whether it is the use of thresholds per se that makes these schemes more likely to show an effect. It may be that threshold-based schemes are perceived as simpler/more clearly obtainable, and so more likely to trigger a willingness to change correlate, or that this design feature correlates with other design features supporting programme success in general.

Many of these threshold-based incentives with an overall beneficial headline study finding (as detailed in Table 8) appeared to target evidence-based processes with low baseline (room for improvement), all factors generally pointed to in studies finding more beneficial results. Interestingly, while penalties and loss aversion mechanisms were frequently highlighted as potential explanations within studies, our quantitative analysis did not find a clear impact of these incentive types over bonus or mixed payments. This again points to the multi-dimensional and complex nature of P4P incentives, that makes it difficult to isolate a clear silver bullet for effectiveness even where one might appear theoretically important.

Programme designs generally aligned poorly with the factors identified as beneficial in the narrative synthesis. On a scale with a theoretical maximum of 12, the highest observed design pathway net score was 5.7 for the English Best Practice Tariff (BPT) programme. Sixteen programmes displayed a negative net design score, 9 of which had no positive design elements.

Poor programme design was not consistently reflected in poor outcomes, nor were positive outcomes confined to programmes with better design scores. Several programmes with net negative design scores, for example, the US HQID and HACRP, reported mixed beneficial and null outcomes in the literature, suggesting that the design factors identified may not consistently capture the mechanisms driving outcomes, not least because the pathway scoring is derived from what authors chose to report rather than a systematic independent assessment of programme design.

The practical implications of the combined analysis are further constrained by the evidence base: all programmes in the positive design/beneficial outcome cell are represented by a single study with the exception of BPT, while the programmes with the largest evidence bases (HRRP, HAC-POA, and HQID) fall in the mixed (net positive mixed or net negative mixed) design categories. This pattern may reflect that, with more studies, there is greater opportunity for authors to flag both beneficial and detrimental design elements. Publication bias favouring studies that find an effect could also contribute to this impression. Taken together, while the framework supports hypothesis generation about beneficial and detrimental design features, the evidence base is insufficient to derive robust prescriptive lessons for future programme design.

Equity results were mixed, and limited to examination of place, race and socioeconomic status as stratifiers. Cost and cost-effectiveness evidence was particularly sparse and likewise mixed. Both dimensions merit further study, not least because rigorous policy assessment of P4P should ultimately address whether it represents value for money and whether it increases or reduces equity in health and access to high-quality care.



FINDINGS IN CONTEXT OF WIDER LITERATURE

Like recent systematic reviews including hospital P4P programmes, we report very mixed, inconclusive (low certainty even when positive results are found) or mostly modest findings.^{21, 26, 29, 31-33, 42, 170, 171} A notable exception might be the Best Practice Tariffs programme in the English NHS, whose effectiveness in improving the quality of hip fracture care was demonstrated in multiple studies, and where we observed the highest programme theory-alignment score. The majority of studies, though, showed mixed to mostly negative evidence when more intensively examined (such as the US studies of the Hospital Readmissions Reduction Program (HRRP), the Hospital-Acquired Condition Present on Admission (HAC-POA) Program, and the Premier Hospital Quality Incentive Demonstration (HQID), where evidence was predominantly negative)(Table 3).

Previous reviews also similarly outlined the importance of the specific P4P design in potentially altering effectiveness, such as absolute versus relative targets (with the caveat of ceiling effects);^{21, 32, 42, 170} penalties as potentially more powerful than bonuses;^{26, 31} that they must be of substantial enough value to matter;^{33, 42} that there must be room for improvement in targets;^{21, 32, 42, 170} that process measures might be more effective than outcomes;^{21, 31-33, 42, 170} simple design is preferable;³³ and best combined with other interventions such as public reporting and feedback.^{21, 32, 42, 170} Similarly, a bottom-up approach may achieve better buy-in and results,^{32, 33, 170} as well as those targets viewed as clinically meaningful.¹⁷⁰ We build on this by outlining how these aspects are seen to fit together and trigger specific mechanisms on the way to better outcomes, but also the feedback loops that might explain this particularly short-termism of any beneficial effect, while not ruling out other spillover effects in combination.

Specifically comparing to the earlier systematic review by Van Herck et al. (2010), which covered both primary care and hospital settings, our results are broadly consistent.⁴² Both reviews find that P4P is associated with a wide spectrum of effects depending on programme design and context, that baseline room for improvement is a key determinant of impact, that process indicators (we specify, clearly linked to outcomes) tend to yield potentially larger improvements, that stakeholder involvement supports effectiveness, and that cost-effectiveness evidence remains scarce. However, our hospital-focused review arrives at a more cautious overall assessment. Where Van Herck et al. found largely positive equity effects driven by the UK primary care Quality and Outcomes Framework, our evidence on equity in hospitals is mixed, with examples of both pro- and anti-equity effects within the same programmes.⁴² Their finding that purely positive (bonus) incentives outperformed competitive approaches is not supported by our quantitative analysis, in which only threshold-based payment emerged as a significant predictor of positive outcomes. These differences likely reflect both the more disappointing hospital-specific evidence base relative to the primary care programmes that dominated Van Herck et al.'s sample, and advances in analytical methods, including our own use of multilevel regression and realist-informed programme theory, that make the limits of attribution and the complexity of causal pathways more transparent.

Our results echo those of other meta-analyses using a vote-counting approach, which, though broader in the types of incentive schemes considered, found that financial incentives implemented in the hospital setting were less likely to be effective than schemes implemented in primary care.^{74, 75} In comparison to these studies, in our sample, we could not replicate the finding that study design was statistically significantly associated with the probability of finding positive effects of P4P, but this could be because, contrary to Scott et al., our search strategy excluded studies without a control group or temporal trend analysis.⁷⁴

Our findings broadly fit with prominent psychological theories of human motivation, such as Self-Determination Theory (SDT), which suggests that intrinsic motivation relies on the satisfaction of basic psychological needs, including autonomy, competence, and relatedness.¹⁷² For instance, foundations of more effective design principles such as clinical involvement and engagement speak to the self-direction (autonomy) of shaping the incentives, the clear evidence-base to the feeling of effectiveness (competence) and mutual professional respect (relatedness). This motivation would be consistent with the 'willingness to change' aspect we outlined as the main potential influence for successful P4P

incentives. Similarly, the complementary Expectancy-Value Theory, which emphasises motivation arising from interaction between a person's expectation of success and the subjective value they place on the outcome, might help bring in the 'ability to change' aspect we highlighted as a key supplementary mechanism.¹⁷³ Similarly, Prospect Theory from behavioural economics would help explain the potential power of penalties, through loss aversion (additional perceived value of avoiding a loss more than acquiring a gain), which could suggest that penalty-based incentive schemes carry a higher risk of inducing unintended effects, such as reclassification of cases due to the perceived strength of such incentives.¹⁷⁴ While this was suggested by many studies within our mechanisms analysis, our quantitative analysis did not find any significant effect of penalties above bonuses, however. Multitask Principal-Agent Theory helps explain the rational reallocation of effort toward measured at the expense of unmeasured activities, the mechanism through which gaming and selective effort might emerge, and both potential positive and negative spillovers.¹⁷⁵

STRENGTHS AND LIMITATIONS

A key strength of this review is the combination of quantitative synthesis with a realist-inspired narrative analysis, allowing us to move beyond whether hospital P4P programmes work on average to examine how, why, and under what conditions they might be more or less effective. This mixed approach better reflects the complexity of P4P as an intervention embedded in diverse institutional and incentive design contexts, and provides more actionable insight for policymakers than either approach alone. We also employed a hybrid extraction approach combining machine with human verification, which enabled efficient processing of a large evidence base while maintaining the accuracy needed for both quantitative and qualitative synthesis.

Several limitations should be noted, however. Our search was limited to exclusively quantitative evaluations, in line with our primary research aim of assessing effectiveness. This means we may have missed qualitative studies offering richer insight into the mechanisms and contextual factors we explored in our narrative synthesis, though the mechanism data we did extract from within quantitative studies proved sufficient to develop coherent CMO configurations. We only included studies published in English, which may have led to the omission of some country-specific evidence, however, given that English is the dominant publication language in this field, the likely impact on conclusions is modest. Our search was also limited by date (from 2008), though this was a deliberate decision to build upon and extend the findings of a previous systematic review covering earlier periods.⁴² The geographic concentration of the evidence base also constitutes an important limitation with implications for the transferability of findings. Finally, the heterogeneity of P4P programmes in terms of design, setting, payment structure, and target conditions limited the extent to which quantitative findings could be pooled, though this heterogeneity is itself informative and was a motivation for the realist-inspired analysis.



4 LESSONS LEARNED FROM THE DESIGN AND IMPLEMENTATION OF FIVE P4P PROGRAMMES IN ACUTE HOSPITALS

KEY POINTS

- **A scoping review and expert consultation** identified approximately 40 OECD hospital P4P programmes (active or ended), from which five were selected for detailed comparative analysis:
 - o Hospital Value-Based Purchasing (HVBP) programme in the US Medicare
 - o Commissioning for Quality and Innovation (CQUIN) scheme in England
 - o Best Practice Tariffs (BPTs) in England
 - o Clinical Practice Improvement Payment (CPIP) and Quality Improvement Payments (QIPs) in Queensland (Australia)
 - o *Incitation financière à l'amélioration de la qualité* (IFAQ) in France
- Although country-specific drivers differed, all **P4P programmes were introduced to address limitations of activity-based and volume-driven hospital financing systems** by explicitly linking part of hospital funding to quality improvement, patient outcomes, and adherence to evidence-based care.
- Across countries, **P4P financing** is not a standalone funding stream but is **embedded within existing hospital payment systems** (e.g. DRGs or blended payments), operating through mechanisms such as budget-neutral redistribution, tariff adjustments, or withheld contract amounts to align quality incentives with overall healthcare financing frameworks.
- P4P schemes use **different incentive logics**, ranging from **pure bonuses** (e.g. QIPs and BPTs) rewarding achievement of defined standards, to **pure or implicit penalties** (e.g. CQUIN through withheld payments), and **hybrid or budget-neutral systems** (e.g. HVBP and IFAQ) that combine rewards and penalties through redistribution of a fixed funding envelope based on performance.
- P4P programmes often combine **achievement-based incentives** (meeting predefined standards or thresholds) and **improvement-based incentives** (progress relative to a hospital's own baseline), although the balance varies: systems like BPTs and QIPs are mainly achievement-focused, while HVBP and (pre-2026) IFAQ explicitly integrate both, and CQUIN has shifted over time from locally tailored improvement towards more standardised achievement-based performance measurement.
- P4P designs mainly fall into **threshold-based systems** (payment for meeting absolute or improvement thresholds) or **tournament-based systems** (relative ranking with a fixed budget redistributed between providers), with most programmes (e.g. CQUIN, BPTs, QIPs) being threshold-based, while others like HVBP and pre-2026 IFAQ include stronger tournament or hybrid elements through budget-neutral redistribution and comparative performance scoring.
- P4P incentives are usually **paid to hospitals as a whole** and absorbed into organisational budgets (e.g. HVBP, CQUIN, BPTs, IFAQ), while only some systems like the CPIP pilot programme in Queensland encourage partial redistribution to clinical units.
- The **selection of quality domains and indicators** is primarily **policy-driven but operationalised through structured clinical and expert processes**, combining national priorities (e.g. safety, effectiveness, efficiency, patient experience) with input from clinical

networks, expert bodies, and stakeholders, and increasingly anchored in evidence-based guidelines and national standard-setting institutions to ensure indicators are clinically relevant, measurable, and aligned with system-level objectives.

- Quality indicators are generally grouped into structure, process, outcome, and patient experience measures, but the balance differs: **process indicators dominate most European schemes, outcome indicators are most developed in the US Medicare HVBP programme** (mortality, complications), and **structure indicators are comparatively limited and mainly used to ensure organisational capacity** (e.g. some English BPTs) **or accreditation requirements** (notably in France). **Patient experience** measures are frequently included as a distinct, survey-based category, but generally receive **less emphasis and financial weight** than process indicators.
- In the programmes studied, **P4P is embedded within wider hospital quality governance systems rather than operating as a standalone mechanism**, and its effects depend on how it interacts with regulatory oversight, performance monitoring, and contractual accountability frameworks. Alongside these governance structures, **P4P coexists with multiple complementary financial and non-financial quality tools**, such as penalties for hospital-acquired conditions and readmissions, bundled payments, mandatory reporting requirements, inspections, benchmarking, and public reporting, meaning that hospital behaviour is shaped by a layered system of quality improvement strategies in which P4P is only one component.
- The **number and composition of P4P indicators** vary widely across programmes and evolve over time in response to changing policy priorities, programme maturity, and integration into broader payment and quality frameworks. While some programmes expanded and became more sophisticated (e.g. English BPTs), others moved towards simplification and standardisation (e.g. CQUIN, QIP), whereas some maintained relatively stable but periodically revised indicator sets (e.g. HVBP).
- P4P participation ranges **from voluntary to mandatory**, but most systems have moved towards more standardised, compulsory integration into national payment systems. All programmes that include penalties are mandatory.
- **Transparency and accountability** are ensured through **combinations of public reporting, internal feedback, and national data systems**, but the level of integration varies.
- Across countries, formal evaluations (government reports and scientific literature) generally find limited effects of P4P on quality of care. **More complex, multi-indicator schemes** (e.g. HVBP, CQUIN, IFAQ) often **show weak or inconsistent improvements** due to small financial incentives, high administrative burden, and limited clinician engagement, while simpler, pathway-based schemes like BPTs and QIPs tend to show more consistent improvements in specific clinical processes (e.g. hip fracture, stroke).
- These evaluation findings have directly shaped policy reforms, with **programmes increasingly moving towards simplification and stronger integration into national payment systems**. Examples include centralising and reducing indicators in England's CQUIN, incremental adjustments rather than major redesign in HVBP, and a major structural reform of France's IFAQ towards more transparent, absolute, per-indicator threshold-based payments and clearer links between performance and funding.



4.1 Introduction

This chapter compares hospital P4P programmes to examine how financial incentives are designed to improve quality and healthcare outcomes. Although P4P programmes pursue similar objectives, they differ in their performance indicators, incentive structures, implementation, and target populations.

A comparative analysis is valuable because programme design influences both effectiveness and unintended consequences. By identifying similarities and differences across programmes, this chapter aims to draw lessons on key design features, highlight policy trade-offs, and identify elements that may inform future P4P development and implementation.

4.2 Method

4.2.1 Selection of P4P programmes

LONG LIST OF PROGRAMMES

A scoping review was carried out to identify a long list of countries with a P4P programme targeting acute hospitals. The scoping review combines a structured web-based search with insights gathered from international contacts. The web search covered both academic databases and general search engines. Particular attention was given to grey literature, including government reports, policy briefs, and technical documents produced by international organisations such as the WHO, the OECD, and the European Observatory on Health Systems and Policies, as well as publications from national health authorities. In addition, international experts were consulted to provide further information and to help identify less well-known programmes. Priority was given to sources that provide sufficient detail on the programme's design, implementation, and outcomes.

The following criteria were used to determine eligibility for inclusion in the long list of P4P programmes:

- Programmes must target acute hospitals or hospital departments, including services related to inpatient care, day care, and emergency care;
- A financial incentive (e.g. bonus, penalty, or adjusted tariff) must be explicitly linked to quality;
- Programmes must rely on measurable indicators of structure, process, or outcomes related to quality (effectiveness, safety, appropriateness of care, continuity of care, and people-centredness);
- Adequate publicly available documentation must be accessible in English, Dutch, or French (or be translatable);
- Programmes may operate at either the national or regional level and must originate from an OECD country;
- Programmes must be either currently active or implemented within the past 20 years to ensure relevance to current policy and practice.

For each programme included in the long list, a set of core characteristics was systematically documented. These include the country of origin, year of launch, current status (active or ended), structure of financial incentives, types of indicators used, targeted quality dimensions, participating providers, governance arrangements, and the status of any evaluation efforts. The selection criteria resulted in the inclusion of approximately 40 programmes.

SHORT LIST OF PROGRAMMES

A predefined set of criteria was applied to reduce the long list to a maximum of five programmes. This limitation was necessary to ensure feasibility within the available time.

The transition from the long list to the short list aimed to include a balanced mix of both successful and less successful initiatives, allowing lessons to be drawn from both effective and less effective approaches. The selection was guided by the following criteria.

Two criteria were mandatory for inclusion:

- Availability of evaluation data (peer-reviewed or grey literature);
- Sufficient duration and maturity of the programme, ensuring it has been in place long enough for meaningful results to have emerged.

Additional selection criteria include:

- Different design types, such as reward- versus penalty-based schemes;
- Variability in scope, ranging from single-condition focus to system-wide quality metrics;
- Diversity of quality domains targeted, such as patient safety, people-centredness, or clinical effectiveness;
- Indicator type diversity, ensuring a balanced mix of process, structure, and outcome indicators to support a comprehensive assessment of healthcare quality.

The following exclusion criteria were applied:

- Financial incentives not directed at hospitals (but regions) or lacking a direct link to quality metrics;
- Too narrow in scope (pathology, number of hospitals) or only aimed at enhancing efficiency;
- Similar types of programmes;
- Programmes that were introduced too recently to allow for a sufficiently meaningful evaluation.

Taking into account the above criteria, the following P4P programmes were included:

- Hospital Value-Based Purchasing (HVBP) programme in the US Medicare (since 2012)
- Commissioning for Quality and Innovation (CQUIN) scheme in England (2009-2024)
- Best Practice Tariffs (BPTs) in England (since 2010)
- Clinical Practice Improvement Payment (CPIP) and Quality Improvement Payments (QIPs) in Queensland (Australia) (2007-2013 and 2012-2022)
- *Incitation financière à l'amélioration de la qualité* (IFAQ) in France (since 2012).

All chapters were reviewed by one or more programme experts.

4.2.2 Information collection process

To increase coherence and comparability between programmes, a detailed template with an elaborate table of contents and a short description of each section for the programme reports was developed (see Table 12).

The structure of the programme reports aligns with the topics outlined in Table 12. Based on the programme reports, a transversal analysis was conducted (sections 4.3 to 4.10). We refer to [Supplement 1](#) to the Scientific Report for the programme reports and for references relating to the programmes in the following sections.



Table 12 – Template for programme assessment

Topic	Short description
Brief overview of the financing methods for hospitals	• Overview of the mix of hospital payment methods, within which the P4P system represents one component • For the P4P system, at least the following topics are covered: - o Start date or duration - o Current status (discontinued, scaled up following a pilot study, etc.) - o Implementing authority (ministry of health, insurance fund, hospital association, etc.)
Objectives and rationale of the programme	• Primary goals (e.g. improving patient outcomes, minimise unwarranted variations in clinical practice and performance across hospitals or regions) • Motivation for implementation (e.g. past quality issues, economic drivers, shift focus from volume-based to value-based healthcare, make quality metrics and performance data more visible to stakeholders)
Design	• Indicator design - o Indicator types: structure, process, outcome - o Scope: hospital-wide (e.g. patient satisfaction, safety culture), specialty-specific (e.g. cardiology), episode-specific (e.g. hip replacements, sepsis) - o Indicator selection approach: basis for selecting indicators (e.g. clinical guidelines, national priorities, expert consensus etc.), and whether indicators are evidence-based or have undergone validation - o Weighting system of indicators: equal across indicators or tailored by importance - o Total number of indicators - o Quality dimensions covered - o Data sources: administrative data, clinical registries, surveys, etc. - o Case-mix adjustment • Target definition and incentive structure - o Target definition approach: basis for setting targets (e.g. scientific literature, national priorities, baseline data analysis, expert consensus, etc.) - o Type of financial incentive: bonus payments for achieving quality benchmarks, penalty deductions for poor performance or underachievement, blended models (e.g. base payment plus quality-linked add-ons), best practice tariff - o Basis of target and payment: absolute thresholds, relative performance, improvement from baseline - o Size and timing of payments: time lag between performance measurement to payment - o Maximum bonus or penalty per provider - o Hospital participation: voluntary or mandatory (by regulation or incentivised through benefits). Are there any exclusion criteria, such as the non-participation of hospitals in remote areas? • Evaluation strategy

- o Adjustment of indicators, weights or targets
- o Reasons for changes to the set of indicators over time
- o Management of unintended effects
- o How are issues like gaming, risk selection, or neglect of non-incentivised areas monitored and addressed
- o Auditing, data verification processes, data quality, etc.
- Integration with hospital strategy: alignment with hospital accreditation or other quality improvement initiatives
- Stakeholder involvement: which stakeholders were involved in the design and/or other stages?
- Methods used for transparency and accountability: are performance results published and accessible?

4.3 Overview of the healthcare system and hospital funding mechanisms

P4P programmes are embedded within broader hospital financing systems and cannot be analysed in isolation. Their design, implementation, and impact are closely shaped by the underlying payment structures, governance arrangements, and policy objectives of each health system. Understanding these wider financing frameworks is therefore essential to contextualise how P4P mechanisms operate across countries.

GOVERNANCE AND ORGANISATION OF HOSPITAL SYSTEMS

Across countries, hospital financing is embedded within broader health system governance structures that shape how services are planned, funded, and delivered. In **England**, the National Health Service (NHS) is centrally funded through general taxation, with resources allocated to regional bodies based on population needs. These bodies, currently called Integrated Care Systems which are led by Integrated Care Boards, are responsible for commissioning most health services for their local populations and for coordinating care. A range of specialised services (tertiary care) are directly commissioned at the national level and managed by NHS England through its regional teams. Since 2024, some specialised services are now commissioned by Integrated Care Boards. Secondary and tertiary care providers in the NHS are organised into trusts. Each trust may manage one or several hospital sites. These hospitals are publicly owned and operated.

Similarly, in **Queensland** (Australia), healthcare governance is decentralised, with state authorities playing a central role. Public hospitals are managed by regional Hospital and Health Services (HHS), which act as providers and budget holders, while the state health department (Queensland Health) functions as purchaser and system manager, including the performance of HHS.

In contrast, the **US Medicare** system operates as a federal social health insurance programme, administered centrally but covering a defined population group (primarily older adults). **France** combines national regulation with regional implementation, where central authorities define payment rules and regional agencies oversee allocation and service delivery.

SHIFT TOWARDS ACTIVITY-BASED PAYMENT AND DRG-BASED CLASSIFICATION SYSTEMS

Across countries, a common reform trend has been the gradual shift from historically based hospital budgets or cost reimbursement towards activity-based funding models. Under earlier arrangements, such as those used in England and Queensland, hospitals were financed through global or block budgets largely determined by historical allocations. Over time, these approaches were either replaced or supplemented by funding mechanisms that more closely reflect actual service delivery.

In **England**, this shift was formalised in 2003 with the introduction of Payment by Results (PbR), which linked hospital reimbursement to activity using nationally determined tariffs. **Queensland** followed a comparable trajectory, introducing case-mix funding in 2007 and progressively aligning with activity-based funding principles as part of broader national reforms from 2011 onwards. **France**



adopted activity-based payment for acute hospitals in the mid-2000s, while the **US Medicare** programme introduced an inpatient prospective payment system in 1983 with the aim of containing hospital costs and improving system efficiency. Although differing in design and implementation, these reforms share the common objective of improving efficiency, transparency, and responsiveness by tying funding more directly to the volume and type of care provided.

A central feature of these activity-based systems is the use of Diagnosis Related Groups or DRG-type classification frameworks to define hospital activity and determine reimbursement. While the underlying principle is consistent across countries, each system applies its own national adaptation. In England, Health Resource Groups (HRGs) are used as the basis for payment, whereas the United States relies on Medicare Severity DRGs (MS-DRGs). Queensland uses the Australian Refined DRG (AR-DRG) system, and France applies *Groupes Homogènes de Malades* (GHM), with corresponding tariffs known as *Groupes Homogènes de Séjours* (GHS).

INCREASING USE OF BLENDED AND FLEXIBLE PAYMENT MODELS

Despite the widespread adoption of activity-based funding, all countries have progressively moved towards more blended payment models. These combine different funding mechanisms to address limitations of pure activity-based systems, such as incentives for volume over quality or insufficient support for integrated care.⁴⁸

In **England**, the Aligned Payment and Incentive (API) model has largely replaced PbR, combining fixed payments with variable components linked to activity and quality. The fixed payment is based on expected activity levels, the variable component is linked to actual service provision, particularly for elective and outpatient care. The fixed element provides financial stability, while the variable part maintains incentives related to activity levels. API has become the main mechanism for funding hospital services.

Queensland also uses a mix of activity-based funding and block payments, particularly for services not well suited to case-mix approaches, such as small hospitals or specialised services.

France has introduced additional funding streams alongside activity-based funding, including bundled payments, global budgets, and targeted funding for specific services. Similarly, in the **US Medicare**, payments include add-ons for new medical services or technologies.

This shift reflects a broader policy trend towards balancing efficiency incentives with financial stability and encouraging coordination across providers.

4.4 Rationale and policy context for the introduction of P4P programmes

P4P programmes in hospital care have generally emerged as policy responses to concerns that traditional funding systems, particularly activity-based or volume-driven payment models, do not sufficiently incentivise quality of care. Across countries, these programmes were introduced to rebalance financial incentives by explicitly linking a portion of hospital revenue to quality improvement, patient outcomes, and adherence to best clinical practices. Beyond their direct financial effects, P4P schemes were also intended to signal the importance of quality within healthcare systems, thereby supporting broader cultural and organisational change.

In **Queensland**, the introduction of P4P was closely linked to a major quality scandal in 2005, which exposed deficiencies in clinical governance and oversight in public hospitals. In response, policymakers sought not only to strengthen monitoring and accountability but also to actively promote high-quality care through positive incentives. The Clinical Practice Improvement Payment (CPIP) programme, implemented as a pilot in 2007-2008, formed part of this strategy. It was introduced alongside a new case-mix funding model, given concerns that activity-based payments could unintentionally reward poor quality by compensating hospitals for treating preventable adverse events. The CPIP programme aimed to improve performance in targeted clinical areas while also fostering a broader cultural shift by recognising and rewarding quality. Although the programme was discontinued

after its pilot and post-pilot phases, it laid the foundation for subsequent Quality Improvement Payments (QIPs), which continued to incentivise best practices within the activity-based funding framework through collaboration with clinical networks.

In **France**, the introduction of the IFAQ programme in 2012 was similarly motivated by perceived limitations of activity-based payment systems. While DRG-based funding improved transparency and efficiency, it was not designed to promote quality and could, in some cases, create incentives to reduce costs at the expense of care standards. At the same time, the development and validation of national quality indicators made it possible to measure and compare hospital performance more systematically. IFAQ was therefore introduced to explicitly integrate quality into hospital financing, with the objective of recognising and rewarding facilities that invest in quality improvement initiatives and perform well on validated indicators.

The Hospital Value-Based Purchasing (HVBP) programme was introduced in the **US Medicare** system as part of broader reforms under the Affordable Care Act (2010), which sought to address longstanding concerns about high healthcare spending and relatively poor performance on key health outcomes. Despite earlier efforts to improve efficiency, such as the introduction of DRG-based payments, hospital costs continued to rise, and quality improvements remained uneven. The HVBP programme reflects a strategic shift from volume-based to value-based purchasing, transforming Medicare into a more active payer that rewards hospitals for delivering better outcomes and patient experiences. It builds on earlier demonstration projects, notably the Premier Hospital Quality Incentive Demonstration (2003-2009), which tested the feasibility of linking financial incentives to quality performance.

In **England**, P4P mechanisms were introduced within a broader policy shift towards a quality-focused National Health Service. Concerns that the Payment by Results system might incentivise activity and cost control at the expense of quality contributed to the development of complementary incentive schemes. The 2008 Darzi Review ('High Quality Care for All') provided the strategic foundation for this shift, emphasising the importance of measuring, reporting, and rewarding quality across the NHS. This led to the introduction of several P4P initiatives, including the Commissioning for Quality and Innovation (CQUIN) framework and Best Practice Tariffs (BPTs), alongside non-payment policies for serious adverse events.

The **CQUIN** framework, launched in 2009, was initially designed to support locally defined priorities and to incentivise improvements in patient outcomes through better safety, clinical effectiveness, patient experience, and more innovative service delivery. Innovation was understood primarily as the adoption and diffusion of existing best practices rather than the development of new interventions.

CQUIN was intended to embed quality improvement into routine commissioner-provider negotiations by linking a small but visible share of provider income to the achievement of agreed goals, thereby reinforcing the importance of quality in organisational decision-making. The scheme was designed to be locally adaptable, allowing commissioners and providers to tailor indicators to local needs, while setting targets that most providers could realistically meet.

Since its introduction, the framework has gradually evolved towards greater standardisation, with increasing use of national indicators to support system-wide priorities and benchmarking. Following the Health and Social Care Act 2012, oversight was transferred to NHS England in 2013, and the scheme was split into two strands: the Clinical Commissioning Group (CCG CQUIN) for locally commissioned services (now within Integrated Care Boards) and the Prescribed Specialised Services (PSS CQUIN) for nationally commissioned specialised care, which were gradually reintegrated into one programme again from 2019 onwards.

In parallel, **Best Practice Tariffs** represent a more direct integration of quality incentives into the core payment system. Unlike CQUIN, which operates at the level of overall contract performance, BPTs adjust the price paid for specific clinical activities within the activity-based payment system. Higher tariffs are paid when care is delivered in accordance with evidence-based best practices, while lower tariffs apply when these standards are not met. Service areas selected for BPTs were typically those



with high clinical or system impact, such as high patient volumes, substantial variation in practice, or important effects on outcomes, combined with a strong evidence base and broad clinical consensus regarding what constitutes best practice care.

4.5 Financing of P4P programmes: macro-level framework and operational rules

This section covers two key aspects of P4P schemes. First, it situates P4P within broader healthcare payment systems, highlighting how it is embedded in existing reimbursement models rather than operating as a standalone financing mechanism (section 4.5.1). Second, it examines how P4P is financed, focusing on the ways incentive payments are generated, allocated, and adjusted within overall healthcare budgets (sections 4.5.2 to 4.5.6). Table 13 summarises the most important characteristics discussed in section 4.5.

4.5.1 Integration with broader hospital payment systems

Across countries, P4P mechanisms are increasingly integrated into broader hospital purchasing arrangements rather than operating as standalone programmes. These incentives are typically embedded within annual contracting cycles or payment systems and are used to steer provider behaviour towards quality improvement and alignment with policy priorities.

In **Queensland**, P4P mechanisms were layered on top of an activity-based funding system, meaning that quality incentive schemes do not replace existing reimbursement structures but instead function as additional, targeted payments within the broader purchasing framework. The payment was restricted to non-recurrent spending on quality improvement and professional development (e.g. data base upgrades, team training). In 2012, the QIPs were incorporated into the hospital purchasing cycle.

In the **Medicare** system, P4P is fully integrated into the existing prospective payment system based on MS-DRGs. The HVBP programme does not operate as an add-on payment but modifies base DRG-payments through upward or downward adjustments. In this sense, it represents a shift within the DRG-based payment structure rather than an additional layer of funding on top of it.

The HVBP programme is explicitly designed to be budget-neutral: penalties imposed on some hospitals finance rewards for others, with the total incentive pool created by withholding a fixed share of Medicare operating DRG-payments (rising gradually to 2%). Each hospital's total performance score is translated into a payment adjustment via a linear exchange function, ensuring that aggregate payouts equal the withheld amount at the national level. Note that the maximum penalty is capped (at 2%, corresponding to the amount withheld) while the maximum bonus is not, allowing high-performing hospitals to receive net gains.

The **CQUIN** programme in **England** links a proportion of provider income (0.5% rising to a maximum of 2.5% between 2012 and 2018; between 2019 and 2024 equal to 1.25%) to the achievement of quality improvement and innovation goals. Rather than providing additional funding, it operated by withholding part of the contract value and making its release conditional on performance, thereby embedding quality incentives within existing payment structures. In 2023, CQUIN payments were calculated as a percentage of the provider's total fixed API payment and the expected annual contract value for elective activity. Payments were issued in full, in advance, through monthly instalments, but any underperformance against nationally defined indicators resulted in end-of-year deductions applied to the variable component of API payments. This approach differed from earlier years, when commissioners had greater discretion over payment timing, including advance payments with retrospective corrections, payments conditional on target achievement, or mixed arrangements combining both approaches.

Best Practice Tariffs were introduced as part of the baseline payment system itself: these adjust the standard HRG tariff level upward or downward depending on whether providers deliver care in line with evidence-based best practice, thereby embedding quality incentives directly into core reimbursement rather than adding separate bonuses.

Over time, these mechanisms have become increasingly integrated within the main payment architecture. With the introduction of the API model and the NHS Payment Scheme, quality incentives are now directly linked to blended payments combining fixed and variable elements, rather than existing as fully separate schemes.

P4P in **France** (IFAQ) is closely integrated with the national activity-based hospital financing system. Initially, quality payments operated as separate bonuses financed through a dedicated envelope alongside DRG-based reimbursement. Over time, however, quality incentives became increasingly embedded within the broader hospital financing framework. Since the 2019 reform, part of the hospital financing envelope has effectively been redistributed according to quality performance, meaning that quality no longer functions solely as an add-on incentive but directly affects the allocation of core hospital resources. Part of the budget was allocated according to both hospital activity and a quality score reflecting comparative performance and improvement. As a result, hospitals with lower scores may receive less funding than under a non-IFAQ system, meaning the mechanism can effectively penalise underperforming providers.

4.5.2 *Type of financial incentive*

Across countries, P4P schemes differ in whether they operate through pure bonuses, pure penalties, or a combination of both. In some settings, such as **Queensland**, incentives have historically taken the form of capped bonus payments on top of existing hospital funding. These payments are awarded when predefined targets are achieved and are explicitly earmarked for quality improvement activities rather than recurrent spending. In the CPIP programme, incentive payments were made to districts (the precursor of the Hospital and Health Services, HHS), which were required to pass at least 80% to the clinical units generating the activity, allowing some overhead retention. District managers decided how to split funds across units, but distribution was weakly controlled, and some units struggled to receive their payments.

Payments were issued twice yearly following six-month performance periods, with fixed timelines for data submission and announcement of results. Incentive payments were primarily set on a per-case basis, though some indicators were rewarded at other levels (e.g. unit or district).

The subsequent QIP programmes (2012-2022) also relied on non-recurrent payments linked to achievement of pre-established targets and paid directly to the HHS. Over time, several QIP schemes were progressively converted into alternative purchasing incentives, reflecting an evolution in payment design as the system matured. The approach shifted away from simple bonus payments for target attainment towards incentives linked to the delivery of best practice care for individual patients, thereby moving closer to the logic underpinning the English BPT model. The final QIP schemes were implemented in 2021-2022, and from 2022-2023 onwards they were replaced by other purchasing approaches, including lump-sum payments, activity volume adjustments, and price modifications embedded within activity-based funding arrangements.

The **BPT** system in **England** provides an important example of a bonus-based pay-for-performance model integrated into routine hospital reimbursement. Introduced within the National Health Service (NHS), BPTs aim to encourage the delivery of evidence-based, high-quality care by rewarding providers that meet nationally defined best practice standards with additional payments or enhanced tariffs. Rather than operating through separate incentive pools, BPTs are embedded directly within the activity-based payment system as tariff uplifts linked to specific clinical criteria. Hospitals that deliver care in accordance with recognised best practice pathways receive a higher reimbursement than under the conventional tariff, thereby creating a direct financial reward for quality improvement. Although providers that fail to meet the criteria receive only the lower tariff, the system primarily functions as a positive incentive mechanism centred on rewarding achievement rather than imposing explicit penalties.

By contrast, systems such as Medicare's HVBP programme in the **US** rely on budget-neutral redistribution mechanisms. Under these models, a proportion of standard payments is withheld in



advance and redistributed according to provider performance. This creates a combined reward-and-penalty structure: higher-performing hospitals receive more than they would receive without P4P, while lower-performing providers receive less. Importantly, the total funding envelope remains fixed at the national level, ensuring that incentive payments do not increase aggregate expenditure.

Also **France** represents a hybrid approach. Since the 2026 reform, hospitals are no longer ranked directly against peers but assessed against predefined absolute thresholds set by the *Haute Autorité de santé* (HAS), with performance translated into category-based remuneration rates. This means that funding is redistributed across providers according to achieved performance levels rather than purely relative comparison, with higher-performing hospitals receiving a larger share of the available envelope. At the same time, the structure of the scheme implies a redistribution effect: lower-performing hospitals receive reduced payments or, in some cases, no payment for specific indicators, so that gains for better-performing providers are indirectly balanced by lower allocations elsewhere within the fixed budget envelope.

Failure to meet nationally defined indicators in the **CQUIN** programme resulted in end-of-year reductions in the variable component of API payments. This implies that if a provider did not achieve its targets, it could not receive the full expected budget. The scheme can therefore be interpreted as a penalty mechanism, relative to the baseline contract value.

4.5.3 Achievement or improvement incentives

A second key design feature across countries is the balance between rewarding absolute achievement and rewarding improvement over time. Many systems now explicitly combine both approaches to ensure that incentives apply to hospitals at different performance levels.

In **Medicare's** HVBP programme, hospitals receive points based on achievement relative to national benchmarks as well as improvement relative to their own historical performance. This dual structure ensures that high-performing hospitals are incentivised to maintain excellence, while lower-performing hospitals are encouraged to improve from their baseline. For each indicator, the higher of the two scores is used in the final calculation.

A similar dual logic was present in **France's** IFAQ programme until 2026. For part of the indicators (based on *indicateurs de qualité et sécurité des soins* or IQSS – see section 4.6.2) achievement- and improvement-based incentives were combined, with most weight placed on achievement. Hospitals were primarily rewarded for reaching predefined targets set by the HAS, with lower thresholds ensuring that partial payment is possible even for lower-performing providers. Hospitals were assessed against other providers, with performance measured through comparative rankings that determined eligibility for funding. When targets were not met, selected indicators also rewarded improvement over time, so hospitals could still receive partial remuneration if performance was stable or improving. However, this improvement dimension did not apply to all indicators. For most remaining indicators, only achievement was considered. For example, for infections and thromboembolic events after hip and knee replacement, remuneration was based on risk-adjusted performance, with full payment only when outcomes were within an acceptable statistical range; otherwise, no payment was made.

Since the 2026 reform, the IFAQ programme has moved towards a more structured achievement-based system while retaining a limited and more targeted improvement component. Hospitals are assessed against absolute performance thresholds defined in advance by the HAS and are allocated to predefined performance categories that determine the level of remuneration for each indicator. The primary logic of the scheme is therefore one of achievement: higher payments are awarded to providers that meet or exceed specified standards of care. At the same time, an improvement element is retained. The improvement component is structured to provide stronger financial incentives for lower-performing hospitals, with higher adjustment rates for poorer

performance classes. Deterioration in performance may also result in penalties, particularly for lower-performing hospitals.

The **CQUIN** framework combined both improvement- and achievement-based incentives, but its emphasis shifted over time. Early schemes were locally negotiated and primarily focused on improvement, with targets set relative to a provider's own baseline or adjusted to local circumstances, encouraging progress rather than absolute performance levels. Payments were based on each provider's performance without peer comparison.

Between 2013 and 2018, as national indicators became more established, the system increasingly incorporated achievement-based targets, often defined by fixed national thresholds, particularly in acute care. However, some areas – such as specialised services – continued to allow local adaptation, and certain indicators blended both approaches by rewarding improvement against baseline as well as attainment of national benchmarks.

From 2019 onwards, the framework moved more clearly towards achievement-based incentives, with payments linked to whether providers met nationally defined lower and upper thresholds. Funding was only triggered once minimum standards were reached, with full payment awarded at higher levels of performance.

Best Practice Tariffs are primarily achievement-based, not improvement-based. They pay a fixed higher tariff when providers meet defined clinical standards or best practice criteria for a given procedure or condition (for example, timely surgery for hip fracture or specific care pathways). If the criteria are met, the provider receives the higher tariff; if not, it receives a lower tariff. There is generally no financial reward for incremental improvement over time unless it results in meeting the specified best-practice standard. So the mechanism is designed to incentivise compliance with evidence-based care pathways (achievement of standards), rather than gradual improvement per se.

In **Queensland**, early P4P schemes focused primarily on absolute achievement of predefined targets. Over time, however, the system progressively shifted from bonuses for target achievement to bonuses for the achievement of best practice for individual patients.

4.5.4 *Tournament- or threshold-based system*

Across the programmes reviewed, P4P designs can be broadly distinguished between threshold-based and tournament-based mechanisms, although several systems combine elements of both. Threshold-based designs link payment to the achievement of predefined performance levels, which may be absolute targets or improvement thresholds. In contrast, tournament-based systems allocate a fixed incentive pool that is redistributed according to relative performance between providers, meaning that one provider's gain is offset by another's loss. A more detailed description of the calculation methods used in each programme can be found in the respective programme reports in [Supplement 1](#) to the Scientific Report.

In **Queensland**, P4P mechanisms were predominantly threshold-based. Payments were awarded when HHS achieve predefined targets set at state level, rather than through ranking hospitals against each other. These thresholds were generally absolute and defined within the Queensland Healthcare Purchasing Framework and associated service agreements. There was no formal competitive ranking or selection of winners and losers across providers. This reflects a deliberate policy choice to avoid tournament-style competition, which was considered potentially discouraging for lower-performing hospitals and risked disengagement.

Best Practice Tariffs in **England** are also threshold-based rather than tournament-based. They provide a fixed higher payment when providers meet clearly defined clinical criteria or best practice standards for specific conditions or procedures. If the required standard is achieved, the provider receives the enhanced tariff; if not, it receives the standard tariff. Importantly, payments are not determined by relative performance between hospitals, and there is no ranking or redistribution of a fixed budget across providers. Instead, each provider is assessed independently against absolute



thresholds, meaning that multiple hospitals can simultaneously qualify for the higher tariff, or none at all if standards are not met.

The design of the **CQUIN** programme has evolved over time from a purely binary threshold-based model to a more graduated threshold-based system. Under earlier CQUIN arrangements, payments were conditional on meeting predefined targets or milestones, with binary thresholds determining whether funding was earned. Over time, the programme moved towards a proportional payment structure in which remuneration varies according to performance between lower and upper thresholds. Providers are therefore no longer assessed in a purely all-or-nothing manner but can receive partial payment depending on their level of achievement within a defined range. Importantly, payments are based on each provider's own performance against predefined standards rather than on comparison or ranking against other providers. As a result, CQUIN remains a threshold-based system, albeit with graduated or linear payment scaling, rather than a tournament-based model.

The **Medicare** HVBP programme is best characterised as primarily tournament-based, although it incorporates structured threshold elements. Hospitals are evaluated relative to national performance benchmarks as well as their own historical performance, and a fixed incentive pool created through withheld DRG payments is redistributed based on comparative performance. This means that hospitals compete indirectly for a share of a fixed budget, a defining feature of tournament design where total payments are constrained and redistributed across providers. At the same time, the use of national achievement thresholds and improvement baselines introduces a structured scoring system within this competitive framework, creating a hybrid between tournament allocation and threshold-based assessment.

Before the 2026 reform, **France** used a hybrid approach that combined threshold-based scoring with elements of tournament-like redistribution. At indicator level, performance could be assessed using absolute targets, linear progression between thresholds, or binary rules depending on the indicator type, while some indicators also incorporated improvement over time. These results were aggregated into a composite hospital-level score. However, funding allocation also depended on relative performance within predefined hospital groups, meaning that hospitals were effectively assessed against peers with similar activity profiles. Because funding was redistributed within a fixed national envelope according to comparative rankings, the system generated a quasi-tournament effect at group level.

Since the 2026 reform, this relative peer-ranking logic has largely been replaced by assessment against predefined absolute thresholds established by the HAS. Hospitals are now classified into performance categories linked to predetermined remuneration rates for each indicator. Nevertheless, the system remains hybrid because, although the explicit peer-ranking mechanism has been removed, funding continues to be distributed from fixed indicator-specific envelopes. As a result, the system has shifted away from a hybrid model towards a predominantly threshold-based approach centred on absolute performance standards, while still retaining indirect redistributive effects within the overall funding envelope.

4.5.5 *Performance measurement and scoring approaches*

Across all systems, individual performance indicators are translated into financial consequences through structured scoring methods, although the degree of complexity and standardisation varies considerably between programmes. These scoring methods determine how measured performance is converted into incentive payments.

Some systems rely primarily on binary or threshold-based scoring, in which payment depends on whether a predefined target or milestone is achieved. Under this approach, providers either receive the incentive or do not, with limited differentiation between levels of performance above or below the threshold. This model characterised earlier **CQUIN** schemes in England, **BPTs** and parts of **Queensland's** incentive programmes, where hospitals received payments only when specific standards or activity targets were met. Queensland historically relied more strongly on

indicator-specific payments, although aggregation sometimes occurred at hospital or district level when indicators covered multiple clinical areas.

Other systems use continuous or proportional scoring methods, where payment increases gradually according to performance between lower and upper thresholds. This creates more nuanced differentiation between providers and rewards partial achievement rather than only full target attainment. **Medicare's** HVBP programme uses this type of approach by assigning scores based on both achievement relative to national benchmarks and improvement relative to a hospital's own baseline performance. Individual weighted domain scores are then aggregated into a total performance score, which determines payment adjustments within a fixed national funding pool. More recent **CQUIN** arrangements similarly moved towards graduated payment structures with proportional remuneration between minimum and maximum thresholds.

The **Medicare** HVBP programme has also been criticised for potentially disadvantaging hospitals serving socioeconomically vulnerable populations. Because performance adjustment accounts for medical but not broader social risk factors, concerns have been raised that the programme may reinforce existing inequalities by reducing payments to hospitals caring for low-income or marginalised groups. Although a Health Equity Adjustment (HEA) was proposed for implementation in 2025-2026, it was ultimately removed from the final payment rules.

Before the 2026 reform, **France** combined several scoring approaches within its IFAQ framework. Some indicators used proportional scoring between thresholds, while others relied on binary or stepwise rules, particularly for accreditation measures and infection-related indicators. In addition, certain indicators incorporated both achievement and improvement dimensions. Individual indicator scores were then aggregated into a weighted composite hospital-level score. However, final remuneration also depended partly on performance relative to comparable hospitals within predefined peer groups, introducing a quasi-competitive component into the allocation process.

Since the 2026 reform, hospitals are no longer ranked against peer groups and indicator results are no longer aggregated into a single composite score. Instead, each indicator is linked to its own funding envelope and remuneration scale based on absolute performance thresholds defined by the HAS.

4.5.6 *Allocation level and internal distribution of incentives*

A further important design dimension concerns where incentives are allocated and how they are distributed within provider organisations.

In the **Queensland** CPIP programme, payments were made to districts (the precursor of the Hospital and Health Services (HHSs)), which were expected to redistribute most of the funding to the relevant clinical units or services responsible for performance. However, this internal allocation process was only partly regulated, resulting in variability between organisations in how incentives were ultimately used. The underlying intention was to link rewards to frontline teams while preserving managerial flexibility at organisational level.

In contrast, in **England** (**CQUIN** and **BPTs**), **France**, the **Medicare** HVBP programme and QIP programmes in Queensland, payments are generally made at provider level and integrated into overall organisational budgets rather than formally earmarked for redistribution within hospitals. Under CQUIN and HVBP, incentive payments or payment adjustments are absorbed into hospital finances as part of broader contractual or reimbursement arrangements. Similarly, BPTs provide enhanced reimbursement when predefined standards are met, but the additional revenue accrues to the hospital as a whole rather than being directly allocated to specific clinical teams. In France, IFAQ payments are likewise made at institutional level, with no formal requirement to redistribute funds internally to departments or individual clinicians.



Table 13 – Key features of selected P4P programmes (most recent status)

	Hospital Value-Based Purchasing (HVBP)	Commissioning for Quality and Innovation (CQUIN)	Best Practice Tariffs (BPTs)	<i>Incitation financière à l'amélioration de la qualité</i> (IFAQ)	Quality Improvement Payments (QIPs)
<i>Country</i>	Medicare (US)	England	England	France	Queensland (Australia)
<i>Introduction year</i>	2012	2009	2010	2012	2012
<i>Status of the programme</i>	Active	On hold	Active	Active	Discontinued
<i>Budget impact</i>	Budget neutral	Budget neutral	Budget neutral	Part additional budget and part budget neutral	Additional budget
<i>Financial incentive</i>	Bonus Penalty	Penalty	Bonus	Bonus Penalty	Bonus
<i>Achievement or improvement</i>	Achievement Improvement	Achievement	Achievement	Achievement Improvement	Achievement
<i>Threshold or tournament</i>	Threshold Tournament	Threshold	Threshold	Threshold	Threshold
<i>Allocation level</i>	Hospital	Hospital	Hospital	Hospital	Hospital
<i>Indicator type</i>	Outcome, process	Process, PREM	Outcome, process	Outcome, structure process,	Process

4.6 Process for identifying and selecting quality domains and indicators

4.6.1 Policy-driven prioritisation of quality domains

The selection of domains is often guided by broader policy priorities.

In **Queensland**, quality domains for P4P schemes were explicitly defined within the state's healthcare purchasing framework and service agreements between the Department of Health and HHSs. From the outset, the selection of domains and indicators was closely linked to policy concerns about patient safety and care quality. This context led to a strong emphasis on identifying areas where care improvement was both measurable and clinically meaningful. Quality domains were therefore not defined as abstract categories, but rather translated into specific clinical priority areas through structured engagement with state-wide clinical networks. These networks ensured that selected domains reflected both system priorities and frontline clinical relevance, covering areas such as acute care pathways, patient safety, and priority populations. Over time, some stand-alone P4P schemes were phased out or absorbed into broader purchasing arrangements, with quality priorities increasingly embedded within routine activity-based funding rather than operating as separate programmes.

In **England**, the early design of the **CQUIN** framework was explicitly structured around four core dimensions of quality: safety, effectiveness, patient experience, and innovation. Commissioners and providers were expected to include indicators covering each of these domains, thereby ensuring a balanced approach to quality improvement. Over time, the framework evolved in line with national policy agendas particularly following a major evaluation in 2013. From 2019 onwards, CQUIN indicators for both acute and specialised acute services were aligned with priority areas defined in the NHS Long Term Plan, such as prevention, patient safety, and best practice pathways. Although the formal structuring of domains has become less explicit in recent guidance, these priorities continue to shape indicator selection.

The **BPT** system reflects policy-driven prioritisation but operates at the level of individual clinical pathways rather than programme-wide domains. Since 2010, the introduction, modification, and removal of BPTs has followed a broadly stable governance model. The underlying policy objective has remained consistent: to reduce unwarranted variation and incentivise evidence-based, high-quality and efficient care. In line with this, BPTs translate national priorities into payment rules applied at the level of specific clinical episodes. Priority service areas are selected where financial incentives are expected to generate measurable improvement, typically in settings characterised by high patient volumes, significant variation in outcomes, or clear scope for efficiency gains. In earlier years, this focused largely on acute hospital pathways such as stroke, hip fracture, and day-case surgery, whereas more recently the scope has broadened to system-wide pressures such as waiting list reduction, care reconfiguration, and shifting activity to more appropriate settings.

BPTs are therefore introduced, revised, or withdrawn in response to changes in clinical evidence, policy priorities, data availability and coding capability, the effectiveness of existing incentives, and broader reforms to NHS payment models.

In the **US**, the Medicare HVBP programme illustrates a more incremental and legislatively structured evolution of quality domains. Initially, the programme focused primarily on clinical process measures and patient experience. Over time, additional domains were introduced: outcome measures were added first, followed by efficiency and cost-related measures, and later safety indicators. This gradual expansion reflects a policy shift from process compliance towards comprehensive value-based care. Since 2017-2018, the programme has consistently evaluated four equally weighted domains: clinical outcomes, patient experience, safety, and efficiency.

In **France**, the IFAQ programme reflects a policy-driven expansion of quality domains over time. The programme is overseen by the Ministry of Health, specifically the Directorate for Healthcare Provision (DGOS), and was initially developed through preparatory work led jointly by DGOS and the French National Authority for Health (HAS), with support from several national agencies. Its quality framework



covers domains such as clinical quality, patient experience, organisational quality, and selected safety outcomes. However, these are not fixed categories: they are implemented through evolving indicator groups that are regularly updated to reflect national priorities for hospital performance and system efficiency.

4.6.2 *Clinical and expert-driven selection processes*

Across countries, the selection of quality domains and indicators for P4P programmes is typically grounded in structured processes. In **Queensland**, state-wide clinical networks played a central role in this process. These networks bring together healthcare professionals, patients, caregivers, and other stakeholders across a wide range of specialties and population groups. They are directly responsible for developing and endorsing indicators within their respective clinical areas, ensuring that selected measures are both evidence-based and clinically relevant. The CPIP pilot programme concentrated on a small set of clinical areas selected based on high disease burden, significant variation in provider performance, strong evidence of effectiveness, clinician support and control, and the availability of real-time (preferably electronic) data collection.

The selection of indicators followed an iterative process. It began with a draft set focusing on diagnoses and procedures where gaps existed between current performance and evidence-based practice, particularly in areas considered to be within clinicians' control. This was followed by extensive stakeholder consultation lasting approximately twelve months, aimed at strengthening the legitimacy and acceptability of the indicators. A further objective was to assess whether the clinical community shared a common understanding of appropriate care processes and of the indicators used to measure their quality. This approach reflects a strong emphasis on aligning financial incentives with professional standards and practice. Early experience showed that clinician scepticism was closely linked to concerns about indicator relevance, highlighting the importance of co-design. Although clinicians generally supported the continuation of the programme, they reported limited perceived impact on their own behaviour, suggesting that engagement does not automatically translate into behavioural change.

In **France**, the HAS is responsible for the development of quality and patient safety indicators (*indicateurs de qualité et sécurité des soins* – IQSS). Most IFAQ indicators are IQSS. Their selection follows a structured process, including literature review, assessment of existing recommendations, and consensus-based validation by an expert group (see further). Not all IQSS indicators are included in IFAQ; the selection is jointly decided by the HAS and DGOS, although the criteria are not publicly specified.

The process is supported by a dual governance structure involving both a multidisciplinary expert working group and a broader stakeholder group. The expert group, composed of clinicians, data specialists, and patient representatives, is responsible for the scientific and clinical validation of indicators, including defining target populations and risk adjustment factors. In parallel, the stakeholder group, comprising professional organisations and patient associations, plays a consultative role by reviewing results, identifying potential issues, and supporting dissemination. This structured approach ensures methodological robustness, although recent evaluations suggest that greater involvement of frontline professionals may be needed to strengthen engagement with the P4P programme.

In the **US**, the selection of indicators within the Medicare HVBP programme also follows a structured and iterative process that combines expert consensus, technical evaluation, and public consultation. A central role in indicator development is played by a consensus-based entity, historically the National Quality Forum (NQF). This multi-stakeholder body brings together representatives from providers, payers, patients, professional associations, and researchers to review and endorse quality measures. Although the 'Centers for Medicare & Medicaid Services' (CMS), which administers the Medicare programme, is not formally required to use only NQF-endorsed measures, endorsement is widely regarded as a marker of methodological robustness and policy relevance, and many HVBP indicators originate from this process.

In parallel, CMS conducts an annual rulemaking process in which proposed indicators, domains, benchmarks, and weighting structures are published in advance. These proposals are open to public comment from hospitals, professional bodies, researchers, and other stakeholders. All submissions are reviewed and may lead to revisions before final rules are issued. This creates a formal mechanism for transparency and iterative refinement of the indicator set.

Indicator selection is also guided by a defined set of technical and policy criteria. For example, the initial HVBP indicator set was drawn from the Hospital Inpatient Quality Reporting (IQR) programme, prioritising measures linked to high-mortality conditions, high-volume and high-cost care, and areas with known variation in practice despite existing clinical guidelines. Several indicators were excluded early on due to limited variability ('topped out' performance), insufficient reporting history, or concerns about unintended consequences (such as promoting inappropriate antibiotic use). Others were removed when structurally unsuitable for payment purposes or when stronger or more clinically relevant alternatives became available.

More broadly, the indicator set is reviewed annually using formal exclusion criteria, including lack of alignment with current clinical guidelines, weak association with patient outcomes, excessive measurement burden, feasibility constraints, or disproportionate administrative costs relative to expected benefits. As a result, the HVBP indicator framework is dynamic, with regular updates reflecting both evolving evidence and changing policy priorities.

In **England**, clinical and expert input is also central to indicator development, although the process combines national standard-setting with elements of stakeholder consultation and, historically, local adaptation. Within the **CQUIN** framework, indicators were initially developed through negotiations between NHS commissioners and providers, drawing on national clinical guidelines, local service priorities, and identified quality gaps. This allowed indicators to be tailored to local contexts, particularly in the early phases of the programme.

Over time, however, the system moved towards greater national standardisation. Indicator design became more centrally defined, with increasing alignment to national policy priorities set by NHS England and related bodies. Clinical expertise continues to play a key role through involvement of advisory groups, guideline development processes, and expert committees that inform the selection, refinement, and technical specification of indicators. This shift has reduced local variation in indicator design while strengthening consistency, comparability, and alignment with national quality objectives across providers.

A particularly strong link between clinical evidence and indicator selection is observed in the development of **BPTs** in England. These tariffs are explicitly grounded in evidence-based clinical guidelines, most notably those produced by the National Institute for Health and Care Excellence (NICE) and translate recommended care pathways into measurable criteria that define high-quality care for specific conditions or procedures. Indicator design therefore begins from clinical guidance and identifies discrete, auditable steps in the care pathway that can be reliably captured using routine hospital data.

BPT indicators typically specify a combination of clinical actions and timing requirements (for example, timely surgery, appropriate diagnostic testing, or adherence to defined perioperative processes), which are then embedded within administrative and clinical datasets. This ensures that compliance can be systematically monitored using existing hospital information systems, without requiring additional data collection.

As a result, the selection process for BPT indicators is highly standardised and tightly anchored in clinical consensus. Unlike more locally adapted or negotiated incentive schemes, BPTs operationalise nationally agreed best practice as a set of uniform rules applied consistently across providers. Financial incentives are therefore directly linked to adherence to validated clinical standards at the level of individual patient episodes, reinforcing guideline-concordant care rather than broader organisational performance targets.



4.7 Types of quality indicators: structure, process and outcome measures

Across the P4P programmes reviewed, indicator sets show a clear differentiation between structure, process, and outcome measures, although their relative balance and evolution over time varies by country and programme maturity. In addition, several systems include broader categories such as organisational or system-level proxies (e.g. accreditation or digital maturity).

Patient experience indicators are treated as a separate category in the following sections, reflecting their specific methodological basis (typically survey-derived data) and their focus on perceived quality of care rather than clinical processes or outcomes (see Table 13).

OUTCOME INDICATORS: CLINICAL RESULTS AND PATIENT OUTCOMES

Outcome indicators are most extensively developed in the **US Medicare** HVBP programme, where they form a distinct clinical outcomes domain. This domain includes multiple risk-standardised mortality indicators covering key conditions, calculated over a long observation period of several years to ensure stability. These indicators measure 30-day mortality following hospital admission for specific conditions, adjusted for patient risk profiles to allow comparison across hospitals. In addition, outcome measurement includes a risk-standardised complication rate following hip or knee replacement, capturing a defined set of post-operative adverse events, including mortality and major complications within a specified time frame.

In **France** (IFAQ), outcome indicators are also prominent but more heterogeneous. In the IFAQ programme in 2025, indicators consist of six outcome indicators, four process indicators, and one structure indicator. Outcome indicators include two patient-reported experience measures (PREMs) (see further) as well as four clinical outcome indicators related to surgical complications, particularly thromboembolic events and infections following hip and knee replacement. For these surgical outcome indicators, risk adjustment is applied using patient characteristics, including age, gender, a set of ten comorbidities, and the time between surgery and discharge. In the 2026 programme, the thromboembolic events and infections following hip and knee replacement indicators are no longer included.

In **CQUIN (England)**, outcome indicators have historically been limited. Early guidance encouraged their use, but most indicators remained process-based, alongside some structure indicators, particularly in specialised services. Over time, outcome measurement has remained a minority component, and, in the final years, was largely limited to PREMs (see further).

Outcome indicators play a limited and indirect role in **BPTs**, which are primarily structured around process measures derived from evidence-based clinical pathways. The design rationale is that many outcomes, such as mortality, complications, or readmissions, are influenced by patient characteristics and case-mix, making them less suitable as direct payment triggers. As a result, BPTs focus financial incentives on adherence to specified processes of care, while using outcome measures mainly for monitoring, validation, and quality assurance.

Where outcome indicators are used, they typically take the form of proxies or risk-related measures rather than direct payment determinants. For example, in the hip fracture BPT, outcomes such as 30-day mortality and length of stay are monitored as indicators of overall quality and system performance, but they do not directly determine eligibility for the enhanced tariff. Instead, key process requirements, such as timely surgery and orthogeriatric assessment, are the primary drivers of payment, with outcomes used to assess whether these processes are associated with improved results.

In the hip and knee replacement BPT (see [Supplement 1](#) to the Scientific Report), outcome-related elements include monitoring of revision surgery rates and postoperative complications, alongside structured prevention processes such as thromboprophylaxis and infection control pathways. Similarly, 30-day readmissions may be tracked as a quality indicator, but they are not typically used

as a direct payment trigger within the tariff. Instead, they serve as signals of resulting performance and are used to assess the broader effectiveness of the care pathway.

Within this framework, Patient-Reported Outcome Measures (PROMs) are also collected for patients undergoing hip and knee replacement through standardised instruments such as the Oxford Hip Score, Oxford Knee Score, and EQ-5D, administered before and after surgery via the national PROMs programme. These measures capture functional improvement and changes in health-related quality of life from the patient perspective. However, PROMs are primarily used for monitoring variation in outcomes across providers, benchmarking performance, and supporting quality improvement rather than directly influencing tariff payments.

Overall, outcome indicators in BPTs, including clinical outcomes such as complications and readmissions as well as PROMs, are primarily used in a supporting role. They help validate that adherence to best-practice processes is associated with improved patient outcomes and provide system-level feedback on quality, but they do not generally form the basis of financial remuneration.

In **Queensland**, outcome indicators were initially less prominent in early CPIP phases, which prioritised measurable clinical processes. However, structure indicators were also included alongside process measures. Outcome measurement was considered an important longer-term objective, with the expectation that programmes would progressively evolve from process-focused indicators toward more patient-centred outcomes such as symptom improvement and access-related results. However, in later QIP schemes, no such outcome indicators were ultimately included.

PROCESS INDICATORS: CLINICAL ACTIVITIES AND CARE PATHWAYS

Process indicators constitute the core of most P4P systems, particularly in England, France, and Queensland. They measure adherence to evidence-based clinical practices and are typically easier to operationalise and attribute to provider behaviour.

In the **Medicare** HVB programme, process indicators are concentrated mainly within early programme phases and have gradually been reduced in relative importance compared to outcome measures. When present, they are embedded within broader domains such as safety processes (e.g. sepsis management protocols introduced later as a process indicator within the safety domain). However, the programme increasingly prioritises outcome and efficiency metrics over pure process measures.

In **Queensland**, process indicators were the primary focus of the CPIP programme. These indicators were often tightly linked to clinical pathways and designed to reflect clearly defined steps in care delivery. Process indicators were deliberately favoured because of their perceived measurability and direct influence on outcomes. This emphasis on process-based measurement was also a defining feature of the QIP programmes, which similarly prioritised structured, observable elements of care delivery within defined clinical pathways (such as the percentage of patients with fractured neck of femur taken to theatre within two days).

In **France** (IFAQ), process indicators remain a central component and are explicitly organised under domains such as appropriateness of care, infection prevention practices, and care coordination. Examples include adherence to pain management guidelines, discharge letter quality, and infection prevention protocols. Process indicators are also used at hospital-wide level and are often linked to organisational practices rather than individual clinical episodes.

Process indicators are the dominant form of measurement in the **CQUIN** programme in **England**, particularly from 2019 onwards when national priorities structured schemes around care pathways. These include timely antibiotic switching, early frailty identification, NEWS2 (National Early Warning Score 2) scoring in critical care, and adherence to cancer diagnostic pathways. In specialised services, process indicators are also used to measure timeliness and adherence to clinical guidelines (e.g. revascularisation timing or hepatitis C treatment initiation). Across all phases, process indicators consistently outnumber outcome measures. The process indicators were not risk-adjusted beyond the delineation of the relevant patient group.



Process indicators are the central design feature of **BPTs in England**. They specify the delivery of particular evidence-based clinical actions within defined care pathways, rather than measuring outcomes directly. These indicators typically relate to the timing, sequencing, and completeness of care, for example, timely surgery for hip fracture, appropriate pre- and post-operative assessments, or adherence to perioperative prophylaxis and infection prevention protocols. The underlying rationale is that these processes are strongly supported by clinical evidence and are within the direct control of providers, making them suitable for financial incentivisation. By tying additional reimbursement to compliance with these specified steps, BPTs aim to standardise care delivery across providers and improve the consistency of evidence-based practice.

STRUCTURE INDICATORS: ORGANISATIONAL CAPACITY AND SYSTEMS

Structure indicators are less common overall but play a targeted role in several programmes, particularly in early-stage or system-building contexts.

Structure indicators were included in early CPIP phases in **Queensland**, such as the COPD pulmonary rehabilitation programme, which was assessed at district level based on compliance with an endorsed clinical standard. These indicators often reflected the presence of services, systems, or protocols rather than direct clinical actions. However, structure indicators were relatively limited compared to process measures.

In **France**, one hospital-level structure indicator assesses both participation in the accreditation procedure and its outcome. Since the 2026 reform, the accreditation indicator has gained significantly increased importance within the IFAQ framework. Rather than aggregating indicators into a single hospital-level score, each indicator is funded through a separate financial envelope, with its size reflecting its relative weight. In this new structure, nearly half of the total funding is now allocated to accreditation, highlighting its central role in the system. In addition, hospitals accredited with a 'high quality' rating receive an increased remuneration rate of 110%. Even non-accredited facilities may still access conditional funding, provided they submit and implement an approved action plan, reflecting the programme's intention to use accreditation as a lever for quality improvement and organisational change.

Structure indicators have been used more frequently in specialised services in the **CQUIN** programme in **England**, particularly in earlier phases, where they supported system development. Examples include establishment of multidisciplinary teams, regional networks, and data infrastructure. These were often considered enabling conditions for quality improvement, for example to support fixed-cost investments associated with service innovation, rather than direct performance measures.

Structure indicators play a still important role in **BPTs in England**, although they are far less prominent than process indicators. An example from current BPTs is the requirement for access to and involvement of specialist multidisciplinary teams, such as diabetes specialist teams in paediatric diabetes care. In these cases, eligibility for the enhanced tariff depends not only on individual clinical processes but also on the presence of an organised care structure capable of delivering coordinated, guideline-based management.

Overall, structure indicators in BPTs are used to ensure that providers have the necessary organisational foundations in place to support high-quality care, but they are less common than process indicators and are typically embedded within broader pathway requirements rather than used as standalone payment determinants.

In **Medicare HVBP**, structural indicators are largely absent from the current framework, reflecting a deliberate shift toward outcomes and process measures with clearer links to patient results.

PATIENT EXPERIENCE-BASED MEASURES

Patient experience measures are typically treated as a distinct category because they capture how patients perceive and experience care, rather than its technical or clinical performance. Unlike structure indicators (resources and organisation), process indicators (delivery of clinical actions), and outcome indicators (clinical results), they focus on interpersonal and service aspects of care delivery.

These measures are usually reported directly by patients and include domains such as communication, involvement in decision-making, dignity and respect, coordination of care, waiting times, and overall responsiveness. They differ from satisfaction surveys by focusing on specific, observable aspects of care rather than overall subjective approval.¹⁷⁶

Because they reflect the patient perspective on care delivery, patient experience measures are increasingly recognised as a core component of patient-centred care and are frequently incorporated into quality improvement and P4P schemes alongside clinical indicators.

Patient-reported measures form a distinct category that appears across all systems, but they are most systematically developed in the Medicare HVBP programme and France's IFAQ framework. In **Medicare**, the 'Hospital Consumer Assessment of Healthcare Providers and Systems' (HCAHPS) survey provides a standardised set of patient experience indicators covering areas such as communication with staff, discharge information, hospital environment, and responsiveness. These measures are risk- and case-mix adjusted to ensure comparability across providers.

In **France**, PREMs are incorporated more explicitly as part of the quality framework, applied both at hospital level and within specific care pathways such as outpatient surgery, reflecting their growing role in defining patient-perceived quality of care. For patient experience and satisfaction indicators, scores are risk-adjusted using two patient-reported factors: overall life satisfaction and perceived health improvement after the hospital stay.

In **England's CQUIN** programme, PREMs are used but are less systematically structured, typically appearing as individual indicators rather than a dedicated domain. In **Queensland**, patient experience measurement was not included in the respective P4P programmes.

Patient experience has an important but largely indirect role in the paediatric diabetes BPT. The tariff requires participation in the National Paediatric Diabetes Audit (NPDA), including the collection of PREMs, thereby incorporating the views of patients and families into the assessment of care quality in diabetes services. More generally, however, patient experience indicators play a relatively limited role across the broader set of **BPTs**, which tend to focus more heavily on clinical processes and outcomes than on experience measures.

4.8 Integration of P4P and other quality initiatives within hospitals

P4P WITHIN BROADER PERFORMANCE GOVERNANCE FRAMEWORKS

P4P is only one element within broader national systems for managing hospital quality. Across countries, hospital performance is primarily governed through formal accountability frameworks that combine monitoring, regulation, and financial and non-financial instruments.

In **Queensland**, this is operationalised through the Health Performance and Accountability Framework applied to Hospital and Health Services, which are accountable for meeting defined performance standards while retaining some autonomy in service prioritisation. Performance is typically monitored through key performance indicators and safety-related measures. **France** relies on national quality indicators and contractual arrangements between regional health agencies and hospitals, while **England** uses broader strategic planning to set priorities, the NHS Standard Contract, national reporting systems, and public dashboards. In the **US**, Medicare combines multiple value-based purchasing and reporting programmes to monitor performance. Across systems, these frameworks define expected standards and escalation pathways, with P4P functioning as one lever among several.



COMPLEMENTARY FINANCIAL AND NON-FINANCIAL QUALITY MECHANISMS

Across systems, P4P programmes operate alongside a broader set of financial and non-financial instruments aimed at improving quality and safety, indicating that P4P constitutes only one element within a wider quality governance framework.

Complementary financial mechanisms include targeted funding adjustments, penalties, and alternative payment arrangements linked to quality objectives. In **France**, for example, bundled payments exist for chronic renal disease patients, while in **Queensland** hospital funding may be adjusted based on safety-related indicators such as hospital-acquired complications, avoidable readmissions, or sentinel events. In the **US**, Medicare combines the HVBP-programme with additional payment adjustments through initiatives targeting hospital-acquired conditions and readmissions, as well as penalties for incomplete reporting of mandatory quality measures. More broadly, these mechanisms are embedded within activity-based funding systems, meaning hospitals are exposed to multiple and interacting financial incentives rather than a single P4P signal.

Non-financial mechanisms focus on monitoring, accountability, transparency, and quality improvement support. Across countries, P4P programmes coexist with inspection systems, patient safety reporting frameworks, benchmarking tools, public reporting platforms, and contractual quality improvement requirements. In England, for example, **CQUIN** operated alongside Care Quality Commission inspections, never events policies, Service Development and Improvement Plans, and Data Quality Improvement Plans. Similarly, **BPTs** are supported by national audits, registries, and administrative datasets used to verify compliance and monitor care pathways.

Performance reporting and feedback mechanisms also vary across systems in their level of integration and transparency. In Queensland, HHS are monitored through indicators covering stewardship, outcomes, and safety/quality, with quarterly review and public reporting, although indicators used for funding, monitoring, and transparency are not always aligned. In the **US**, hospital performance data are collected through the Hospital Inpatient Quality Reporting programme and publicly reported through Hospital Compare, while hospitals additionally receive confidential feedback reports to validate and improve performance before final scoring. In **France**, quality indicators related to IFAQ and IQSS are publicly available and hospitals are required to disseminate results locally, although patient use of these data appears limited. In **England**, providers report performance through multiple channels, including contractual reporting requirements, annual Quality Accounts, audits, and national registries. Under **BPTs**, providers must submit standardised clinical and administrative data regardless of payment eligibility, enabling monitoring, benchmarking, and commissioner verification of compliance.

Taken together, these findings illustrate that quality improvement strategies rely on interacting regulatory, transparency, monitoring, and incentive-based approaches. P4P programmes therefore function within a broader ecosystem of complementary mechanisms rather than acting as standalone levers for changing provider behaviour.

4.9 Additional design and implementation features of P4P programmes

NUMBER OF INDICATORS

The number of indicators used in P4P programmes varies widely both between different programmes and within the same programme over time. Some schemes rely on a relatively small set of indicators, focusing on a few high-priority outcomes, while others use large and complex indicator frameworks covering multiple dimensions of performance.

The number of indicators used in **CQUIN** evolved substantially over time, reflecting a broader shift from locally tailored schemes to greater national standardisation. When introduced in 2009, CQUIN gave providers and commissioners considerable flexibility to develop their own goals and indicators, resulting in highly heterogeneous schemes. By 2010-2011, acute care providers collectively used 1 546 distinct indicators, with hospital-level schemes containing a median of 16 indicators, ranging

from 3 to 52. Over time, concerns about excessive variation, duplication, and difficulties in benchmarking led NHS England to centralise indicator development. Nationally mandated indicators were gradually expanded, and from 2014 a national 'pick-list' of indicators was introduced, later reduced in 2016 to a shorter set of 30 validated indicators across seven priority areas. Local discretion was increasingly restricted, with limits placed on the number of locally selected goals. From 2019 onwards, all indicators had to be selected from a national list, although commissioners and providers retained discretion to negotiate and select up to five indicators per contract. A similar simplification occurred within the specialised services (PSS) CQUIN framework, where centrally designed schemes became progressively fewer and less complex over time before being integrated with the main CQUIN programme in 2023. Overall, the programme moved from a large and diverse set of locally developed indicators towards a smaller, more standardised and centrally managed indicator framework.

Since their introduction in 2010, the number of **BPTs** in England has expanded progressively as the policy has matured and been embedded within national payment arrangements. The initial phase focused on a small number of high-impact clinical areas, including stroke care, hip fracture, cataract surgery, and cholecystectomy, designed as demonstrator tariffs to test the feasibility of linking payment to evidence-based care processes. Over time, the portfolio was steadily broadened to include a wider range of elective, urgent, and long-term condition pathways, as well as emergency care services. By the mid-2010s, BPTs had become a more established component of the tariff system, with additional indicators added annually. In the period leading up to 2026, the system further expanded and was reorganised into distinct categories (including activity-based elective, urgent and emergency care, and annual BPTs), reflecting both the growth in coverage and the increasing sophistication of the design. In the 2026/27 programme, about 20 indicators were related to quality in the hospital setting. Overall, the trajectory from 2010 to 2026 is characterised less by a fixed number of tariffs and more by gradual expansion, refinement, and reclassification within an evolving national payment framework.

The **Medicare** HVBP programme uses a more structured and comparatively stable indicator framework than some other P4P schemes, although the number and composition of indicators have still evolved over time. The programme currently includes indicators across four domains: clinical outcomes, person and community engagement, safety, and efficiency and cost reduction. Together, these domains comprise around twenty indicators, including mortality measures, patient experience indicators from the HCAHPS survey, healthcare-associated infection measures, a severe sepsis management process indicator, and a single efficiency measure related to Medicare spending per beneficiary. While the overall framework has remained relatively consistent, indicators have been periodically added, removed, or revised in response to changing policy priorities, evidence and measurement approaches. For example, the safety domain was recently expanded with the inclusion of a sepsis management indicator from 2025-2026 onwards.

In the initial CPIP pilot phase in **Queensland**, the scheme included seven indicators covering five clinical areas. During the second pilot phase, the framework expanded considerably, with additional clinical areas such as maternity care and diabetes management introduced, alongside new indicators. At the same time, some existing indicators were removed or adapted, resulting in a more comprehensive set of 19 indicators overall. Further modifications occurred during the post-pilot phase, reflecting ongoing refinement of priorities and measurement approaches. Subsequently, within the QIP programmes between 2013 and 2022, the number of indicators gradually declined from five to one, although there were fluctuations between individual years as indicators were added, revised, or retired. This reduction was largely linked to the transformation of QIP arrangements into other purchasing incentives, demonstrating how indicator sets may decline as programmes are integrated into wider health system reform strategies.

In **France**, the number of indicators peaked at 13 in 2022, fell to 7 in 2023, and then increased again to 11 in 2024-2025, and decreased to 8 in 2026.



VOLUNTARY VERSUS MANDATORY HOSPITAL PARTICIPATION

The **Medicare** HVBP programme is mandatory for acute hospitals paid under the Inpatient Prospective Payment System (IPPS). Several groups are excluded, including specialised hospitals such as cancer, children's, and certain long-term care and rehabilitation facilities. In addition, hospitals that do not meet minimum data or case-volume requirements, or fail to qualify in at least three of four performance domains, are excluded from HVBP. Overall, about 80% of eligible IPPS hospitals participated in HVBP in 2024-2025, indicating broad but not universal coverage.

Participation in the CPIP programme in **Queensland** was voluntary, with eligibility for bonus payments requiring clinical units to collect and submit performance data. The main barriers to participation included limited resources (time and staffing), challenges in collecting accurate and complete data, and insufficient leadership and organisational support. By contrast, the QIP programmes became formally embedded in service agreements between Queensland Health and each Hospital and Health Service after the 2012 reform. This meant participation was no longer optional, as QIP indicators were integrated into routine purchasing and funding arrangements that applied to all HHS, making programme involvement effectively mandatory across the system.

Participation in the IFAQ programme in **France** initially began on a voluntary basis in 2012 and remained voluntary through two experimental phases until 2016, when it was generalised and became mandatory for acute hospitals. The second experimental phase remained voluntary but expanded the indicator set to include specialty areas such as stroke and diabetes, with higher financial rewards and broader participation. From 2016, participation became mandatory for acute hospitals, with a simplified system based on existing quality indicators and no new data collection. Although coverage expanded, financial incentives remained relatively limited and only a share of hospitals received payments. In 2019, the programme was reformed and integrated into hospital funding, with a substantial increase in budget and gradual expansion of scope.

BPTs are universally applied where relevant clinical activity falls within their scope, meaning that all eligible episodes of care are assessed against the same nationally defined criteria. However, meeting the associated best practice standards is not a formal requirement for service provision; rather, it determines the level of reimbursement received. Providers that comply with the specified criteria receive the higher tariff, while those that do not meet them are paid at a lower baseline rate. As a result, adherence to best practice is not mandatory from a regulatory perspective, but is strongly encouraged through financial incentives embedded in the payment system.

While participation in the **CQUIN** framework is mandatory, the choice of indicators within the framework illustrates a clear shift over time from a largely voluntary, locally determined approach towards a more mandatory, nationally standardised model with constrained choice. In the early phase (from 2009), providers and commissioners had substantial autonomy to design local schemes, effectively selecting or developing their own indicators and defining associated thresholds and payment conditions. Although national guidance required coverage across broad quality domains, participation in specific indicators was essentially voluntary and driven by local negotiation, resulting in wide variation in indicator sets.

4.10 Evaluations of P4P programmes

The following section reviews formal evaluations of the selected hospital P4P programmes. It synthesises evidence from government reports and scientific literature to assess their design, implementation, and impact on quality of care. While extensive formal evaluations have been conducted in England, France, and the US, no comprehensive official evaluation has been identified for Queensland's P4P programmes. Particular attention is given to common challenges, such as limited effectiveness, complexity, and stakeholder engagement, as well as to the reforms introduced in response to these findings. Further evidence from peer-reviewed evaluations is presented in the systematic review in Chapter 3.

EX-ANTE ASSESSMENT AND PROGRAMME DESIGN CHOICES

In some cases, formal evaluation activities began before implementation.

In **England**, an ex-ante impact assessment of **CQUIN** compared no programme, full national standardisation, and a locally flexible framework. It concluded that locally developed schemes were likely to be most effective, as they better aligned with local priorities, encouraged clinician engagement and ownership, and reduced perceptions of top-down targets. While cost implications were unclear, risks were generally seen as lower under local approaches due to stronger engagement and more realistic goal-setting.

However, concerns were raised that local flexibility could lead to information asymmetry and less ambitious targets, while national standardisation risked overlooking local priorities and imposing uneven levels of ambition. The assessment also noted that locally defined indicators could weaken national benchmarking, although reliance on existing national measures was expected to support gradual standardisation and comparability over time.

In the **US**, the Affordable Care Act mandated prospective evaluation of Medicare's value-based purchasing programmes, embedding evaluation requirements into programme design from the outset, including the HVBP programme. It specifically required the Government Accountability Office to assess the programme's impact by 1 July 2017, with an interim report due by 1 October 2015, focusing on quality of care, expenditure, and differential effects on safety-net hospitals as well as small rural and small urban hospitals.

EMPIRICAL IMPACT EVALUATIONS

Across countries, empirical evaluations generally find limited or mixed effects on quality.

Early evaluations of the **CQUIN** programme (2009-2012) found that its main strength was enabling locally tailored quality priorities, with substantial variation across schemes reflecting local needs. However, there was little evidence of measurable improvements in quality, with only one clear positive effect identified (return-to-home after hip fracture). Key barriers included limited clinical involvement and engagement, highly complex and heterogeneous schemes, weak standardisation and benchmarking, poor adherence to national guidance, heavy data collection burdens, and perceptions that incentives were considered a financial risk rather than driver for quality. Recommendations emphasised clearer separation between national technical design and local priority-setting, fewer and more focused indicators, and avoiding all-or-nothing payment structures.

A later evaluation of CQUIN specialised services (PSS CQUIN, 2016-2019) showed similarly modest and inconsistent effects, with most outcomes showing no significant impact and some unintended negative effects. Although centralisation improved consistency, schemes remained complex, administratively burdensome, and weakly justified in terms of indicator selection and sustainability.

In contrast, **BPT** schemes show more consistent improvements in selected clinical processes, likely due to their simpler design, clearer clinical focus, and stronger alignment with care pathways.

The 2024 *Inspection générale des affaires sociales* (IGAS) evaluation of the IFAQ programme in **France** identified multiple structural limitations:

- Fragmented governance and weak coherence: poor articulation between IFAQ and other quality-improvement levers, reflecting a lack of unified national quality strategy and fragmented responsibilities across actors.
- Complex and opaque incentive design: hospitals often find it difficult to understand how their performance translates into financial rewards. This is largely due to the use of aggregated scoring systems, where multiple indicators are combined into a single composite measure, making it unclear how individual improvements affect overall payments. In addition, the reliance on relative thresholds means that payments depend not only on a hospital's own performance but also on its



performance compared with other hospitals, which further reduces transparency and predictability.

- Weak professional engagement: many indicators are defined at hospital level and therefore have limited direct relevance for frontline clinical practice. As a result, clinicians often find it difficult to translate these measures into actionable improvements within their daily work. Evidence from evaluations suggests that less than half of hospital governing bodies report that the indicators meaningfully mobilise clinical teams around quality and patient safety. In addition, awareness of the indicator sets among staff remains relatively low, with many clinicians having only a partial understanding of the measures being used or their implications.
- Limited financial impact: IFAQ represents only ~1.1% of hospital activity; many hospitals use payments for budget balancing rather than quality investment.
- Limited perceived effectiveness: only 40% of hospitals report positive impact, while 50% report unintended effects (administrative burden, administrative view of quality).
- Limitations of patient-reported measures: low response rates (7-12%), representativeness concerns, and limited interpretability at clinical level.
- Redundancy and missing dimensions: high correlation between indicators, duplication across settings, and missing domains such as staff working conditions.

The IGAS report therefore called for a deep reform of quality governance, including a unified national framework, simplified and transparent incentive design, per-indicator payments, absolute thresholds, stronger outcome orientation (including PROMs), and improved data infrastructure and clinician involvement.

In the **US**, multiple studies on Medicare's HVBP programme found no significant impact on mortality, process quality, or patient experience overall.

An early (2012) evaluation of **BPTs** found mixed effects after their introduction. Uptake varied for locally agreed tariffs (around 52% for stroke, 53% for hip fracture, and 26% for cataract), while cholecystectomy was automatically implemented. Clinicians generally viewed BPTs more positively than CQUIN due to their stronger evidence base, although implementation was sometimes limited by administrative burden and commissioning capacity. Quantitative results showed a 7 percentage point rise in day-case cholecystectomy with no clear quality loss aside from longer waits, no additional short-term impact for stroke beyond existing trends, and modest improvements in hip fracture care, including faster surgery and small gains in outcomes. Overall, early evidence suggests BPTs can change clinical behaviour, but effects vary by pathway and depend on implementation and baseline performance.

DESIGN AND IMPLEMENTATION CHALLENGES

Evaluations consistently highlight complexity and design limitations. In England, **CQUIN** schemes were criticised for excessive complexity, lack of standardisation, and weak benchmarking, particularly due to the large number of locally defined indicators. Later reforms attempted to address these issues by centralising technical design while maintaining some local flexibility. **BPTs**, by contrast, were designed as simpler, condition-specific incentives, reducing administrative burden.

In **France**, the IFAQ system has been criticised for its complexity, lack of transparency, and difficulty in linking performance to financial outcomes, particularly due to aggregated scoring and relative thresholds.

In the **US**, the HVBP programme faces similar concerns, including indicator proliferation, volatility of financial adjustments, and limited incentive size. Financial stakes are considered too small to drive major investments in quality improvement. Because the programme is budget-neutral, all adjustments are centred around zero, meaning hospitals must outperform peers to gain financially, while others inevitably lose. Year-to-year payment adjustments can be highly variable, complicating financial and

operational planning. Moreover, performance differences between hospitals are persistent: top performers tend to remain at the top, while lower-ranked hospitals struggle to close the gap even when improving. As a result, hospitals may still face penalties despite absolute gains in performance.

Concerns also relate to the indicator set and weighting. Some authors argue for fewer, more meaningful and better risk-adjusted measures, while others suggest increasing the weight of mortality indicators to reduce disadvantages for safety-net hospitals. Current weighting may also misalign with patient preferences, as beneficiaries place greater value on clinical outcomes than on other measured domains.

DATA, MEASUREMENT, AND BENCHMARKING ISSUES

Reliable data and comparability are recurring challenges. In **England**, locally defined **CQUIN** indicators limited benchmarking across providers, whereas BPTs, based on national standards, facilitate comparability. Data limitations and reporting burden were also major concerns, particularly in early phases.

In **France**, extensive data verification processes were introduced to ensure reliability, including external audits to detect over-scoring. However, issues remain regarding indicator relevance, delays in data availability, and the limited usefulness of aggregated hospital-level indicators for clinical teams.

In the **US**, data from reporting programmes such as IQR feed into multiple P4P schemes, improving standardisation.

REFORMS AND RECENT DEVELOPMENTS

All systems have undergone reforms in response to evaluation findings.

In **England**, the 2019 **CQUIN** reform aimed to simplify the programme and align it more closely with evidence-based practice. NHS England reframed the scheme to reduce administrative burden and focus on the adoption of clearly defined, nationally supported clinical practices, with the intention of limiting additional implementation costs while promoting wider uptake of proven interventions. In 2024, the CQUIN was paused as part of a broader review of quality incentives, with NHS England highlighting the need to ensure that such mechanisms are proportionate and effective. The decision also reflected concerns that financial penalties were not always applied consistently in practice. More generally, sector feedback indicated that the administrative workload associated with CQUIN was considered disproportionate to its relatively limited financial impact, and the pause was widely supported.

BPTs have remained a stable component of the incentive landscape. Recent reforms mainly reflect a shift towards integrating them into broader NHS payment systems rather than operating as standalone incentives. Under the National Tariff Payment System and the NHS Payment Scheme, BPTs have increasingly been embedded within blended funding models combining fixed and variable elements, alongside a clearer distinction between elective, urgent, and annual tariffs. There has also been a move towards simplifying criteria and relying more on routinely available administrative data. More recently, BPTs have been more closely aligned with national priorities such as reducing waiting times and improving system efficiency, alongside their original focus on clinical quality. Overall, reforms have strengthened their integration into system-wide payment and performance frameworks.

Adjustments to the HVBP programme in the **US** have included changes in domain weighting and eligibility rules, though core design features remain.

Following the IGAS evaluation, the IFAQ programme in **France** was substantially reformed for 2026 to improve transparency and predictability. A key change is the move away from relative ranking: hospitals are no longer grouped, and payment thresholds are no longer based on peer performance. Instead, performance is assessed against absolute thresholds defined in advance by the HAS, which determine fixed performance classes and remuneration rates.



The incentive design was also simplified. Each indicator now has its own dedicated budget envelope, eliminating the previous overall weighted score. The accreditation indicator was strengthened, accounting for nearly half of the total envelope, and high-performing accredited hospitals can receive up to 110% of the standard rate, while non-accredited hospitals may receive conditional payments linked to action plans.

At the same time, the indicator set was revised: orthopaedic indicators were removed, and digital health indicators were updated by the national digital health authority.

Financial incentives are now more sensitive to improvement, particularly for lower-performing hospitals, and allocations depend on both hospital activity and performance class, with additional adjustment coefficients based on prior-year results. Unused funds are redistributed across hospitals, and payment calculations are based on data from the previous year.

4.11 Limitations

A limitation of this comparative analysis concerns the selection of programmes included in the study. The five programmes were not intended to represent the full range of possible P4P design choices available to Belgian policymakers. Rather, they were selected as illustrative and sufficiently documented cases that enable meaningful comparison across key programme features. This selection implies that programmes are geographically concentrated and disproportionately drawn from settings where P4P initiatives have been extensively studied and documented in the literature. Consequently, less well-documented programmes and experiences, including potentially relevant alternative design approaches, are excluded from the analysis. The findings should therefore be interpreted as informing comparative reflection and identifying potential lessons for policy design, rather than providing an exhaustive overview of all possible P4P approaches.

Another limitation concerns the absence of primary consultation with frontline healthcare professionals involved in the selected programmes. Given the comparative and international scope of the analysis, systematically engaging clinicians across different countries and health systems was not feasible. In addition, clinicians represent a heterogeneous group, both within and across healthcare settings, making representative ad hoc consultation difficult to achieve. As a result, clinician perspectives were captured indirectly through programme evaluations and existing literature, which synthesise stakeholder experiences and report on issues such as clinician engagement, perceived behavioural impact, and implementation barriers. While these sources provide an important empirical basis for the transversal analysis, they may not fully reflect the diversity of frontline experiences or contextual nuances across programmes.

5 IMPLEMENTATION OF A P4P PROGRAMME IN BELGIUM: POINT OF VIEW OF EXPERTS AND STAKEHOLDERS

KEY POINTS

- Two expert panels were organised (one Dutch-speaking and one French-speaking) using a **nominal group technique** to assess desirability and feasibility of P4P in Belgian acute hospitals.
- **P4P is considered both desirable and feasible**, but only under specific conditions. Experts generally view it as conceptually attractive because it links funding to quality of care, and they also consider its implementation achievable in practice, **provided that key structural, design, and implementation requirements are adequately addressed** and **appropriate safeguards** are in place.
 - **P4P requires adequate funding and should complement stable structural hospital financing.** It functions as an additional incentive rather than a substitute for core financing, which must remain protected.
 - **Multiannual stability is essential for implementation.** Predictable budgets and indicator cycles are required to enable planning, investment, and sustained improvement.
 - **Provider engagement is a core success factor.** Active ownership by clinicians and hospital staff is necessary for meaningful uptake and behavioural change.
 - **Stakeholder involvement and buy-in is essential.** Co-development with frontline professionals ensures relevance and usability for improvement.
 - **Evidence-based, actionable process indicators in well-defined clinical domains are most effective.** P4P works best in targeted, actionable areas rather than broad hospital-wide measures.
 - **Financial incentives must be clearly aligned with system objectives.** They should directly support defined quality improvement goals and desired outcomes.
 - **Incentive design should prioritise rewarding improvement over punishment.** Mechanisms should avoid punitive effects and focus on supporting performance improvement.
 - **Threshold-based approaches are preferred** over a tournament approach. Hospital rankings and competition over resources are seen as harmful to collaboration.
 - **Reliable, timely, and accessible data systems are a precondition for effective P4P.** This requires high-quality coding, robust validation and audit mechanisms to safeguard against miscoding and gaming, and enabling data-sharing frameworks to support learning and continuous improvement.
 - **Transparent governance and fair allocation methods are required.** Robust, methodologically sound allocation increases credibility and acceptance.
 - **Collaboration and structured learning drive improvement beyond measurement.** Peer learning, feedback, a quality-supportive organisational culture and leadership are essential complements to incentives.
 - **Policy coherence across instruments is necessary.** Alignment between P4P, benchmarking, accreditation, and reporting avoids duplication and conflicting incentives.



- o **Scientific and clinical leadership is essential for legitimacy.** Scientific institutions, professional societies and clinical leaders help ensure evidence-based and widely supported indicator selection.
 - According to the experts, the **development and implementation** of a coherent and effective P4P programme, should:
 - o **Adopt a hybrid governance model** for P4P development. Define priority domains top-down while co-developing indicators bottom-up with clinicians and field experts to ensure both strategic alignment and clinical relevance.
 - o **Target priority areas based on evidence and performance gaps.** Use data on variation in care, benchmarking, and scientific input to identify high-impact pathologies and domains where improvement is most needed.
 - o **Develop actionable, evidence-based, and clinically relevant indicators.** Focus on pathology-specific and feasible measures that are co-created with frontline professionals and directly linked to quality improvement.
 - o **Embed P4P in a structured, multiannual and coordinated framework.** Ensure inter-federal alignment, long-term planning, and a balance between stability and adaptability in indicator selection and programme design.
- Link financial incentives to meaningful and measurable quality improvement.** Align incentives with real care outcomes and ensure that funding effectively supports quality initiatives rather than administrative or volume-driven activities.

5.1 Introduction

In this chapter, we take the existing Belgian P4P programme as a starting point and reflect on how it aligns with insights from the systematic review and the analysis of international P4P models. Building on these findings, we engaged with a panel of Belgian experts to discuss the desirability of moving towards a P4P approach that better incorporates these design principles. At the same time, we explore whether, according to experts, alternative or complementary policy instruments might be more effective in improving hospital quality.

In addition, we assess the feasibility of introducing a more robust P4P framework in Belgium by identifying potential barriers and facilitators, including institutional, technical, and stakeholder-related factors. Based on this approach, our research questions are:

- How do Belgian experts perceive the desirability of further pursuing a P4P programme for acute hospitals?
 - o Given the limited impact observed in the literature and given the identified preconditions, does P4P have a meaningful role within Belgium's hospital quality policy? Are there alternative approaches that might be more effective in improving quality of care?
- How do Belgian experts perceive the feasibility to design a P4P programme for acute hospitals that may have a meaningful impact on quality of care? Which barriers and facilitators influence this process?

5.2 Method

5.2.1 Introduction

Addressing the research questions necessitates a robust set of technical skills, including expertise in Belgian and international hospital payment systems and P4P designs, quality measurement and data systems. To facilitate this, we employed a modified 'nominal group approach', recruiting content

experts to participate. The nominal group received the results from the literature review and the international comparison, delivered through a PowerPoint presentation with voiceover. Their task was to evaluate the desirability and feasibility of a P4P programme for acute hospitals within the Belgian healthcare context.

Box 2 – Definition nominal group technique

‘Nominal group technique (NGT) is a structured variation of a small-group discussion to reach consensus. NGT gathers information by asking individuals to respond to questions posed by a moderator, and then asking stakeholders to prioritize the ideas or suggestions of all group members. The process prevents the domination of the discussion by a single person, encourages all group members to participate, and results in a set of prioritized solutions or recommendations that represent the group’s preferences’.¹⁷⁷

5.2.2 Targeted profiles for the modified nominal groups

We conducted two nominal group sessions: one with Dutch-speaking experts (on 3 April 2026) and another with French-speaking experts (on 7 April 2026). To ensure a diverse range of insights, we outlined necessary profiles and recruited participants from our professional network. We aimed to strike a balance between the following profiles:

- Hospital managers
- Hospital umbrella organisations
- Organisations specialised in measuring quality of care
- Experts in health services research with expertise in hospital payment systems and quality of care
- Experts from the Federal Public Service Public Health and from the National Institute for Health and Disability Insurance (RIZIV – INAMI).

Invitations were sent by email.

5.2.3 Approach used during the nominal group meetings

To prepare for the meeting, the KCE team set up one whiteboard to capture the pros and cons related to the desirability of a P4P programme, and a second whiteboard to record perceived facilitators and barriers related to its feasibility.

The two-hour nominal group meeting was moderated by a KCE expert and was structured as follows:

- Welcome and brief introduction outlining the process of the expert meeting (~15 minutes).
- Individual reflection (~10 minutes): Each respondent was asked to consider the pros and cons of P4P’s potential role in Belgium’s hospital quality policy, taking into account the results presented from the literature review and the international comparison, as well as possible alternative approaches for improving quality of care. Pros were written on green sticky notes, cons on red sticky notes, using one sticky note per idea.
- Plenary session (~30 minutes): Participants took turns presenting one idea at a time, ensuring that previously mentioned ideas were not repeated.
- Individual reflection (~10 minutes): Each respondent was asked to reflect on the barriers and facilitators to design a P4P programme that may have a meaningful impact on quality of care in Belgium. Facilitators were written on green sticky notes, barriers on red sticky notes, using one sticky note per idea.
- Plenary session (~50 minutes): Participants took turns presenting one idea at a time, ensuring that previously mentioned ideas were not repeated.



- Wrap-up (~5 minutes): Each expert received 3 page flags to indicate the most important barriers and/or facilitators.

Four (Dutch-speaking group) or three (French-speaking group) observers recorded notes during the meeting, which was also audio-recorded. This material was used to compile the results section, written by one researcher and reviewed by the KCE team members who attended the expert meetings. The draft was then sent to the participating experts for feedback, allowing them to add insights or correct any misinterpretations.

In addition to the nominal groups, a comprehensive stakeholder meeting was conducted, featuring an in-depth presentation of the international study findings (literature review and selected programmes), the expert panel's report, and recommendations. Stakeholders (see colophon for participants) had the opportunity to engage during a Q&A session and could submit comments to the KCE in the week following the meeting. Comments emerging from the stakeholder meeting that had not already been raised during the expert panels will be explicitly reported in the results section (see section 5.3).

5.3 Results

Disclaimer. The reader should be aware that this chapter is based on the opinions of the consulted experts and stakeholders. Hence, statements are solely based on expert and stakeholder opinions stated during the expert panels and stakeholder meeting and not on verified facts.

5.3.1 *Participants of the expert panels*

In total 18 experts (8 Dutch-speaking and 10 French-speaking experts) participated, representing the following profiles: 2 health service researchers; 9 (ex) hospital managers; 3 representatives of public authorities, 3 representatives of an institute for healthcare quality measurement and 1 representative of a hospital umbrella organisation (see colophon).

5.3.2 *Desirability of further pursuing a P4P programme for acute hospitals in Belgium*

As an introduction to the expert meeting, it was clarified that the objective was not to evaluate the current P4P programme. However, throughout the discussion, participants frequently referred to and commented on the current programme, often using it as a concrete example to illustrate broader views on P4P.

Participants were also encouraged to consider a broader perspective, including elements related to P4P that fall outside the scope of the official programme. One such approach is the newly proposed project, called AIM, to embed antimicrobial stewardship (AMS), infection prevention and control (IPC), and microbiological surveillance, which partly incorporates a P4P element. Another example are existing conventions for chronic kidney patients (see Box 1 in Chapter 2).

5.3.2.1 *Conditional support for P4P in Belgian Hospitals*

In general, experts expressed support for the idea of implementing P4P in Belgian hospitals, recognising its potential to focus attention on quality improvement. However, this support was immediately accompanied by discussion of the necessary conditions for success. Experts emphasised that without careful design and clear preconditions, the introduction of P4P could risk limited effectiveness or unintended consequences.

One expert raised the question of why P4P is still being implemented if the literature does not show clear effects, and stressed that quality of care should be and is a core professional obligation, and should not require financial incentives.

5.3.2.2 *Potential role of P4P programme in Belgian hospitals and conditions for effective implementation*

APPEALING IN THEORY

A P4P approach is often considered intuitively attractive, as it links financial incentives to quality of care. It was emphasised that quality improvement requires resources, and therefore rewarding quality financially is legitimate. Financial incentives can contribute to behavioural change, as seen in other domains such as accreditation of physicians, where clinicians are generally committed to attending continuing education courses to maintain their accreditation. In this sense, P4P could help shift funding away from rewarding non-performance towards incentivising quality. Moreover, attaching financial stakes to quality may increase its visibility within hospitals and stimulate internal discussions and engagement around quality improvement. Finally, financial support can help sustain attention for quality improvement. Without dedicated funding, there is a risk that quality initiatives receive less priority. P4P gives quality a more prominent place in healthcare systems that traditionally reward activity volume rather than care quality. Some participants also considered P4P a useful driver of a broader quality culture and a way to maintain momentum around continuous improvement. Others noted that, if designed appropriately, it may improve patients' quality of life.

In theory, P4P can help standardise practices by encouraging providers to align with agreed standards and evidence-based care pathways.

DIFFICULT TO VALIDATE IN PRACTICE

Evidence on the effectiveness of P4P programmes may be influenced by publication bias. This bias is particularly relevant given the methodological challenges in evaluating P4P, as P4P can be considered a 'complex intervention', and its effects may be distorted by other factors. This also applies to sub-indicators, such as the hospital standardised mortality ratio (HSMR), which reflects the combined effect of a hospital's activities and case mix.

There is positive publication bias, whereby studies showing positive effects are more likely to be published than those reporting null or negative results. As a result, the apparent impact of P4P in the literature could be overestimated, giving stakeholders an overly optimistic view of its effectiveness. Additional challenges include difficulties in isolating its effects from other concurrent quality improvement initiatives, limited or no programme evaluations in some countries, and the potential for unmeasured outcomes. Others argued that the effects of P4P may in fact be underestimated, as some positive effects cannot be measured and are therefore overlooked. Recognising these biases is important when interpreting international evidence and considering the introduction or expansion of P4P schemes in Belgian hospitals.

The distinction between P4P and incentives embedded within the existing nomenclature is not always straightforward. While P4P explicitly links financial rewards to measurable quality indicators, nomenclature-based incentives may also include conditions or financial adjustments that indirectly influence provider behaviour (for example, care protocols). In the opinion of the experts, this overlap can create ambiguity about whether observed changes in practice are driven by P4P mechanisms or by pre-existing financial structures. Clarifying these differences is important for designing, evaluating, and interpreting the impact of P4P programmes, and for ensuring that incentives are aligned with the intended quality improvement goals.

EFFECTIVENESS OF P4P IN WELL-DEFINED, ACTIONABLE CLINICAL AREAS

According to the consulted experts, P4P may be particularly effective in well-defined clinical areas with clear, actionable targets. Examples include antibiotic stewardship, specific care pathways, or time-sensitive interventions. In such contexts, combining measurable indicators with financial incentives may support targeted improvements. However, there was concern that programmes may focus on narrow measurable aspects of care while neglecting broader dimensions of quality.



Hospital-wide mortality, for example, is too broad to serve as a practical target. Experts noted that using overall hospital mortality as a quality indicator is not specific enough to guide actionable improvement efforts, as it aggregates many different patient populations and conditions, making it difficult to identify where changes would have an impact. Moreover, hospital-wide approaches such as mortality do not foster a sense of (co-)ownership among department heads or physicians, as it is unclear whether their own service or specialty contributes positively or negatively to the overall hospital outcome.

PROCESS INDICATORS AS ACTIONABLE MEASURES FOR P4P

Process indicators are regarded as more actionable than outcome indicators, particularly in the context of P4P programmes. They measure specific steps in care delivery, such as adherence to clinical protocols or completion of recommended interventions, which hospitals can directly influence. When combined with benchmarking against peer institutions, process indicators provide clear targets for improvement and foster accountability. Experts highlighted that focusing on process measures allows hospitals to identify concrete areas for intervention, engage clinicians more effectively, and track progress in a timely and practical way, without the confounding factors that can affect outcome indicators.

Outcome-based metrics were considered to carry a higher risk of gaming and often necessitate risk-adjustment. To minimise gaming in outcome measures, it is important to focus on more objective indicators such as (re)admissions, emergency department visits, etc., following specific procedures or in chronic patient populations (e.g. heart failure, respiratory failure). Hence, if outcome indicators are included, they should be defined with strict conditions and safeguards.

HOSPITAL CHOICE IN SELECTING P4P INDICATORS

Allowing hospitals to select which indicators to focus on from a longer list can enhance ownership and engagement in P4P programmes. Local indicator selection is a key feature of the English CQUIN programme (see [Supplement 1](#) to the Scientific Report). Experts noted that giving hospitals this flexibility enables them to prioritise areas where they see the greatest need for improvement or where they have the capacity to act, making targets more relevant and actionable. This approach can also increase clinician buy-in, as staff are more likely to support initiatives they have had a role in choosing. However, some experts cautioned that if each hospital selects only the areas where they already perform well, attention may be diverted away from less quality-minded clinicians or weaker domains, potentially slowing broader improvement efforts. Therefore, hospitals should be able to demonstrate a margin for improvement and provide evidence of progress. Careful oversight is needed to balance flexibility with accountability.

MULTIANNUAL PLANNING AS A KEY TO EFFECTIVE P4P IMPLEMENTATION

Experts emphasised that implementing P4P programmes within a multiannual, structured plan is crucial for success. Multiyear planning allows hospitals to set clear, measurable objectives, track progress over time, make necessary investments, and adjust strategies as needed. It provides continuity and predictability, encourages sustained quality improvement, and ensures that resources and incentives are aligned with long-term goals. Experts also noted that multiannual planning helps make indicators more actionable, supports benchmarking, and fosters a culture of continuous improvement across institutions. They stressed that indicators should remain in the programme for several years to allow sufficient time for improvement efforts to take effect and to ensure meaningful evaluation.

NECESSITY OF ADEQUATE FINANCING FOR P4P

Experts highlighted the importance of financial incentives, as they help focus attention on quality improvement and signal the importance of achieving measurable targets.

Funding that is too low risks having little impact, particularly in financially performant hospitals that already invest in quality. Adequate funding ensures that P4P initiatives can be properly implemented, sustained, and linked to actionable quality improvements, while supporting hospitals in achieving long-term goals.

For some experts, the current P4P budget in Belgium is considered substantial and sufficient to support quality improvement initiatives. Others, however, argued that the financial incentive is too limited to meaningfully influence hospital behaviour, particularly in institutions that are already financially performant and actively engaged in quality improvement. This divergence highlights differing views on the added value and impact of financial incentives in improving quality of care.

Yet funding alone was not considered sufficient. While budgets need to be substantial enough to meaningfully influence hospital behaviour and clinician engagement, several experts emphasised that financial incentives are only one component of effective quality improvement strategies. Intrinsic motivation, such as professional commitment, organisational culture, and the desire to provide high-quality care, was also identified as an important driver by some participants.

SETTING CLEAR TARGETS FOR P4P FINANCING

The objective of financial incentives must be clearly defined. Experts noted that quality improvement efforts can sometimes lead to a reduction in activity. For example, specialised heart failure nurses may help prevent emergency visits or readmissions and thus reduce activity-based financing. In such cases, linking funding to these efforts requires a specific, well-articulated goal to ensure that the incentive aligns with the intended outcome. Clear objectives help hospitals understand what behaviours are being rewarded and ensure that financial support effectively drives the desired improvements in care.

FINANCING AND INCENTIVES

Rankings between hospitals were strongly criticised, with several participants preferring threshold-based approaches over competitive league tables. Competition between hospitals was seen as undermining collaboration on quality improvement, as financial or reputational rivalry makes transparent peer learning more difficult. This concern was particularly raised in relation to the scoring of the HSMR, which is currently based for half of its score on a linear ranking of hospitals. In this context, programmes focused solely on measurement, without accompanying support mechanisms, were considered insufficiently effective.

Views on incentives were mixed. Some participants supported rewarding high performers in order to incentivise good practice, while others argued that rather low performers need to be incentivised and supported to improve and thus not be punished, expressing clear opposition to punitive approaches and malus systems. This raised the broader question of whether penalties should be excluded altogether. Some, however, favoured the idea of no pay for no quality.

Several participants emphasised the need for adequate additional resources rather than budget-neutral or closed-envelope systems, which could effectively function as financial sanctions. Overall, the emphasis was placed on supporting improvement rather than penalising low performance.

More broadly, critics argued that many P4P programmes mainly redistribute existing budgets rather than providing additional funding, thereby creating winners and losers. In such a context, poorly performing hospitals may receive fewer resources, further limiting their ability to improve and potentially creating a vicious cycle. Concerns were also raised about hospital-wide penalties affecting staff who are not directly responsible for performance outcomes. Finally, it was noted that reducing base funding for clinical costs (nursing and care personnel and medical equipment, the B2-part of the hospital budget) while offering relatively small quality bonuses is putting strain on base quality, internally inconsistent and unlikely to lead to meaningful improvements in care quality.



DATA TIMELINESS AND RELIABILITY

Some participants highlighted that healthcare systems now possess large amounts of data that can be used for P4P programmes. P4P can therefore encourage systematic measurement and analysis of quality. Shared indicators may also create a common language for benchmarking and improvement.

However, delays in data availability can undermine the credibility and effectiveness of P4P programmes. Experts noted that rewarding hospitals based on past performance may be perceived as unfair, especially when there is little opportunity to act on current indicators, which may have already changed.

Many concerns were raised regarding indicator validity and data quality. There is also a risk of 'gaming' specific diagnoses to improve measured performance. To address this, robust coding controls and regular audits of patient records are essential. Analyses of care quality should account for coding differences between hospitals to avoid distorted results and ensure that incentives truly reflect meaningful improvements.

UNINTENDED EFFECTS

Beyond gaming, several other unintended effects were highlighted. Participants pointed out that P4P may lead to a narrowing of clinical focus, where attention is concentrated on incentivised indicators while other important aspects of care are relatively neglected. This selective attention can distort clinical priorities. There was also concern that performance measurement can create perverse incentives, such as performing unnecessary procedures or over-treating patients in order to meet thresholds or improve scores. In some cases, this may result in care that is not fully aligned with clinical need.

INCENTIVES FOR CLINICIANS AND THE ROLE OF LOCAL NETWORKS

The current system was described as providing too few direct incentives for individual healthcare providers. In this context, experts emphasised the importance of fostering learning and collaboration through quality-focused clinical forums, such as evolving from multidisciplinary oncology consultation meetings (MOCs) towards broader quality-oriented consultation (QOC) frameworks, in order to strengthen knowledge exchange and mutual improvement. One proposal was to anchor part of the P4P programme at the loco-regional network level, thereby encouraging hospitals and clinicians to engage more actively in joint quality initiatives. However, experts also noted challenges with existing local networks (LOK – GLEM)ⁱ, where only first-line practitioners participate, rather than hospitals, which has limited the effectiveness of these initiatives. Strengthening these networks and embedding P4P incentives at this level was viewed as a potential way to enhance engagement and promote shared responsibility for quality improvement.

At the same time, a broader structural issue was discussed regarding incentives for individual physicians. A question was raised about whether physicians would be willing to forgo fees or the accreditation-related part of their fee in cases of low-value care. It was noted that, in the Belgian system, quality-related mechanisms currently affect hospital-level funding rather than individual physician remuneration, which was seen as problematic because it may indirectly penalise professionals who are not responsible for performance shortcomings. It was therefore argued that quality incentives should be more directly linked to the fee for the individual act. Given that

ⁱ A LOK (*lokale kwaliteitskring*) or GLEM (*groupe local d'évaluation médicale*) is a local quality group of physicians or pharmacist–biologists who share and critically review each other's clinical practice with the aim of improving care quality through peer review. It functions as part of the accreditation requirements for individual healthcare providers, who must be a member of a LOK – GLEM to obtain or maintain their accreditation. Membership involves regular meetings where clinical practice and quality issues are discussed, and participation contributes to continuing professional development.

remuneration of physicians in Belgium is not integrated into the DRG-system but organised through a separate funding stream, this institutional separation was identified as a key constraint.

THE COMPLEMENTARY ROLE OF P4P TO LICENSING

Healthcare licensing based on structural elements, patient volumes and outcomes (for example, a breast clinic with multidisciplinary team achieving specific case numbers and clinical results) raises the question of whether additional P4P incentives are necessary. Experts noted that even in such settings, variation in care quality may persist, suggesting that licensing alone may not fully drive improvement. Effective quality programmes require alignment between programming, licensing and financing. While licensing is typically a one-time award, or should at a minimum be periodically monitored and enforced, financing mechanisms are more flexible and can be used strategically to guide ongoing behaviour and encourage continuous quality improvement.

ALTERNATIVES TO P4P

Experts debated which mechanisms are most effective in driving quality improvement, particularly comparing direct financial incentives with benchmarking and public reporting. While public reporting may encourage improvement, it was noted that it requires appropriate safeguards and opportunities for contextualisation or rebuttal to ensure fairness, which can be difficult to operationalise. Some experts also pointed out that such reporting is often of limited relevance to patients, who may prioritise practical aspects such as waiting times over (small) differences in performance scores.

In the Belgian context, a general lack of transparency was highlighted, with calls for making results publicly available, including hospital names, as is already done by the Flemish Institute for Quality of Care (*Vlaams Instituut voor Kwaliteit van Zorg – VIKZ*) and in some sickness fund reports. At the same time, it was emphasised that this should not apply uniformly to all indicators. Broad consensus on the choice and interpretation of indicators was considered essential, since certain measures, such as mortality rates, may unfairly disadvantage specific hospitals depending on their case mix. Outcome indicators as used by the Flemish Institute for Quality of Care are always case-mix adjusted.

Benchmarking, particularly within a limited peer group, was seen as a powerful driver of improvement, as hospitals are motivated to avoid repeatedly ranking poorly (e.g. being identified as having the highest mortality rates over consecutive periods). Benchmarking may also be unfair when hospitals differ in case mix or when some institutions had prior knowledge of indicators.

A complementary approach would be to place greater emphasis on patient-reported measures such as PROMs and PREMs, although their use was described as still insufficiently developed, despite being included in the current P4P programme. In this context, it was stressed that both clinical-reported outcomes (CROMs) and PROMs should be systematically collected and transparently reported. However, reporting alone was seen as insufficient; the value of these data depends on their integration into broader collaborative learning systems. This includes structured workshops to discuss variations in outcomes, on-the-ground support to interpret results, exchange of best practices, and non-judgemental peer review. Such an approach was considered to require a non-competitive environment between hospitals in order to function effectively.

More generally, collaborative workshops were viewed as more effective than measurement alone (see also the AIM project, Box 1 in Chapter 2). Data collection was not considered sufficient without accompanying support to help translate results into practice change. A collaborative model was therefore seen as essential both within and across hospitals.

As an illustration, it was mentioned that the FPS Public Health is funding a maternal mortality project (covering both mothers and newborns) that includes field-based support and joint analysis of results with frontline professionals. However, it was acknowledged that implementing such approaches requires considerable infrastructure and resources.

One proposal raised by experts was to reduce the P4P budget back to the budget at the start of the P4P programme (€ 6 million) and redirect resources toward internal quality improvement projects



within hospitals. There are already numerous initiatives in place that could benefit from targeted support, allowing hospitals to focus on meaningful, actionable improvements in care. This highlights the delicate balance required in designing incentives.

Several other alternatives or complements to P4P were proposed:

- Accreditation systems: some participants argued that accreditation had already produced more tangible improvements than public reporting or P4P. In this context, it was also noted that compliance with accreditation criteria should be ensured, which is not always straightforward due, for instance, to workforce shortages.
- Financing specific improvement projects (e.g. antibiotic stewardship through infectious disease specialists)
- Clinical pathway-based contracts
- Continuous professional training
- Incident learning systems with transparent reporting.

Finally, others noted that, although the effects of P4P are considered mixed, it may still produce positive spillover effects. More broadly, they emphasised that it is the combination of quality-related measures, including P4P, that can improve overall quality through diffusion and indirect effects beyond the directly targeted areas.

5.3.3 Feasibility to design a P4P programme that has a meaningful impact on quality of care

This section explores the feasibility of designing a P4P programme that can have a meaningful impact on quality of care. The following subsections present the key barriers and facilitators identified by experts, highlighting the conditions that may hinder or support effective implementation. Some overlap exists with the previous section on desirability, as several of the issues discussed inevitably touch upon both the principles and the practical conditions of implementation.

5.3.3.1 Barriers

LIMITED EVIDENCE OF EFFECTIVENESS AND RESULTING LACK OF BUY-IN

A perceived lack of clear and consistent evidence of effectiveness in the literature was identified as an important barrier. Participants noted that the absence of demonstrable results makes it more difficult to justify and support P4P approaches, particularly among clinicians and stakeholders expected to engage with them. This uncertainty also undermines acceptance and adherence, as confidence in the expected benefits of the programme remains limited when supporting evidence is weak or inconclusive.

LACK OF ALIGNMENT BETWEEN OBJECTIVES AND INDICATOR DESIGN

A lack of clearly defined health system objectives prior to indicator development was identified as a structural barrier, as it leads to measurement tools that are not fully aligned with intended outcomes. Without clear goals, indicators risk lacking relevance, failing to reflect clinical priorities, and not being accepted by frontline professionals, thereby limiting their usefulness for driving quality improvement. In addition, the absence of explicit objectives was seen to hinder alignment with other ongoing quality initiatives, potentially leading to fragmentation, duplication of efforts, and reduced overall coherence in quality policy.

LACK OF CLINICAL GUIDELINES

The absence of well-defined and widely implemented clinical guidelines remains a barrier to effective P4P implementation. Without clear standards of care, it becomes difficult to define appropriate indicators and assess performance. Some experts also raised the question of whether financial consequences, such as withholding accreditation incentives in cases of low quality (for example, for physicians at the 97th percentile) could be considered, though this would require a robust and widely accepted framework of guidelines.

COST OF IMPLEMENTATION

A major barrier identified was the cost of implementation, particularly the administrative burden of data entry, registry maintenance, and monitoring systems. The time required to set up registries and ensure ongoing monitoring must be taken into account, as this represents a significant cost. In several cases, the cost of measurement was considered higher than the value of the financial incentives. This burden was further exacerbated by duplicated data collection efforts, highlighting the importance of adhering to a 'register once, use multiple times' principle, whereby data are recorded at the source and reused across different systems and purposes. Insufficient staffing levels and resource constraints were also seen as significant obstacles to effective implementation.

DATA AVAILABILITY, QUALITY, AND TIMELINESS

Data-related challenges were identified as a major barrier. Concerns were raised about the validity, completeness, and timeliness of available data, as not all relevant aspects of care are systematically captured (e.g. timing of interventions in sepsis or specific clinical actions). Poor data quality was also highlighted, including inconsistencies and underreporting that distort reality and undermine credibility (e.g. some hospitals never report hospital-acquired urinary tract infections).

Experts emphasised the need for more recent data to ensure relevance and fairness in P4P programmes. While hospitals often have access to their own data in real time, official datasets may only become available after significant delays (e.g. two to three years), which undermines the usefulness and credibility of performance measurement. Legal barriers further compound these issues, as federal data cannot always be returned to hospitals for validation, limiting feedback loops, learning, and improvement. Strengthening legislation to enable data exchange, validation, and linkage with other sources (e.g. mortality data) was therefore seen as essential to improve accuracy and enable meaningful analysis. Stakeholders emphasised the need for actors responsible for data processing to be provided with a clear, robust and sufficiently comprehensive legal basis for carrying out these activities.

At the same time, improving data infrastructure was considered an important facilitator. This includes investment in national systems for collecting PROMs, PREMs, and CROMs, as well as mechanisms that enable scientific analysis and structured feedback to hospitals. Platforms for data exchange, such as an inter-federal system, were seen as potentially valuable, particularly if focused on a limited set of key performance indicators. Participants stressed the importance of standardised methodologies to avoid fragmentation, where institutions develop divergent approaches.

DATA REGISTRATION AND USE OF RESOURCES

There was debate about whether financial resources should be used to improve data registration. Proper funding for data collection was also considered essential, given the significant workload and associated costs. However, most argued that improving data registration and quality cannot be funded by taking resources from the B2-part of the hospital budget, which is used to finance the hospital core business. The focus should rather be on how existing data can be used to drive quality improvement. Experts also cautioned against designing indicators that reward registration itself rather than actual quality of care, stressing that indicators must always reflect meaningful quality objectives.



CODING PRACTICES AND RISK OF GAMING

Concerns were raised about overcoding, undercoding, miscoding and gaming, particularly when financial incentives are linked to coded data. Evidence suggests that coding practices are already influenced by existing financing mechanisms, which may distort the reliability of data for quality measurement. For example, the coding of complications can affect the severity of illness (SOI) and, in turn, indicators potentially used in P4P schemes. These effects may operate in opposing directions: overcoding of complications may increase the SOI, leading to higher activity-based payments and financial gains, while simultaneously resulting in penalties or reduced payments under P4P complication-related indicators. However, it was noted that the financial gains associated with a higher SOI are likely to outweigh the potential losses linked to P4P incentives, thereby creating a misalignment in incentives.

Hospitals are aware that coding is not always strictly controlled, further increasing the risk of strategic behaviour. Data linkage (e.g. with billing data) and validation processes may partially address these issues, but are not sufficient on their own. Systematic auditing of coding practices and patient records is essential. According to the consulted experts, no effective control mechanisms are currently in place. It was mentioned that while the principal diagnosis is generally coded accurately, secondary diagnoses are often less reliably recorded. Experts also noted that strong financial incentives aimed at counteracting overcoding could unintentionally shift gaming behaviour toward quality indicators themselves.

COMPLEXITY OF INDICATORS

At a minimum, results should be shared between hospitals to enable peer discussion and comparison, and the calculation methods underlying these results should be fully transparent, as there is currently considerable debate and uncertainty regarding the methodologies used. In addition, overly complex indicator definitions were seen as a barrier to effective implementation, as indicators that are not easily understandable or actionable are less likely to influence clinical practice. Simplicity and clarity were therefore highlighted as essential design principles to ensure usability, facilitate engagement, and support meaningful use of P4P information.

LIMITED FINANCIAL MARGINS AND NEED FOR PREDICTABILITY

Cost containment policies and limited financial margins within hospitals were identified as a constraint. For P4P to be effective, funding mechanisms should provide predictability, combining stable operational financing with additional incentives for achieving quality targets. Without sufficient and predictable financial margins, hospitals may struggle to invest in quality improvement initiatives.

Ensuring adequate staffing levels around patients was considered essential, as insufficient resources and cost-containment measures were identified as significant barriers to improving quality of care.

ACTIVITY-BASED PAYMENT AS A BARRIER TO QUALITY

Activity-based payment (fee-for-service) was identified as a barrier to quality of care, as it may incentivise volume over appropriateness of care. Participants noted that such a system can encourage unnecessary tests and procedures, as well as defensive medicine driven by risk avoidance or patient expectations rather than clinical need. This focus on activity rather than value may distort clinical decision-making and divert attention away from delivering the most appropriate care for patients.

FRAGMENTED HEALTHCARE LANDSCAPE

The increasingly diffuse healthcare landscape raises questions about the scope of P4P programmes. As care is increasingly delivered across different settings (for example, through integrated care models), it becomes less clear which providers should be targeted by financial incentives. For example, in chronic conditions like diabetes, responsibilities may be shared across multiple actors, complicating the design of hospital-based P4P programmes.

DISTRIBUTION OF COMPETENCIES

A major structural barrier lies in the fragmented division of responsibilities across different levels of governance (federal and regional). This institutional separation can hinder the alignment of policies, financing mechanisms, and quality initiatives, thereby complicating the design and implementation of coherent and effective P4P programmes. Misalignment of objectives between entities further reinforces this fragmentation and limits the feasibility of a unified approach. In contrast, a coherent and integrated governance structure was considered essential for success, with a unified quality strategy across federal and regional level, supported by aligned training, funding mechanisms, and a clear policy framework, seen as necessary to ensure consistency and effective implementation.

5.3.3.2 *Facilitators*

P4P AS A COMPLEMENT TO ADEQUATE BASE FINANCING

A key facilitator identified is the presence of sufficient structural financing, with P4P functioning as an additional incentive rather than a substitute for core funding (B2). Experts stressed that base financing should not be reduced to fund P4P, as this could undermine essential hospital operations. Instead, P4P should operate as a bonus system that rewards good performance, while carefully considering whether and how to apply penalties. Avoiding linear budget cuts and ensuring that incentives genuinely stimulate quality improvement are essential design principles.

The overall cost of the system should, however, not outweigh its benefits.

BUDGET PREDICTABILITY AND MULTIANNUAL PLANNING

Participants consistently highlighted the instability of the current system, noting that frequent changes in indicators and rules make it difficult for hospitals to adapt and invest, undermine trust in the programme, and hinder the standardisation of practices. In contrast, predictability and long-term stability were seen as key enablers of an effective P4P approach. Establishing multiannual cycles, such as five-year plans, with clear rules defined well in advance and ideally preceded by pilot phases, was considered essential to support strategic priority setting, gradual implementation, alignment of indicator development, and sustained investment in quality improvement. Within this framework, it was suggested that indicators could be maintained for a defined period of around four to five years, after which they could be discontinued once predefined thresholds are reached and replaced by new priorities. It was emphasised that the rules governing this transition should be established *ex ante*, including how success is assessed and when an indicator can be withdrawn. It was also noted that discontinuing indicators after such a period would require safeguards to ensure that achieved quality levels are maintained beyond the programme itself.

ROBUST AND TRANSPARENT ALLOCATION METHODOLOGY

A scientifically sound and transparent methodology for allocating P4P funds is essential. Experts pointed to the need for appropriate statistical approaches (e.g. identifying significant deviations from expected performance) to ensure fairness and credibility. Concerns were raised that current allocation mechanisms may not adequately reflect true performance differences, underscoring the importance of methodological rigour.

ENGAGEMENT AND OWNERSHIP OF HEALTHCARE PROVIDERS

Active involvement of healthcare providers is a crucial facilitator. Ensuring ownership among clinicians and hospital staff increases the likelihood that P4P initiatives will lead to meaningful changes in practice. Engaging providers in the design and implementation of the programme helps build trust and supports sustained quality improvement efforts.



ACTIONABLE AND EVIDENCE-BASED INDICATORS

Indicators used in P4P programmes should be actionable, evidence-based, and focused on areas where variation in care is substantial and well understood. Experts emphasised the importance of selecting domains where there is clear evidence on how outcomes can be improved, ideally starting with less complex areas where the drivers of variation are already established to facilitate implementation. Indicators were seen as more useful when they are pathology-specific rather than hospital-wide, as overly broad measures such as overall mortality often fail to provide actionable insights. In this logic, variation itself should not be tied to financial incentives, but rather used to identify priority areas for improvement.

Well-designed outcome indicators were considered a key facilitator when they are risk-adjusted and grounded in clinical evidence. Strong support was also expressed for process indicators with a clear evidence base, as well as for PROMs and CROMs, provided they are meaningful and can guide improvement. At the same time, several barriers were identified, including the use of structural indicators and PREMs, which may be more easily influenced without accurately reflecting the quality of care. Moreover, PREM scores are often already very high, leaving little room for further improvement.

It was also stressed that indicators should be co-developed with frontline professionals and tailored to specific clinical pathways.

COLLABORATION, LEARNING AND SUPPORT

Collaboration was consistently identified as a key facilitator of quality improvement. This encompasses team-based care, coordination across providers, and even hospital mergers that can support the harmonisation of clinical practices. Structured support mechanisms, such as field-based programmes, peer learning initiatives, and ongoing exchange of best practices, were seen as particularly important. Leadership and organisational culture were also highlighted as central, albeit difficult-to-measure, determinants of quality. Overall, effective programmes were viewed as those that go beyond measurement alone and include strong, hands-on support in clinical practice.

ALIGNMENT ACROSS QUALITY POLICY INSTRUMENTS

Consistency between different policy instruments, such as P4P, public reporting, and benchmarking, is essential. Collecting the same information through multiple channels or creating conflicting financial incentives can reduce effectiveness and create confusion. A coherent approach that aligns data collection, reporting, and incentives across policy tools is therefore a key facilitator.

INVOLVEMENT OF CLINICAL LEADERSHIP

Involving clinical leadership, particularly chief medical officers or heads of a medical department, in identifying priority areas for improvement is important. These actors are well positioned to pinpoint key quality challenges within hospitals and to support the implementation of targeted interventions.

ROLE OF SCIENTIFIC INSTITUTIONS AND SECTOR INVOLVEMENT

Scientific societies and professional organisations play a crucial role in defining priority domains and developing indicators. One expert referred to the Federation of Medical Specialists in the Netherlands as an example of good practice (see <https://richtlijnendatabase.nl/>). Their involvement helps ensure that P4P programmes are evidence-based and broadly supported, reducing the risk of decisions being driven by lobbying or short-term political considerations. At the same time, engaging the broader healthcare sector in identifying a long list of potential improvement areas can support a more inclusive and balanced selection process.

5.3.3.3 *Prioritised barriers and facilitators for P4P implementation*

All of the above barriers and facilitators were captured on a whiteboard, and each expert was then asked to flag the three factors they believed were most critical overall. The following sections summarise the factors that received the highest priority from the expert panels. Because facilitators and barriers are often opposites of one another, the results are analysed and presented together. We present the results separately for Dutch-speaking and French-speaking experts, because the voting patterns were not fully consistent between the two groups.

The Dutch-speaking experts identified the lack of involvement from healthcare providers the greatest barrier and integration with complementary incentives, such as public reporting, as one of the key facilitators for a successful P4P programme (both four votes). The following factors each received three votes: the availability and quality of relevant data, and financing of P4P not at the expense of a reduction in core funding (B2 part). Two votes were given to: P4P based on a core set of evidence-based, actionable and measurable indicators and aligned with other systems, forming an overarching framework rooted in health(care) policy. Finally, one vote was given to the problem of upcoding, the need to understand the drivers of variation between hospitals, and the need to work in a more targeted way through clearly defined projects with one-to-one funding.

For the French-speaking experts, the main facilitator was the predictability and stability of indicators, for example, maintaining them over a four-year period before they are discontinued or replaced (six votes). Four votes were given to the need to define the intended goal to be achieved with the introduction of P4P, followed by a focus on the culture and leadership of a team within the hospital (three votes). The following topics received two votes: the necessity of public reporting and of evidence-based threshold values for process indicators. Finally, one vote was given to the following factors: budgetary constraints, financing of P4P not at the expense of a reduction in core funding (B2), integration of PROMs and CROMs, investing in public administration, the need to have a global approach (federal, regional, etc.).

5.3.4 *Development of a P4P programme in Belgium*

A key consideration in the development of a P4P programme is determining how to initiate the process. Although this was not explicitly addressed as a separate question during the expert meetings, participants considered it an important topic. It includes identifying priority domains for intervention, selecting appropriate and actionable indicators, and establishing a structured approach for their development.

Many participants supported a hybrid governance model: priority domains could be identified top-down by authorities, while indicators should be developed bottom-up with clinicians and field experts. Indicators should be evidence-based, e.g. based on guidelines and evidence-based practice with clear link to outcomes, and aligned with real care problems.

GOVERNANCE IN THE SELECTION OF PRIORITY PATHOLOGIES

Determining which pathologies to target is a critical step in developing a P4P programme. It was suggested that action should start where the main problems are identified, such as antibiotic use, and that the Inter-federal Platform should play a central role in defining and coordinating priority areas for quality improvement (see Box 3). However, the platform was described as an initiative with limited involvement from field practitioners, except through occasional expert invitations. It should be mentioned that the platform was only recently established and its primary objective is to align policies across the different administrations.

The importance of involving frontline professionals and hospitals, particularly chief medical officers, was emphasised, with the platform seen as a potential foundational structure to which field actors could subsequently be invited to contribute. It was further suggested that priority domains be identified top-down, while indicators should be defined bottom-up in collaboration with field experts, as they are best placed to develop measures grounded in clinical practice.



Experts noted that current consensus-based decision-making processes at the federal level can dilute suitable indicators, and instead recommended that the selection process be guided by data on variation in care and performance gaps, drawing on national and international benchmarking sources. Scientific institutions (such as KCE, Sciensano or universities) were identified as key actors in this process, providing evidence-based guidance, alongside existing platforms (such as VIKZ or PAQS, the Platform for the Continuous Improvement of Care Quality and Patient Safety) and sector input to ensure relevance while avoiding interest-driven choices.

Current governance was described as unclear, fragmented, and lacking coherence across Belgian regions. Participants criticised the limited involvement of frontline actors in defining quality programmes and questioned whether institutional responsibilities for healthcare quality are sufficiently clear.

Box 3 – Inter-federal Platform for Quality of Care in Hospitals

The **Inter-federal Platform** was created to provide a common forum for developing a shared approach to hospital care quality in Belgium. It brings together federal and federated administrations, as well as advisory bodies, to develop a joint policy framework for monitoring and improving quality of care, while respecting the autonomy of each entity. Its work focuses on common methodologies, reliable data quality, and structured cooperation across levels of governance. Initially covering general and university hospitals, the scope may later expand to other healthcare actors. Its governance combines political guidance, administrative coordination, and expert input from quality institutes such as VIKZ and PAQS.

FOCUS ON ACTIONABLE AND RELEVANT INDICATORS

A central element in the development of indicators is the active involvement of clinicians, which helps ensure that indicators are feasible, relevant, and supported in practice. Experts emphasised that field input is essential but must be balanced with structured guidance to avoid overly complex, vague, or unmeasurable proposals.

Indicators should be pathology-specific where possible, such as in oncology (e.g. time to diagnosis), while recognising that not all hospitals treat all conditions. At the same time, certain hospital-wide domains (e.g. infection control) may also be relevant. Experts stressed the importance of selecting areas where hospitals can realistically influence outcomes, avoiding domains that are difficult to act upon by hospitals (e.g. suicide prevention). A combination of top-down priority setting and bottom-up input through co-creation was seen as the most effective approach.

USE OF INDICATOR SETS AND CHOICE MECHANISMS

Different approaches to indicator selection were discussed. One option is to provide hospitals with a limited menu of indicators, from which they can select a subset to focus on. Indicators should be sufficiently broad to avoid overly narrow optimisation (e.g. focusing only on a single time metric), yet remain clear and transparent. While some experts favoured composite indicators based on care pathways, others stressed the importance of simplicity and transparency, particularly if indicators are also used for public reporting. Regardless of the approach, indicators must be actionable at the hospital level.

STRUCTURED INDICATOR DEVELOPMENT AND INTERNAL GOVERNANCE

Experiences from hospitals show that indicator development benefits from central coordination at hospital level, in collaboration with medical councils. A more structured approach, combined with internal debate, can lead to more robust and actionable indicators, often focusing on process measures (e.g. documentation practices or discharge communication). Importantly, performance must be clearly defined to ensure that financial incentives effectively support quality improvement rather than administrative optimisation.

BALANCING DYNAMISM AND PREDICTABILITY

The development of a P4P programme requires a dynamic process that allows for continuous adaptation, while also ensuring sufficient predictability for hospitals. Experts highlighted the importance of striking a balance between flexibility in updating indicators and stability in programme design, enabling institutions to plan and implement improvements over time.

MULTIANNUAL AND INTER-FEDERAL PLANNING

A multiannual planning framework, such as a five-year cycle, was considered essential for setting priorities and guiding implementation. This planning should be informed by areas where Belgium performs relatively poorly in international comparisons. Experts suggested a rolling system, whereby new multiannual plans are launched each year, ensuring continuity while allowing for gradual expansion. Inter-federal coordination is crucial to align priorities in broader quality policy across levels of governance and to provide the necessary resources for implementation.

LINKING INCENTIVES TO MEANINGFUL QUALITY IMPROVEMENT

Finally, experts noted that current financing systems may still reward non-quality, for example through activity-based funding mechanisms. A well-designed P4P programme should aim to correct such incentives by linking funding to meaningful quality improvement efforts. This includes ensuring that financial rewards are effectively used for quality initiatives, rather than absorbed into general hospital budgets.

5.4 Limitations

A limitation of the expert consultation concerns the composition of the participant group and the perspectives represented. Participants were intentionally selected as system-level stakeholders with a broad overview of the Belgian payment and quality landscape, and the consultation did not aim to include a clinically representative sample of frontline healthcare professionals. As a result, the findings primarily reflect a system-design and governance perspective on the desirability, priorities, and future development of P4P, rather than the experiences and views of frontline clinical practice. Although some participating experts had prior clinical experience, the perspectives captured may differ from those that would emerge from a broader or more clinically diverse group.

This limitation is particularly relevant given the importance of clinician engagement identified in the transversal analysis of international P4P programmes, where implementation experiences frequently highlighted clinician buy-in as a key determinant of effectiveness. At the same time, the focus on system-level stakeholders reflects the governance context in which decisions on the introduction, design, and prioritisation of P4P schemes are typically made, whereas clinicians are more directly involved in the operationalisation and refinement of indicators in the context of clinical practice. Finally, assembling a representative group of clinicians would itself be challenging given the heterogeneity of clinical roles, specialties, and professional contexts across healthcare settings.



6 DISCUSSION

Healthcare systems, and quality improvement initiatives in particular, ultimately aim to improve patient outcomes. P4P is one such initiative, linking financial incentives to the achievement or improvement of measured quality. However, the desirability of using extrinsic motivation as well as the effectiveness and cost-effectiveness of P4P remain contested.²⁶

A P4P programme was introduced for acute hospitals in Belgium in 2018. Although it represents only a small share of hospital financing (approximately 0.4% of the hospital budget and 0.15% of total hospital revenue), the substantial budget increase from € 7.2 million in 2023 to € 40 million from 2024 onwards has renewed stakeholder interest and prompted calls for programme revision.

Against this background, a clearer understanding of if, when and how P4P can contribute to quality improvement is needed. This report provides a comprehensive and up-to-date assessment of whether P4P programmes improve the quality of care in acute hospitals and under which conditions they are most likely to be effective. It examines not only their effects but also the design features and contextual factors shaping outcomes, and contextualises these findings for the Belgian setting.

6.1 Is P4P an effective and appropriate strategy for improving hospital care quality?

6.1.1 *P4P is not a universally applicable quality improvement strategy*

NO CONSISTENT STRATEGY TO IMPROVE PATIENT OUTCOMES

In line with previous systematic reviews, the findings of this report do not suggest that P4P is a consistent lever for improving the quality of care in acute hospitals.^{21, 26, 29, 31-33, 42} The inclusion of more recent evidence does not alter the overall picture, which remains characterised by very mixed, inconclusive, or mostly modest effects. While positive results are reported in some contexts, these are often based on single studies and may be susceptible to publication bias. In contrast, programmes with larger evidence bases tend to show more mixed results, with short-term improvements in patient outcomes that are not consistently sustained over time or under more rigorous evaluation designs.

The programmes examined in greater detail in this report largely reflect these findings. Partial positive effects are observed for the Queensland QIP programme, particularly in stroke care, whereas limited or no impact on patient outcomes is found for HVBP (US), IFAQ (France), and CQUIN (England), and formal evaluations of the latter two highlighted important shortcomings. A notable exception to this pattern is the English Best Practice Tariffs (BPT) programme, whose effectiveness in improving the quality of hip fracture care was clearly demonstrated.

LIMITED LEARNING EFFECTS OVER TIME

Importantly, the lack of consistent impact cannot be explained simply by insufficient refinement of programme design. Even programmes that have evolved or were redesigned over time (such as CQUIN) or that were explicitly informed by theory or prior experience (such as the Medicare P4P programmes and Advancing Quality), do not systematically demonstrate stronger or more durable effects. This points to more fundamental challenges in aligning financial incentives with meaningful improvements in patient outcomes.

WITH A RISK OF UNINTENDED CONSEQUENCES, BUT ALSO POSITIVE SPILLOVERS POSSIBLE

Beyond effectiveness, the findings highlight a range of unintended and dynamic effects. Targeted incentives may generate spillovers in both directions: some studies report positive effects through shared learning and improvements across care pathways, while others document effort diversion away from non-incentivised areas. Mixed results often reflect this trade-off between gains in targeted domains and losses elsewhere. Similarly, behaviours commonly described as “gaming” emerged in

some cases as an alternative explanation of quality improvements, for example to explain improvements in trauma care in CQUIN. Gaming may arise either opportunistically, particularly where monitoring is weak, or as a rational response when genuine improvement appears unachievable. For example, concerns about over-scoring have led France to introduce specific control mechanisms prior to linking indicators to financial incentives.

NO EVIDENCE ON VALUE FOR MONEY

Finally, evidence on equity and cost-effectiveness remains a critical gap. Available findings suggest that P4P may generate both favourable and unfavourable equity effects, sometimes within the same programme. For example, both positive and negative impacts have been reported in CQUIN, while HVBP has been criticised for disproportionately penalising hospitals serving more disadvantaged populations. Although cost-effectiveness was not specifically searched, cost and cost-effectiveness evidence was sparse in the studies included in the systematic review, with mixed findings and no full cost-effectiveness analysis identified. Similarly, the programmes in the international comparison provide little basis for assessing whether P4P represents value for money. This is a key limitation for policymaking, as decisions to revise, expand or discontinue programmes require evidence not only on effectiveness, but also on whether the resources invested are justified by the benefits achieved.

Taken together, these findings reinforce the growing consensus that P4P cannot be considered a generalisable strategy for improving hospital care quality with a clearly reproducible pathway to success. Compared to the previous KCE report 118 on pay-for-quality programmes, which drew partly on more favourable evidence from primary care, this hospital-focused analysis leads to a more cautious interpretation of P4P effectiveness.^{42, 43}

6.1.2 *But might work under narrow conditions*

In the introduction, P4P was described as a complex combination of design choices that jointly influence its ability to improve quality of care. While the importance of design features and contextual factors is well established in the literature, the realist-informed approach adopted in this report allows these elements to be integrated into a coherent programme theory. Grounded in psychological theory, this framework explains how a set of design and context features interact to trigger mechanisms that may lead to improved outcomes, while also outlining feedback loops that might explain this particularly short-termism of any beneficial effect, and potential spillover effects.

Our findings suggest that hospital P4P is most likely to be effective under a narrow set of circumstances, shaped by providers' willingness and ability to change, the opportunity for improvement, and the presence of complementary policies. This combination of factors may trigger sustained institutional change rather than superficial compliance.

ENGAGEMENT AND CREDIBILITY ARE PREREQUISITES FOR CHANGE

A first key element is **willingness to change**. This requires that targeted outcomes can be meaningfully influenced by clinical and managerial activity within the hospital, that well-established and evidence-based practices are available, and that the causal link between these practices and patient outcomes is clear and demonstrable. Under such conditions, well-designed incentives targeting controllable processes and developed with strong clinical input are more likely to support genuine behavioural change.

Stakeholder buy-in is not merely a procedural step but helps avoid an accountability gap,²³ where quality initiatives are perceived as externally imposed and disconnected from clinical practice, as observed in more top-down approaches such as the IFAQ programme. As quality is ultimately delivered by healthcare professionals, engaging clinicians and frontline staff is essential to foster ownership, support professional autonomy, and align intrinsic and extrinsic motivation. This is particularly important because P4P incentives are generally directed at hospitals and can only lead to meaningful quality improvement if they are translated into sustained changes in clinical practice and



care organisation. This creates a structural tension that current P4P designs have not consistently overcome.

International experience illustrates this challenge. Strong stakeholder involvement was identified as a key driver of success in Queensland, particularly for stroke care, while limited clinical engagement in CQUIN – despite initial intentions to promote local ownership – was highlighted as a barrier in early evaluations. Similarly, IFAQ evaluations pointed to limited mobilisation of clinical teams and low awareness of indicators.

Approaches to stakeholder involvement vary, but are often organised in a multidisciplinary way, such as through state-wide clinical networks in Queensland or consensus-based indicator development through the National Quality Forum in HVBP. In some cases, involvement extends across the full programme lifecycle, from development to revision and phase-out, as in the English BPT. Across programmes, barriers to effective involvement include limited time for consultation, particularly in the context of frequent programme or indicator changes, lack of clear mechanisms to involve clinicians and disengagement driven by complex and non-transparent indicators and incentives.

Beyond stakeholder involvement, the use of evidence-informed indicators is equally critical. For example, programmes such as BPT and CQUIN draw on established clinical (NICE) guidelines, and in France, literature review forms a core component of indicator development for the majority of indicators.

INCENTIVES MUST BE MEANINGFUL, SIMPLE AND BUILT ON CLEAR THRESHOLDS

Well-designed incentives are typically characterised by being sufficiently large to motivate behavioural change, simple and focused rather than diffuse, and associated with transparent financial implications. Attempting to incentivise too many objectives simultaneously risks diluting impact. When incentives are explicit and predictable, they can support internal decision-making by justifying investments in quality improvement, but also prompt organisations to weigh potential revenues against implementation costs, including potential losses when quality improvements reduce avoidable activity in activity-based payment systems. This may provide a rationale for inaction when incentives are too small or poorly aligned with the underlying financing model.

Clear guidance on sufficient or optimal incentive size is largely lacking in the literature, with recommendations often reduced to the 'Goldilocks principle': incentives should be neither too small nor too large.¹⁷⁸ Economic work on optimal incentives provides some direction, suggesting that incentive size should not be set as a simple percentage of hospital revenue, but should be linked to the marginal costs of improvement, the expected health gains, the weight providers give to patients' benefits relative to profits, and the opportunity cost of public funds.^{178, 179} In practice, this means that incentive size is context-specific and should take account of the expected benefits and the investments required from hospitals to achieve change, which in turn depend on the scope of the programme, the targeted processes and outcomes, and baseline performance. This is important for interpreting the international evidence base on hospital P4P, which is largely derived from programmes with small or modest incentives. Limited effects may therefore partly reflect insufficient financial leverage rather than the absence of any potential effect of financial incentives as such.

This difficulty in setting incentives is reflected in practice: concerns that the administrative burden of CQUIN was disproportionate to its limited financial value contributed to its pause, while IFAQ and HVBP incentives have been judged insufficient to motivate substantial investment or change. By contrast, although BPT for hip fracture was also small in financial size (around 1%), its simple design and direct link to the care provided meant that it was perceived as a stable and predictable source of income, supporting longer-term investments, for example, enabling the recruitment of orthogeriatricians to deliver best practice care.^{180, 181} A related tension was highlighted by Belgian experts when discussing potential future applications of P4P. For example, nurse-led management of heart failure patients may improve outcomes and reduce emergency contacts and hospitalisations,

but could also reduce hospital revenue, thereby weakening the financial incentive to invest in such approaches.¹⁸²

The quantitative analysis identified threshold-based payment as the only individual design feature associated with a higher likelihood of positive outcomes, with several programmes using a range with lower and upper thresholds. Clear, predefined thresholds give providers greater certainty about the payment consequences of their own performance. By contrast, this design can make expenditure levels less predictable for policymakers.

No clear conclusions could be drawn regarding other financial design elements, such as the relative effectiveness of bonuses versus penalties, or whether incentives should reward achievement, improvement, or both. Belgian experts generally preferred bonus-based schemes funded through additional resources, reflecting a broader sector preference for financial rewards over penalties.²¹ However, from a theoretical perspective, penalty-based approaches may exert stronger behavioural effects through loss aversion and may be more appropriate in cases of sustained sub-normal quality. Such approaches may, however, require mandatory participation. Regarding the choice between incentivising improvement or achievement, programmes vary: some, such as HVBP and IFAQ, combine both, while others focus on a single dimension. For example, CQUIN evolved over time from rewarding improvement to focusing more on performance.

TARGETS MUST BE ACHIEVABLE AND WITHIN PROVIDER CONTROL

A second key element is the **ability to change**, which depends on expectations of success and perception of targets as realistically achievable. Even where willingness is present, incentives are unlikely to translate into improved outcomes if providers lack sufficient time, funding, workforce, or operational flexibility to implement change. In addition, perceived or real external constraints – such as targets outside the control of individual hospitals, because not directly or unclearly related to their activity or influenced by morbidity, demographic or socioeconomic factors outside the control of the provider – may not only undermine the perceived feasibility of improvement, but also encourage gaming responses. This may argue for incentivising actionable processes rather than outcomes, even when outcomes are risk-adjusted.

In response to such concerns, the 2019 reform of CQUIN shifted towards more actionable process indicators. BPTs also predominantly incentivise clearly specified evidence-based care processes, such as timely surgery in hip fracture care, rather than outcomes directly. Evidence from the IFAQ and HVBP programmes suggests that high-performing hospitals may maintain a persistent advantage, while lower-performing providers may disengage when performance gaps are perceived as unbridgeable. Some programmes have sought to address this by adapting targets to starting positions. For example, the Queensland stroke QIP introduced differentiated target tracks, allowing hospitals with newly established stroke units to improve more gradually, while maintaining higher targets and faster improvement trajectories for more advanced providers.

TARGETS MUST FOCUS ON AREAS WITH ROOM FOR IMPROVEMENT

A third key element is the **opportunity for improvement**, meaning that incentive targets are set in areas with a low baseline and clear room for improvement. In Queensland, the success of the stroke QIP was notably attributed to this combination of low baseline access to stroke units and strong evidence for their use, which allowed sustained gains in outcomes.

In contrast, when baseline performance is already high, incentives are subject to ceiling effects, and further gains, both in process adoption and patient outcomes, are inherently limited, resulting in small or negligible impact. This highlights the need to regularly adapt targets over time and to phase out or retire indicators in a timely and managed way, as incentives lose effectiveness once practices become routine. This approach is reflected, for example, in the HVBP programme, where indicators are removed once they are 'topped out', reflecting uniformly high performance across hospitals. Similarly, in England, CQUIN indicators with consistently high performance may be retired or incorporated into mandatory contract standards.



P4P MUST BE EMBEDDED IN A BROADER QUALITY IMPROVEMENT STRATEGY

A fourth and final element is the **presence of complementary policies**. Financial incentives alone are rarely sufficient to drive quality improvement. Across programmes, incentives were most effective when embedded within broader quality improvement infrastructures. Benchmarking (e.g. CQUIN) and public reporting (e.g. HVBP, IFAQ) are commonly used complementary approaches. These elements reinforce the incentive signal; for example, benchmarking and public reporting may introduce reputational incentives alongside financial ones, further encouraging performance improvement. To reduce the administrative burden of data collection, preference should be given to the use of existing administrative data, following a 'register once, use multiple times' principle. In addition, data collected for P4P should be used as broadly as possible, including for audit, public reporting, and benchmarking across providers, as well as for more granular analysis within organisations. Conversely, P4P can also reinforce other quality initiatives by strengthening quality measurement and reporting by requiring systematic data collection, improving data completeness and coding quality, and enabling indicator use beyond the payment scheme itself. These secondary benefits are relevant, but should be weighed against the costs and administrative burden of achieving them through financial incentives.

A clear policy framework setting out national strategic priorities for quality improvement can support the selection of appropriate policy levers and ensure that different initiatives operate in a complementary way. Such medium-term strategic planning is a feature of the English NHS and programmes such as BPT and CQUIN.

6.1.3 *P4P to be used with caution and only selectively*

A critical finding is how few existing programmes simultaneously meet all of these conditions. Not all care domains are equally suited, although stroke unit care may represent a potential exception as it combines a clear bundle of evidence-based practices, strong links to outcomes, and substantial room for improvement. Many outcomes that policymakers may wish to ultimately improve, such as patient experience, long-term chronic disease management, or health equity, either lack a clear evidence-based and actionable process or arise from complex interactions across the broader care system and social determinants of health, rather than from discrete, controllable actions within a single institution.

Targeting such outcomes through P4P therefore falls outside these favourable conditions and risks unintended consequences, including gaming, selective effort towards measurable proxies, and neglect of unmeasured but important aspects of care. The fact that even well-targeted incentives showed diminishing effects over time, and that spillover effects could be simultaneously positive and negative, further underscores the limits of P4P as a standalone lever for quality improvement.

Importantly, even adherence to many of these design principles does not guarantee success. Following the 2019 reform of CQUIN, guidance emphasised simple, evidence-based, and widely supported interventions aligned with national priorities and other national quality initiatives. Despite this, the programme was paused in 2024, without clear evidence of sustained improvements in patient outcomes.

Policymakers should therefore consider P4P not as a universal quality improvement strategy but as one tool suited to very specific, well-defined circumstances. They should invest as much attention in assessing system readiness, including organisational capacity, administrative support, complementary policies, and clinical engagement, as in the design of the incentive itself. Where the conditions for effective P4P are not met, alternative or complementary approaches to quality improvement may be more appropriate, less likely to generate perverse consequences, and potentially represent better use of scarce healthcare resources.

This perspective was also reflected in the views of the consulted Belgian experts and stakeholders. While most supported the principle of P4P, they did not support its current design. P4P was generally seen as a potential tool to counterbalance activity-based financing, which may inadvertently reward poor quality, and to raise the visibility of quality in organisational decision-making. However, the lack of clear evidence on effectiveness was considered a barrier to engagement among healthcare

professionals. Experts broadly endorsed targeted use within the narrow set of conditions identified in this report. They emphasised the need for strong stakeholder involvement, evidence-based and actionable indicators with room for improvement, and alignment with complementary policies such as benchmarking and public reporting (see section 6.3). Sufficient organisational capacity to act, including adequate base funding, organisational culture and leadership supportive of quality improvement, was seen as essential.

6.2 Limitations of the Belgian P4P programme

When the current Belgian P4P programme is assessed against the narrow set of conditions identified in this report, several limitations become apparent. Many of these issues were also raised during the consultations with Belgian experts and stakeholders.

First, the programme is dominated by structure indicators, which account for around 60% of the budget and focus primarily on data registration and accreditation. While robust data collection is a necessary foundation for quality improvement, it is not sufficient to improve patient outcomes. The evidence on accreditation suggests limited or no direct impact on patient outcomes, despite some positive effects on processes of care and professional development.¹⁸³ As such, the current design appears closer to a mechanism for reimbursing data collection and accreditation efforts than to a system that incentivises improvements in quality of care.

Second, since 2023, the programme has moved away from more actionable, clinically grounded process indicators, such as those previously used in oncology registration and antibiotic prophylaxis. This reduces its actionability and clinical relevance. Although the programme initially benefited from stakeholder involvement through a multidisciplinary strategic working group and co-creation of indicators with clinicians and domain experts, this engagement has weakened over time, particularly following the pause of the strategic P4Q working group in 2021.

Third, the main outcome indicator currently used in the Belgian programme, hospital-wide mortality, remains debated as a performance metric in the literature.¹¹ As a broad measure influenced by patient case-mix, demographic, and socioeconomic factors, it may be perceived as outside the control of individual hospitals or not clearly linked to specific processes. This can undermine its perceived fairness and achievability and may encourage gaming rather than foster genuine quality improvement. Moreover, the design and financing of this indicator lack transparency and do not rely on clear performance thresholds.

Fourth, despite a substantial budget increase in 2024, the financial incentives remain relatively small. Moreover, the use of a tournament-based approach, where hospitals compete for a share of a fixed budget, may discourage collaboration and shared learning, which are important drivers of quality improvement.

Fifth, frequent changes in indicators, combined with the use of outdated data and a lack of clarity on the duration and continuity of indicators, provide limited incentives for sustained investment and long-term behavioural change.

Finally, the programme operates largely in isolation from complementary policies. Results for indicators are not publicly reported, and feedback to providers does not enable benchmarking against peers. More broadly, the Belgian P4P programme is not embedded within a coherent national quality strategy aligning different policy levers around shared priorities, which limits its potential impact.

While no formal impact evaluation is currently available, these features indicate that the Belgian P4P programme does not meet the conditions under which P4P is most likely to be effective and is therefore unlikely to deliver meaningful and sustained improvements in quality of care.



6.3 Future role of P4P in Belgium: a selective and structured approach

Chapter 5 contextualises the findings of this report to the Belgian setting, identifying key barriers, facilitators, and preconditions for a revised role of P4P based on input from Belgian experts and stakeholders. Together with the findings from the systematic literature review and the lessons learned from the international comparison, this provides a basis for sketching out a possible future role for P4P in Belgium.

START WITH PRIORITIES, NOT WITH P4P

A first step is to establish a clear strategic vision of what needs improvement. Any future use of P4P should start from clearly identified quality priorities, focusing on areas where improvement is needed and where meaningful gains are still possible. This prioritisation should ideally take place within a broader national quality strategy. In this respect, the newly established Interfederal Platform for Quality of Care in Hospitals (see Box 3 in Chapter 5) has already provided a first step towards such a strategic vision. This vision should now be further developed and refined, informed by existing health and healthcare objectives, scientific evidence on quality gaps and variation, and reported quality problems from the field, including input from clinical leadership, such as chief medical officers and heads of medical departments.

A second step is to determine for which, if any, of these priorities P4P is an appropriate instrument. P4P should be considered as one tool among others, to be used only in specific, well-defined circumstances, as outlined above. In many cases, alternative quality improvement strategies may be more appropriate. P4P is most suited to accelerating the uptake of well-established, evidence-based practices that are not yet consistently delivered, potentially due to (large) upfront costs, and where a clear and demonstrable link exists between incentivised actions and patient outcomes. These conditions are more likely simultaneously met within pathology-specific domains (e.g. stroke), but can in some cases be hospital-wide (e.g. infection control).

GOOD DESIGN IS NOT OPTIONAL

Where these conditions are met, the design of a P4P scheme should follow a number of key principles to ensure relevance, ownership, and a clear and compelling case for sustained improvement. Building on the elements discussed in sections 6.1.2 and 6.1.3 – willingness and ability to change, opportunity for improvement, and presence of complementary policies – the consulted experts emphasised the importance of defining clear objectives in terms of patient outcomes, combined with the co-creation of indicators and targets with relevant experts and, importantly, with frontline healthcare professionals. Incentives should focus on actionable process indicators within the control of providers. Financial design must be transparent, fair, and predictable, preferably based on clear thresholds, and should avoid a competitive tournament approach that undermines collaboration and shared learning. The design process could be coordinated under the auspices of the Interfederal Platform for Quality of Care in Hospitals, as envisaged by consulted experts.

PLAN THE EXIT FROM THE START

P4P schemes should be incentivised over a sustained period to allow sufficient time for implementation, investment, and organisational and behavioural change. A stable but dynamic multiannual framework can provide both direction and predictability. The AMR trajectory within the newly established AIM project (see Box 1 in Chapter 2) follows such a multiannual approach. This may include gradually increasing thresholds, as in the stroke QIP in Queensland. It also facilitates alignment with other quality improvement initiatives such as collaborative learning, benchmarking, or public reporting.

Crucially, exit strategies should be defined from the outset to ensure that improvements are sustained once incentives are withdrawn and to avoid regression in quality. Options include continued data collection for benchmarking or public reporting, integration of successful practices into standard financing (e.g. through adjusted DRG tariffs, as in the Queensland stroke programme after its exit from

the QIP and more generally in BPT), or incorporation into mandatory quality requirements, as seen in CQUIN. Where improvements generate costs for providers but benefits accrue elsewhere, continued funding outside P4P may be needed to sustain best practices.

To maintain momentum while keeping the system manageable, consulted experts pointed to the potential value of a rolling approach, whereby new programmes are introduced over time as others are phased out. This allows priorities to evolve, supports continuous improvement, and avoids overburdening providers with abrupt changes.

SUFFICIENT RESOURCES ARE NEEDED FOR P4P TO WORK

Adequate resources and system capacity are critical. Budget constraints may therefore justify prioritising other quality improvement strategies over P4P. Indeed, P4P requires substantial investment, first and foremost to fund the incentives linked to performance. These need to be sufficiently meaningful relative to implementation costs to trigger investment and real change. Budgets should therefore reflect both the scale of activity concerned and the resources required to improve performance. If incentives are too small, providers are less likely to invest, with poorer-performing and financially constrained hospitals most at risk of opting out, potentially creating a vicious cycle of underperformance.⁵⁹ Conversely, very large incentives are unlikely to be cost-effective.

Additional resources are also needed for administrative support, the development of new data collection where required, and complementary policies. At a minimum, benchmarking should be in place to reinforce incentives through reputational effects and to highlight the relevance of data registration for clinical staff. Public reporting and collaborative learning can further strengthen impact.

Importantly, funding for P4P should not come at the expense of the budget for clinical services, as this risks undermining safe staffing levels, baseline quality of care, and the organisational capacity needed to deliver improvement. This should not be interpreted as a statement on the adequacy of current hospital financing levels, but rather as a recognition that effective P4P requires dedicated resources in addition to those needed for care delivery.

LOCAL FLEXIBILITY IS AN OPTION, BUT WITH CLEAR BOUNDARIES

Some experts advocated for a degree of local discretion to better align nationally defined programmes with hospital needs and quality priorities. This could take the form of a limited menu of schemes or goals from which hospitals can select, a priori, the most relevant subset for their context, as is currently done in the AMR trajectory within the AIM project (see Box 1 in Chapter 2). Local flexibility was also a key feature of the CQUIN programme and can help account for heterogeneous needs, while fostering clinical engagement, ownership, and motivation, and reducing the risk of disengagement associated with strictly top-down targets. However, lessons from this experience highlight the need to clearly define how such flexibility is organised and bounded in practice. Indicators should remain defined at the national level to ensure consistency. Flexibility should not undermine benchmarking, public reporting, or programme evaluation, and data should therefore be collected for all indicators across all hospitals, including those not selected (except where activities are not performed). To avoid excessive administrative burden, the range of choices should therefore remain limited.

ROBUST DATA AND EVIDENCE-BASED STANDARDS ARE ESSENTIAL PRECONDITIONS

Finally, two key preconditions were identified, not so much related to programme design, but to its ability to function effectively. First, a stronger development and use of clinical guidelines and quality standards are needed to identify relevant processes and define meaningful indicators and targets. Second, improvements in data infrastructure are essential, ensuring timely, valid, and linkable data, as well as safeguards against gaming through auditing and regulatory mechanisms. In Belgium, data availability is often delayed by several years and should be made more timely, with most evaluated programmes working with a lag of at most one year. Data collection should capture key aspects of incentivised processes or outcomes (e.g. timing of interventions), with clear definitions of what needs to be registered, support meaningful feedback to hospitals, be linked to patient outcomes to assess



programme impact, and enable the sharing of data for scientific and policy-oriented research. Strengthening the legal framework for data exchange and linkage is therefore essential. Without these foundations, the effectiveness of P4P is likely to remain limited.

6.4 Conclusion

P4P is best understood as a complex combination of design choices that together determine the extent to which it can improve quality of care. Overall, the available evidence points to a limited impact of national or regional hospital P4P programmes on care quality, with effects that are often mixed, unclear, modest, or temporary. In addition, there is limited evidence on cost-effectiveness and a risk of unintended consequences. At the same time, the findings suggest that P4P is more likely to have a positive impact under a narrow set of conditions. P4P should therefore be approached as a selective and complementary policy instrument, applied only where clearly justified, and policymakers should exercise caution when considering its use.

P4P may have a role in supporting targeted improvements in areas with well-established evidence-based practices, measurable and actionable processes with low uptake, and sufficient organisational capacity to act. However, even under favourable conditions, its effects are often modest and time-limited, and depend on sufficient engagement from healthcare professionals and hospitals to translate incentives into sustained improvement. In practice, P4P is rarely effective as a standalone lever and is typically embedded within a broader quality strategy, alongside complementary policies, regulation, and improvements in data and information systems.

In the Belgian context, the current programme does not meet many of these conditions, and its design limits its potential impact. A revised approach would therefore require not only changes to programme design, but also broader investments in system readiness, data infrastructure, and policy coherence. Greater emphasis should be placed on identifying priority areas for quality improvement and selecting the most appropriate mix of policy levers. In some cases, P4P may be appropriate; in others, alternative or complementary strategies are likely to be more effective and represent better use of scarce healthcare resources.

For any revised Belgian P4P programme, policymakers should prioritise rigorous evaluation of both intended and unintended effects, including impacts on targeted processes, patient outcomes, equity, and cost-effectiveness. This is particularly important given the limited evidence on value for money.

Box 4 – Limitations of the study

Limited and uneven evidence base. The evidence on P4P remains limited and geographically concentrated, which constrains the transferability of findings and calls for caution in deriving prescriptive design lessons.

Scope of the literature review. The literature review focused on quantitative studies published in English, potentially excluding qualitative evidence and country-specific insights that could further illuminate mechanisms and contextual factors. This is partly counterbalanced by the international comparison drawing on a wider set of evidence.

Heterogeneity of P4P programmes. Wide variation in design, payment structures, targets, and contexts limits comparability and the ability to pool quantitative results and generalise findings, although this heterogeneity informed the realist-inspired analysis.

Limited number of countries and programmes in the international comparison. The inclusion of a small set of five in-depth programme analyses in the international comparison restricts the generalisability of findings to other healthcare systems. Despite their geographical concentration, the selected programmes capture considerable system diversity.



Contextual and policy interactions. P4P is often implemented alongside other policies (e.g. public reporting or benchmarking, payment reforms), making it difficult to isolate its impact on observed outcomes.

Dynamic programme design. Frequent changes in programme design and indicators over time complicate the assessment of long-term effects.

Expert consultation and healthcare professional perspectives. Insights from the expert consultation are based on a limited number of system-level experts and may not fully capture the diversity of perspectives across stakeholders, including frontline healthcare professionals. In addition, no primary consultation was conducted with frontline healthcare professionals involved in the selected international programmes. Their perspectives were captured indirectly through programme evaluations and existing literature.



■ REFERENCES

1. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*. 2018;391(10125):1023-75.
2. Gerkens S, Lefèvre M, Bouckaert N, Levy M, Maertens de Noordhout C, Obyn C, et al. Performance of the Belgian health system: Report 2024. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2024 02/2024. KCE Reports 376C (376) Available from: <https://doi.org/10.57598/R376C>
3. OECD. Health at a Glance 2025: OECD Indicators. Paris: OECD; 2025.
4. WHO, OECD, The World Bank. Delivering quality health services: a global imperative for universal health coverage. Geneva: World Health Organization, Organisation for Economic Co-operation and Development, and The World Bank; 2018. Available from: <https://www.who.int/publications/i/item/9789241513906>
5. Bates DW, Levine DM, Salmasian H, Syrowatka A, Shahian DM, Lipsitz S, et al. The Safety of Inpatient Health Care. *N Engl J Med*. 2023;388(2):142-53.
6. Panagioti M, Khan K, Keers RN, Abuzour A, Phipps D, Kontopantelis E, et al. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. *BMJ*. 2019;366:l4185.
7. Van Wilder A, Bruyneel L, Cox B, Claessens F, De Ridder D, Janssens S, et al. Call for Action to Target Interhospital Variation in Cardiovascular Mortality, Readmissions, and Length-of-Stay: Results of a National Population Analysis. *Med Care*. 2024;62(7):489-99.
8. Van Wilder A, Bruyneel L, Cox B, Claessens F, De Ridder D, Vanhaecht K. Rates Of Patient Safety Indicators In Belgian Hospitals Were Low But Generally Higher Than In US Hospitals, 2016-18. *Health Aff (Millwood)*. 2024;43(9):1274-83.
9. Van Wilder A, Bruyneel L, Cox B, Claessens F, De Ridder D, Vanhaecht K. Identifying high-impact-opportunity hospitals for improving healthcare quality based on a national population analysis of inter-hospital variation in mortality, readmissions and prolonged length of stay. *BMJ Open*. 2025;15(1):e082489.
10. Van Wilder A, Cox B, De Ridder D, Tambour W, Maertens P, Stijnen P, et al. Unwarranted Between-hospital Variation in Mortality, Readmission, and Length of Stay of Urological Admissions: An Important Trigger for Prioritising Quality Targets. *Eur Urol Focus*. 2022;8(5):1531-40.
11. Bouckaert N, Lefèvre M, Van den Heede K, Detollenaere J, Castanares-Zapatero D, Van de Voorde C. Organisation and financing of intensive care services in Belgium. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2025 04/07/2025. KCE Reports 404 Available from: <https://doi.org/10.57598/R404C>
12. Coeckelberghs E, Vanhaecht K, Seys D, Cox B, Bislenghi G, Wolthuis AM, et al. A Breakthrough Improvement Collaborative Significantly Reduces Hospital Stay After Elective Colectomy for Cancer Across a Healthcare System. *Ann Surg*. 2022;276(5):890-6.
13. van Zelm R, Coeckelberghs E, Sermeus W, De Buck van Overstraeten A, Weimann A, Seys D, et al. Variation in care for surgical patients with colorectal cancer: protocol adherence in 12 European hospitals. *International Journal of Colorectal Disease*. 2017;32(10):1471-8.
14. Lens C, Coeckelberghs E, Seys D, Demeestere J, Weltens C, Vanhaecht K, et al. Variation in stroke care at the hospital level: A cross-sectional multicenter study. *Frontiers in Neurology*. 2022;Volume 13 - 2022.
15. Van Wilder A, Cox B, De Ridder D, Tambour W, Vanden Boer G, Brouwers J, et al. A comprehensive analysis of temporal trends of between-hospital variation in mortality, readmission and length of stay using logistic regression. *Healthcare Analytics*. 2022;2:100123.

16. Wytinck AEJ, Najjar S, Seys D, Van Wilder A, Kesteloot K, Ramaekers D, et al. The Economic Burden of Unsafe Care: A Systematic Review With Narrative Synthesis on the Economic Cost of Unsafe Care in Hospitals Worldwide. *Journal of Patient Safety*. 2026;10.1097/PTS.0000000000001521.
17. Slawomirski L, Auraaen A, Klazinga N. The Economics of Patient Safety: Strengthening a Value-based Approach to Reducing Patient Harm at National Level. Paris: OECD Publishing; 2017. OECD Health Working Papers 96
18. Slawomirski L, Klazinga N. The Economics of Patient Safety: From Analysis to Action. Paris: OECD Publishing; 2022. OECD Health Working Papers 145
19. Gerkens S, Devos C, Van Deynse H, Bouckaert N. Health system performance assessment (HSPA): a first step in the assessment of Environmental sustainability. 2024-13. Brussels: Belgian Health Care Knowledge Centre (KCE); 2026 23/04/2026. KCE Reports 419 (419C) Available from: <https://doi.org/10.57598/R419C>
20. Bates DW, Singh H. Two Decades Since To Err Is Human: An Assessment Of Progress And Emerging Priorities In Patient Safety. *Health Affairs*. 2018;37(11):1736-43.
21. Milstein R, Schreyoegg J. Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries. *Health Policy*. 2016;120(10):1125-40.
22. WHO. Handbook for national quality policy and strategy: a practical approach for developing policy and strategy to improve quality of care. Geneva: World Health Organization; 2018. Available from: <https://www.who.int/publications/i/item/9789241565561>
23. Brouwers J, Cox B, Van Wilder A, Claessens F, Bruyneel L, De Ridder D, et al. The future of hospital quality of care policy: A multi-stakeholder discrete choice experiment in Flanders, Belgium. *Health Policy*. 2021;125(12):1565-73.
24. Busse R, Panteli D, Quentin W. An introduction to healthcare quality: defining and explaining its role in health systems. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. *Improving healthcare quality in Europe. Characteristics, effectiveness and implementation of different strategies*. Copenhagen: European Observatory on Health Systems and Policies; 2019. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK549277/>
25. Busse R, Klazinga N, Panteli D, Quentin W. *Improving healthcare quality in Europe. Characteristics, effectiveness and implementation of different strategies*. Copenhagen: European Observatory on Health Systems and Policies; 2019.
26. Eckhardt H, Smith P, Wilm Q. Pay for Quality: using financial incentives to improve quality of care. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. *Improving healthcare quality in Europe. Characteristics, effectiveness and implementation of different strategies*. Copenhagen: European Observatory on Health Systems and Policies; 2019. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK549278/>
27. Shaughnessy PW, Kurowski B. Quality assurance through reimbursement. *Health Serv Res*. 1982;17(2):157-83.
28. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington (DC): National Academies Press (US); 2001.
29. Eijkenaar F. Pay for performance in health care: an international overview of initiatives. *Med Care Res Rev*. 2012;69(3):251-76.
30. Paris V, Devaux M, Wei L. *Health Systems Institutional Characteristics: A Survey of 29 OECD Countries*. OECD, Directorate for Employment, Labour and Social Affairs, OECD Health Working Papers. 2010;50.
31. Mathes T, Pieper D, Morche J, Polus S, Jaschinski T, Eikermann M. Pay for performance for hospitals. *Cochrane Database of Systematic Reviews*. 2019(7).
32. Eijkenaar F, Emmert M, Scheppach M, Schöffski O. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy*. 2013;110(2-3):115-30.



33. Slawomirski L, Hensher M, Campbell J, deGraaff B. Pay-for-performance and patient safety in acute care: A systematic review. *Health Policy*. 2024;143:105051.
34. Roland M. Pay-for-performance: not a magic bullet. *Ann Intern Med*. 2012;157(12):912-3.
35. Epstein AM. Will pay for performance improve quality of care? The answer is in the details. *N Engl J Med*. 2012;367(19):1852-3.
36. Eijkenaar F. Key issues in the design of pay for performance programs. *Eur J Health Econ*. 2013;14(1):117-31.
37. Frakt AB, Jha AK. Face the Facts: We Need to Change the Way We Do Pay for Performance. *Ann Intern Med*. 2018;168(4):291-2.
38. Cattel D, Eijkenaar F, Schut FT. Value-based provider payment: towards a theoretically preferred design. *Health Econ Policy Law*. 2020;15(1):94-112.
39. Papanicolas I, Milstein R, Ryan AM, Figueroa JF. Are alternative payment models the answer to the failures of pay-for-performance? *BMJ*. 2024;386:e077941.
40. Emmert M, Eijkenaar F, Kemter H, Esslinger AS, Schöffski O. Economic evaluation of pay-for-performance in health care: a systematic review. *Eur J Health Econ*. 2012;13(6):755-67.
41. Maynard A. The powers and pitfalls of payment for performance. *Health Econ*. 2012;21(1):3-12.
42. Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res*. 2010;10:247.
43. Annemans L, Boeckxstaens P, Borgermans L, De Smedt D, Duchesnes C, Heyrman J, et al. Advantages, disadvantages and feasibility of the introduction of 'Pay for Quality' programmes in Belgium. *Health Services Research (HSR)*. Brussels: Belgian Health Care Knowledge Centre (KCE); 2009 16/11/2009. KCE Reports 118C (D/2009/10.273/52) Available from: <https://doi.org/10.57598/R118C>
44. Institute of Medicine. Medicare: A Strategy for Quality Assurance: Volume 1. Washington (DC): National Academies Press (US); 1990.
45. Federale Raad voor Ziekenhuisvoorzieningen. Advies van de FRZV over een P4Q programma 2018. 2018 30/01/2018. FRZV/D/466-2 Available from: <https://overlegorganen.gezondheid.belgie.be/nl/documenten/frzvd466-2-advies-van-de-frzv-over-een-p4qprogramma-2018>
46. FPS Public Health. Key data in healthcare. General hospitals. Brussels: 2025. Available from: <https://www.gezondbelgie.be/images/2024%20FOD%20Volksgezondheid%20-%20Blikvanger%20-%20ENG%20interactief.pdf>
47. Brouwers J. Exploring the future of hospital quality management and policy in Flanders. A multi-method approach. Leuven: KU Leuven; 2023. Available from: https://flaquum.org/wp-content/uploads/2024/04/final-phd_jonasbrouwers_feb2023-1.pdf
48. Van de Voorde C, Lefèvre M, Van den Heede K. Expenditure control measures in DRG-based hospital payment systems. *Health Services Research (HSR)*. Brussels: Belgian Health Care Knowledge Centre (KCE); 2024 20/12/2024. KCE Reports 392 (382C) Available from: <https://doi.org/10.57598/R392C>
49. Bouckaert N, Savoye I, Zeevaert R, Devos C, San Miguel L, Van den Heede K. Shifting care from inpatient towards day-care and ambulatory care: impact on readmission rates & post-discharge mortality. *Health Services Research (HSR)*. Brussels: Belgian Health Care Knowledge Centre (KCE); 2026 08/01/2026. KCE Reports 412 Available from: <https://doi.org/10.57598/R412C>
50. Audit Ziekenhuizen. Rapport Audit Zware Medische Beeldvorming. 2022. Available from: https://www.riziv.fgov.be/nl/themas/zorgkwaliteit/Paginas/audit_medische_beeldvorming/Index.html

51. Audit Ziekenhuizen. Heupvervangende ingrepen: het zorgproces in Belgische ziekenhuizen verbeteren. 2024. Available from: <https://www.riziv.fgov.be/nl/pers/kwaliteit-heupvervangende-ingrepen-in-belgische-ziekenhuizen-onderzocht>
52. RIZIV-INAMI. Indicator over arthroscopische meniscectomie bij patiënten ouder dan 50 jaar [Web page].2026 [cited 23/04/2026]. Available from: <https://www.riziv.fgov.be/nl/themas/kwaliteitszorg/indicatoren/indicator-over-arthroscopische-meniscectomie-bij-patienten-ouder-dan-50-jaar>
53. Koninklijk besluit betreffende de vaststelling en de vereffening van het budget van financiële middelen van de ziekenhuizen, B.S. 25 April 2002. Available from: <https://www.ejustice.just.fgov.be/eli/besluit/2002/04/25/2002022335/justel>
54. RIZIV-INAMI. Overeenkomst betreffende de financiering van dialyse (2024-2026). 2025. Available from: https://www.riziv.fgov.be/SiteCollectionDocuments/overeenkomst_financiering_hemodialyse.pdf
55. Audit Ziekenhuizen. Chronische nierinsufficiëntie: audit en aanbevelingen voor een meer gediversifieerd aanbod van patiëntgerichte zorg. 2025. Available from: <https://www.riziv.fgov.be/nl/pers/chronische-nierinsufficiëntie-audit-en-aanbevelingen-voor-een-meer-gediversifieerd-aanbod-van-patientgerichte-zorg>
56. FPS Public Health. AIM project document. Brussels: 2026 10/04/2026. Available from: <https://www.gezondbelgie.be/images/2024%20FOD%20Volksgezondheid%20-%20Blikvanger%20-%20ENG%20interactief.pdf>
57. FPS Public Health. P4P-indicatorenset 2026. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2026. Available from: https://www.health.belgium.be/sites/default/files/media/files/2026-04/nota_p4p-2026_nl_def.pdf
58. Vlayen A. Pay-for-performance (P4P). [Preprint]. 2019]Available from: https://overlegorganen.gezondheid.belgie.be/sites/default/files/documents/1-vlayen annemie - hilde peleman 2019-10-2-workshop_antibioticabeleid_hpn_definitief.pptx.pdf
59. Brouwers J, Seys D, Claessens F, Van Wilder A, Bruyneel L, De Ridder D, et al. Effect on hospital incentive payments and quality performance of a hospital pay for performance (P4P) programme in Belgium. J Healthc Qual Res. 2024;39(3):147-54.
60. FPS Public Health. P4P-indicatorenset 2019. Pay for Performance programma voor algemene ziekenhuizen - versie 2.0. Brussels: 2019.
61. FPS Public Health. Pay for Performance-programma 2018 voor algemene ziekenhuizen. Begeleidende nota. Brussels: 2018.
62. FPS Public Health. P4P-indicatorenset 2020. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2020.
63. FPS Public Health. P4P-indicatorenset 2021. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2021.
64. FPS Public Health. Belonen van verbetering in het programma Pay-for-Performance 2022. Tweede voorstel aan de strategische werkgroep P4P (impactanalyse). Brussels: 2021.
65. FPS Public Health. P4P-indicatorenset 2025. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2025.
66. FPS Public Health. P4P-indicatorenset 2024. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2024.
67. FPS Public Health. P4P-indicatorenset 2023. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2023.
68. FPS Public Health. P4P-indicatorenset 2022. Pay for Performance programma voor algemene ziekenhuizen. Brussels: 2022.



69. Jegers M, Kesteloot K, De Graeve D, Gilles W. A typology for provider payment systems in health care. *Health Policy*. 2002;60(3):255-73.
70. McKenzie J, Brennan S. Synthesizing and presenting findings using other methods [last updated October 2019]. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5. cochrane.org/handbook: Cochrane; 2024.
71. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*. 2021;372.
72. Van den Noortgate W, Lopez-Lopez JA, Marin-Martinez F, Sanchez-Meca J. Three-level meta-analysis of dependent effect sizes. *Behav Res Methods*. 2013;45(2):576-94.
73. Moeyaert M, Ugille M, Natasha Beretvas S, Ferron J, Bunuan R, Van Den Noortgate W. Methods for dealing with multiple outcomes in meta-analysis: a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*. 2017;20(6):559-72.
74. Scott A, Liu M, Yong J. Financial Incentives to Encourage Value-Based Health Care. *Medical Care Research and Review*. 2016;75(1):3-32.
75. Zaresani A, Scott A. Is the evidence on the effectiveness of pay for performance schemes in healthcare changing? Evidence from a meta-regression analysis. *BMC Health Services Research*. 2021;21(1).
76. Pawson R, Tilley N. *Realistic evaluation*. London; Thousand Oaks, Calif: Sage; 1997.
77. Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 2006;3(2):77-101.
78. Merton RK. *Social Theory and Social Structure*. Free Press; 1968.
79. Slawomirski L, Otahal P, Hensher M, Campbell J, Newell S, de Graaff B. Financial penalty associated with a decline in hospital-acquired complications in Australia. *Health Policy*. 2025;161.
80. Grimley RS, Collyer TA, Andrew NE, Dewey HM, Horton ES, Cadigan G, et al. Impact of pay-for-performance for stroke unit access on mortality in Queensland, Australia: an interrupted time series analysis. *The Lancet Regional Health–Western Pacific*. 2023;41.
81. Grimley R, Kim J, Dewey HM, Andrew NE, Collyer TA, Horton ES, et al. The impact of pay-for-performance incentives for stroke unit access on public hospital costs and use, Queensland, 2012–17: interrupted time series analysis. *Medical Journal of Australia*. 2025;222(5):249-58.
82. Vermeulen MJ, Stukel TA, Boozary AS, Guttman A, Schull MJ. The Effect of Pay for Performance in the Emergency Department on Patient Waiting Times and Quality of Care in Ontario, Canada: A Difference-in-Differences Analysis. *Annals of Emergency Medicine*. 2016;67(4):496-505.e7.
83. Chan Y-CL, Hsu SH. Target-Setting, Pay for Performance, and Quality Improvement: A Case Study of Ontario Hospitals' Quality-Improvement Plans. *Revue Canadienne des Sciences de l'Administration/Canadian Journal of Administrative Sciences*. 2019;36(1):128-44.
84. Staples JA, Liu G, Brubacher JR, Karimuddin A, Sutherland JM. Physician Financial Incentives to Reduce Unplanned Hospital Readmissions: an Interrupted Time Series Analysis. *Journal of General Internal Medicine*. 2021;36(11):3431-40.
85. Meacock R, Rud Kristensen S, Sutton M. The Cost-Effectiveness of Using Financial Incentives to Improve Provider Quality: A Framework and Application. *Health Economics*. 2014;23(1):1-13.
86. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *New England Journal of Medicine*. 2012;367(19):1821-8.

87. Britteon P, Kristensen SR, Lau YS, McDonald R, Sutton M. Investigating the Spillover Mechanisms of Payment Incentives on the Outcomes for Non-Targeted Patients. *Health Economics (United Kingdom)*. 2025;34(7):1274-94.
88. Kristensen SR, Meacock R, Turner AJ, Boaden R, McDonald R, Roland M, et al. Long-term effect of hospital pay for performance on mortality in England. *New England Journal of Medicine*. 2014;371(6):540-8.
89. 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. 2015.
90. Metcalfe D, Zogg CK, Judge A, Perry DC, Gabbe B, Willett K, et al. Pay for performance and hip fracture outcomes: An interrupted time series and difference-in-differences analysis in England and Scotland. *Bone and Joint Journal*. 2019;101 B(8):1015-23.
91. Grasic K, Villasenor A, Gaughan J, Gutacker N, Siciliani L. Incentivizing Hospital Quality Through Care Bundling. *Health Economics*. 2025;34(11).
92. Zogg CK, Metcalfe D, Judge A, Perry DC, Costa ML, Gabbe BJ, et al. Learning from England's Best Practice Tariff: Process Measure Pay-for-Performance Can Improve Hip Fracture Outcomes. *Annals of Surgery*. 2022;275(3):506-14.
93. Núñez-Elvira A, Feng Y, Kristensen SR, Lorgelly P, Meacock R, Siciliani L, et al. Does pay for performance affect socioeconomic inequalities in access? Evidence from hospital specialised care in England. *Health Policy*. 2025;105365.
94. Feng Y, Kristensen SR, Lorgelly P, Meacock R, Nunez-Elvira A, Rodes-Sanchez M. Pay-for-Performance Incentives for Specialised Services in England: A Mixed Methods Evaluation. *European Journal of Health Economics*. 2024;25(5):857-76.
95. Lalloue B, Jiang S, Girault A, Ferrua M, Loirat P, Minvielle E. Evaluation of the effects of the French pay-for-performance program-IFAQ pilot study. *International Journal for Quality in Health Care*. 2017;29(6):833-7.
96. Herbst T, Foerster J, Emmert M. The impact of pay-for-performance on the quality of care in ophthalmology: Empirical evidence from Germany. *Health Policy*. 2018;122(6):667-73.
97. Peluso A, Berta P, Vinciotti V. Do pay-for-performance incentives lead to a better health outcome? *Empirical Economics*. 2019;56(6):2167-84.
98. Minato K, Kunisawa S, Imanaka Y. Early effect of a financial incentive for surgeries within 48 h after hip fracture on the number of expedited hip fracture surgeries, in-hospital mortality, perioperative morbidity, length of stay and inpatient medical expenses. *Journal of Evaluation in Clinical Practice*. 2025;31(3):e14189.
99. Honda Y, Shin JH, Kunisawa S, Fushimi K, Imanaka Y. Impact of a financial incentive on early rehabilitation and outcomes in ICU patients: A retrospective database study in Japan. *BMJ Quality and Safety*. 2025;34(9):601-8.
100. Park Y, Kim JH, Lee KS. Changes in cesarean section rate before and after the end of the Korean Value Incentive Program. *Medicine (United States)*. 2022;101(33):E29952.
101. Kuo RNC, Chung KP, Lai MS. Effect of the pay-for-performance program for breast cancer care in Taiwan. *Journal of Oncology Practice*. 2011;7(3 SUPPL.):e8s-e15s.
102. Chu Y-J, Shih H-M, Chu Y-R, Kung P-T, Tsai W-C. The emergency department quality improvement program improves the prognosis of patients with stroke in Taiwan: A nationwide cohort study. *Scientific reports*. 2025;15(1):35473.
103. Chakravarty S, Lloyd K, Cantor JC. The Impact of Payment Reforms on the Safety Net: Examining Effects of the New Jersey Delivery System Reform Incentive Payment Program on Quality of Care Among Medicaid Beneficiaries. *Population Health Management*. 2022;25(6):703-11.



104. Vaz LE, Kleinman KP, Kawai AT, Jin R, Kassler WJ, Grant PS, et al. Impact of Medicare's Hospital-Acquired Condition Policy on Infections in Safety Net and Non-Safety Net Hospitals. *Infection Control and Hospital Epidemiology*. 2015;36(6):649-55.
105. Kim KM, White JS, Max W, Chapman SA, Muench U. Evaluation of Clinical and Economic Outcomes following Implementation of a Medicare Pay-for-Performance Program for Surgical Procedures. *JAMA Network Open*. 2021;4(8):e21115.
106. Morgan DJ, Meddings J, Saint S, Lautenbach E, Shardell M, Anderson D, et al. Does nonpayment for hospital-acquired catheter-associated urinary tract infections lead to overtesting and increased antimicrobial prescribing? *Clinical Infectious Diseases*. 2012;55(7):923-9.
107. Gidwani-Marszowski R, Bhattacharya J. CMS Reimbursement Reform and the Incidence of Hospital-Acquired Pulmonary Embolism or Deep Vein Thrombosis. *Journal of General Internal Medicine*. 2015;30(5):588-96.
108. Kwong JZ, Weng Y, Finnegan M, Schaffer R, Remington A, Curtin C, et al. Effect of Medicare's Nonpayment Policy on Surgical Site Infections Following Orthopedic Procedures. *Infection control and hospital epidemiology*. 2017;38(7):817-22.
109. Waters TM, Daniels MJ, Bazzoli GJ, Perencevich E, Dunton N, Staggs VS, et al. Effect of medicare's nonpayment for hospital-acquired conditions: Lessons for future policy. *JAMA Internal Medicine*. 2015;175(3):347-54.
110. Schuller K, Probst J, Hardin J, Bennett K, Martin A. Initial impact of Medicare's nonpayment policy on catheter-associated urinary tract infections by hospital characteristics. *Health Policy*. 2014;115(2-3):165-71.
111. Lee GM, Kleinman KP, Soumerai SB, Tse A, Cole D, Fridkin SK, et al. Effect of nonpayment for preventable infections in U.S. Hospitals. *New England Journal of Medicine*. 2012;367(15):1428-37.
112. Calderwood MS, Vaz LE, Tse Kawai A, Jin R, Rett MD, Grant PS, et al. Impact of Hospital Operating Margin on Central Line-Associated Bloodstream Infections Following Medicare's Hospital-Acquired Conditions Payment Policy. *Infection control and hospital epidemiology*. 2016;37(1):100-3.
113. Peasah SK, McKay NL, Harman JS, Al-Amin M, Cook RL. Medicare non-payment of hospital-acquired infections: infection rates three years post implementation. *Medicare & Medicaid Research Review*. 2013;3(3).
114. Thirukumaran CP, Glance LG, Rosenthal MB, Temkin-Greener H, Balkissoon R, Mesfin A, et al. Impact of Medicare's Nonpayment Program on Venous Thromboembolism Following Hip and Knee Replacements. *Health Services Research*. 2018;53(6):4381EP-402.
115. Thirukumaran CP, Glance LG, Temkin-Greener H, Rosenthal MB, Li Y. Impact of Medicare's Nonpayment Program on Hospital-acquired Conditions. *Medical Care*. 2017;55(5):447-55.
116. Arntson E. Changes in hospital acquired conditions and mortality associated with the hospital acquired condition reduction program. *J Gen Intern Med*. 2019;34(2 Supplement):S159.
117. Rude TL, Donin NM, Cohn MR, Meeks W, Gulig S, Patel SN, et al. Analysis of National Trends in Hospital Acquired Conditions Following Major Urologic Surgery Before and After Implementation of the Hospital Acquired Condition Reduction Program. *Urology*. 2018;119(wsy, 0366151):79-84.
118. Hsu HE, Wang R, Broadwell C, Horan K, Jin R, Rhee C, et al. Association between Federal Value-Based Incentive Programs and Health Care-Associated Infection Rates in Safety-Net and Non-Safety-Net Hospitals. *JAMA Network Open*. 2020;3(7) (no pagination).
119. Banerjee S, Paasche-Orlow MK, McCormick D, Lin M-Y, Hanchate AD. Association between Medicare's Hospital Readmission Reduction Program and readmission rates across hospitals by medicare bed share. *BMC health services research*. 2021;21(1):248.



120. Alrawashdeh M, Rhee C, Hsu H, Wang R, Horan K, Lee GM. Assessment of Federal Value-Based Incentive Programs and In-Hospital *Clostridioides difficile* Infection Rates. *JAMA Network Open*. 2021;4(10) (no pagination).
121. Bischoff K, Goel A, Hollander H, Ranji SR, Mourad M. The housestaff incentive program: Improving the timeliness and quality of discharge summaries by engaging residents in quality improvement. *BMJ Quality and Safety*. 2013;22(9):768-74.
122. Epstein AM, Jha AK, Orav EJ. The impact of pay-for-performance on quality of care for minority patients. *Am J Manag Care*. 2014;20(10):e479-86.
123. Nicholas LH, Dimick JB, Iwashyna TJ. Do hospitals alter patient care effort allocations under pay-for-performance? *Health Services Research*. 2011;46(1 PART 1):61-81.
124. Shih T, Nicholas LH, Thumma JR, Birkmeyer JD, Dimick JB. Does pay-for-performance improve surgical outcomes An evaluation of phase 2 of the premier hospital quality incentive demonstration. *Annals of Surgery*. 2014;259(4):677-81.
125. Ryan AM, Blustein J, Casalino LP. Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Affairs*. 2012;31(4):797-805.
126. Ryan AM. Has pay-for-performance decreased access for minority patients?: Hospitals and nursing homes. *Health Services Research*. 2010;45(1):6-23.
127. Ryan AM. Effects of the premier hospital quality incentive demonstration on medicare patient mortality and cost: Quality and performance. *Health Services Research*. 2009;44(3):821-42.
128. Epstein AM, Joynt KE, Jha AK, Orav EJ. Access to coronary artery bypass graft surgery under pay for performance: evidence from the premier hospital quality incentive demonstration. *Circulation. Cardiovascular quality and outcomes*. 2014;7(5):727-34.
129. Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcomes. *New England Journal of Medicine*. 2012;366(17):1606-15.
130. Jha AK, Orav EJ, Epstein AM. The effect of financial incentives on hospitals that serve poor patients. *Annals of Internal Medicine*. 2010;153(5):299EP-306.
131. Werner RM, Kolstad JT, Stuart EA, Polsky D. The effect of pay-for-performance in hospitals: Lessons for quality improvement. *Health Affairs*. 2011;30(4):690-8.
132. Bonfrer I, Figueroa JF, Zheng J, Orav EJ, Jha AK. Impact of Financial Incentives on Early and Late Adopters among US Hospitals: observational study. *BMJ (Clinical research ed.)*. 2018;360:j5622.
133. Sabbatini AK, Joynt Maddox KE, Liao JM, Basu A, Parrish C, Kreuter W, et al. Accounting for the Growth of Observation Stays in the Assessment of Medicare's Hospital Readmissions Reduction Program. *JAMA Network Open*. 2022;5(11):E2242587.
134. Ziedan E, Kaestner R. Did the Hospital Readmissions Reduction Program Reduce Readmissions? An Assessment of Prior Evidence and New Estimates. *Evaluation review*. 2021;45(6):359-411.
135. Mehtsun WT, Papanicolas I, Zheng J, Orav EJ, Lillemoe KD, Jha AK. National Trends in Readmission Following Inpatient Surgery in the Hospital Readmissions Reduction Program Era. *Annals of Surgery*. 2018;267(4):599-605.
136. Gupta A. Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program. *American Economic Review*. 2021;111(4):1241-83.
137. Banerjee S, Paasche-Orlow MK, McCormick D, Lin MY, Hanchate AD. Readmissions performance and penalty experience of safety-net hospitals under Medicare's Hospital Readmissions Reduction Program. *BMC health services research*. 2022;22(1):338.
138. Ryan AM, Krinsky S, Adler-Milstein J, Damberg CL, Maurer KA, Hollingsworth JM. Association between hospitals' engagement in value-based reforms and readmission reduction in the hospital readmission reduction program. *JAMA Intern. Med*. 2017;177(6):862-8.



139. Nathan AS, Martinez JR, Giri J, Navathe AS. Observational study assessing changes in timing of readmissions around postdischarge day 30 associated with the introduction of the Hospital Readmissions Reduction Program. *BMJ Quality and Safety*. 2020;30(6):493-9.
140. Chhabra KR, Ibrahim AM, Thumma JR, Ryan AM, Dimick JB. Impact of medicare readmissions penalties on targeted surgical conditions. *Health Affairs*. 2019;38(7):1207-15.
141. Desai NR, Ross JS, Kwon JY, Herrin J, Dharmarajan K, Bernheim SM, et al. Association between hospital penalty status under the hospital readmission reduction program and readmission rates for target and nontarget conditions. *JAMA - Journal of the American Medical Association*. 2016;316(24):2647-56.
142. Wasfy JH, Bhambhani V, Healy EW, Choirat C, Dominici F, Wadhera RK, et al. Relative Effects of the Hospital Readmissions Reduction Program on Hospitals That Serve Poorer Patients. *Medical Care*. 2019;57(12):968-76.
143. Ramaswamy A, Reitblat C, Marchese M, Friedlander DF, Newell P, Schoenfeld AJ, et al. Association of the hospital readmission reduction program with readmission and mortality outcomes after coronary artery bypass graft surgery. *Journal of Cardiac Surgery*. 2021;36(9):3251-8.
144. Huckfeldt P, Escarce J, Sood N, Yang Z, Popescu I, Nuckols T. Thirty-Day Postdischarge Mortality Among Black and White Patients 65 Years and Older in the Medicare Hospital Readmissions Reduction Program. *JAMA network open*. 2019;2(3):e190634.
145. Salerno AM, Horwitz LI, Kwon JY, Herrin J, Grady JN, Lin Z, et al. Trends in readmission rates for safety net hospitals and non-safety net hospitals in the era of the US Hospital Readmission Reduction Program: A retrospective time series analysis using Medicare administrative claims data from 2008 to 2015. *BMJ Open*. 2017;7(7) (no pagination).
146. Zuckerman RB, Sheingold SH, Orav EJ, Ruhter J, Epstein AM. Readmissions, observation, and the Hospital Readmissions Reduction Program. *New Engl. J. Med*. 2016;374(16):1543-51.
147. Gupta A, Allen LA, Bhatt DL, Cox M, DeVore AD, Heidenreich PA, et al. Association of the hospital readmissions reduction program implementation with readmission and mortality outcomes in heart failure. *JAMA Cardiol*. 2018;3(1):44-53.
148. Popescu I, Sood N, Joshi S, Huckfeldt P, Escarce J, Nuckols TK. Trends in the Use of Skilled Nursing Facility and Home Health Care Under the Hospital Readmissions Reduction Program: An Interrupted Time-series Analysis. *Medical care*. 2019;57(10):757-65.
149. McGarry BE, Blankley AA, Li Y. The Impact of the Medicare Hospital Readmission Reduction Program in New York State. *Medical care*. 2016;54(2):162-71.
150. Ibrahim AM, Nathan H, Thumma JR, Dimick JB. Impact of the Hospital Readmission Reduction Program on Surgical Readmissions Among Medicare Beneficiaries. *Annals of surgery*. 2017;266(4):617-24.
151. Deore A, Krishnan R, Nair A. Virtuous spillover effects of quality penalties on the continuity of health care. *Production and Operations Management*. 2023;32(9):2846-63.
152. Mellor J, Daly M, Smith M. Does It Pay to Penalize Hospitals for Excess Readmissions? Intended and Unintended Consequences of Medicare's Hospital Readmissions Reductions Program. *Health Economics (United Kingdom)*. 2017;26(8):1037-51.
153. Hasan MM, Noor EAM, Wang X, Zepeda ED, Young JG. Hospital Readmissions to Nonindex Hospitals: Patterns and Determinants Following the Medicare Readmission Reduction Penalty Program. *Journal for Healthcare Quality*. 2020;42(1):E10-E7.
154. Waters TM, Burns N, Kaplan CM, Graetz I, Benitez J, Cardarelli R, et al. Combined impact of Medicare's hospital pay for performance programs on quality and safety outcomes is mixed. *BMC health services research*. 2022;22(1):958.



155. Figueroa JF, Tsugawa Y, Zheng J, Orav EJ, Jha AK. Association between the Value-Based Purchasing pay for performance program and patient mortality in US hospitals: observational study. *BMJ*. 2016;353.
156. Ryan AM, Burgess JF, Pesko MF, Borden WB, Dimick JB. The early effects of Medicare's mandatory hospital pay-for-performance program. *Health Services Research*. 2015;50(1):81-97.
157. Besstremyannaya G, Golovan S. Provider Altruism in Incentives Contracts: Medicare's Quality Race. *HSE Economic Journal*. 2022;26(3):375-403.
158. Kyalwazi AN, Narasimmaraj P, Xu J, Song Y, Wadhera RK. Medicare's Value-Based Purchasing And 30-Day Mortality At Hospitals Caring For High Proportions Of Black Adults. *Health Affairs*. 2024;43(1):118-24.
159. Papanicolas I, Figueroa JF, John Orav E, Jha AK. Patient hospital experience improved modestly, but no evidence medicare incentives promoted meaningful gains. *Health Affairs*. 2017;36(1):133-40.
160. Banerjee S, McCormick D, Paasche-Orlow MK, Lin M-Y, Hanchate AD. Association between degree of exposure to the Hospital Value Based Purchasing Program and 30-day mortality: experience from the first four years of Medicare's pay-for-performance program. *BMC Health Services Research*. 2019;19(1):921.
161. Ryan AM, Krinsky S, Maurer KA, Dimick JB. Changes in hospital quality associated with hospital value-based purchasing. *New England Journal of Medicine*. 2017;376(24):2358-66.
162. Barbash IJ, Pike F, Gunn SR, Seymour CW, Kahn JM. Effects of physician-targeted pay for performance on use of spontaneous breathing trials in mechanically ventilated patients. *American Journal of Respiratory and Critical Care Medicine*. 2017;196(1):56-63.
163. Calikoglu S, Murray R, Feeney D. Hospital pay-for-performance programs in Maryland produced strong results, including reduced Hospital-Acquired conditions. *Health Aff*. 2012;31(12):2649-58.
164. Ryan AM, Blustein J. The effect of the masshealth hospital pay-for-performance program on quality. *Health Services Research*. 2011;46(3):712-28.
165. Herrin J, Nicewander D, Ballard DJ. The effect of health care system administrator pay-for-performance on quality of care. *Joint Commission journal on quality and patient safety / Joint Commission Resources*. 2008;34(11):646-54.
166. Andriole KP, Prevedello LM, Dufault A, Pezeshk P, Bransfield R, Hanson R, et al. Augmenting the Impact of Technology Adoption With Financial Incentive to Improve Radiology Report Signature Times. *Journal of the American College of Radiology*. 2010;7(3):198-204.
167. Kalata S, Reddy RM, Norton EC, Clark MJ, He C, Leyden T, et al. Quality improvement mechanisms to improve lymph node staging for lung cancer: Trends from a statewide database. *The Journal of thoracic and cardiovascular surgery*. 2024;167(4):1469-78.e3.
168. Benzer JK, Young GJ, Burgess Jr JF, Baker E, Mohr DC, Charns MP, et al. Sustainability of Quality Improvement Following Removal of Pay-for-Performance Incentives. *Journal of General Internal Medicine*. 2013:1-6.
169. Feng Y, Kristensen SR, Lorgelly P, Meacock R, Sanchez MR, Siciliani L, et al. Pay for performance for specialised care in England: Strengths and weaknesses. *Health Policy*. 2019;123(11):1036-41.
170. Kondo KK, Damberg CL, Mendelson A, Motu'apuaka M, Freeman M, O'Neil M, et al. Implementation Processes and Pay for Performance in Healthcare: A Systematic Review. *Journal of General Internal Medicine*. 2016;31(Supplement 1):61-9.
171. Mendelson A, Kondo K, Damberg C, Low A, Motu'apuaka M, Freeman M, et al. The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review. *Annals of Internal Medicine*. 2017;166(5).



172. Deci EL, Ryan RM. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology / Psychologie canadienne*. 2008;49(3):182-5.
173. Wigfield A, Eccles JS. Expectancy-Value Theory of Achievement Motivation. *Contemp Educ Psychol*. 2000;25(1):68-81.
174. Vlaev I, King D, Darzi A, Dolan P. Changing health behaviors using financial incentives: a review from behavioral economics. *BMC Public Health*. 2019;19(1):1059.
175. Holmstrom B, Milgrom P. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*. 1991;7:24.
176. Desomer A, Van den Heede K, Triemstra M, Paget J, De Boer D, Kohn L, et al. Use of patient-reported outcome and experience measures in patient care and policy. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2018 07/2018. KCE Reports 303 Available from: <https://doi.org/10.57598/R303C>
177. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. *Int J Clin Pharm*. 2016;38(3):655-62.
178. Kristensen SR, Siciliani L, Sutton M. Optimal price-setting in pay for performance schemes in health care. *Journal of Economic Behavior & Organization*. 2016;123:57-77.
179. Siciliani L, Gaughan J, Gutacker N, Gravelle H, Chalkley M. Paying for Health Gains Using Patient Reported Outcome Measures. *Health Econ*. 2026;35(5):831-47.
180. McDonald R, Kristensen SR, Sutton M, Zaidi S, Todd S, Konteh F, et al. Evaluation of the Commissioning for Quality and Innovation Framework. Final report. Nottingham: University of Nottingham; 2013.
181. Slawomirski L. Pay-for-performance: the case for trying something different. *Health Services Daily* 2024 16 April 2024.
182. Qiu X, Lan C, Li J, Xiao X, Li J. The effect of nurse-led interventions on re-admission and mortality for congestive heart failure: A meta-analysis. *Medicine (Baltimore)*. 2021;100(7):e24599.
183. Van Wilder A, Bruyneel L, De Ridder D, Seys D, Brouwers J, Claessens F, et al. Is a hospital quality policy based on a triad of accreditation, public reporting and inspection evidence-based? A narrative review. *Int J Qual Health Care*. 2021;33(2).