Health policy

Pay-for-performance and patient safety in acute care: a systematic review

--Manuscript Draft--

Manuscript Number:	HEAP-D-23-00634R1
Article Type:	Review article
Section/Category:	Hospital policies
Keywords:	Pay-for-performance; Systematic Review; patient safety; hospital; acute care; financial incentives
Abstract:	Pay-for-performance (p4p) has been tried across all healthcare settings to address ongoing deficiencies in the quality and outcomes of care. The evidence for the effect of these policies has been inconclusive, especially in acute care. This systematic review focused on patient safety p4p in the hospital setting. Using the PRISMA guidelines, we searched five biomedical databases for studies using at least one outcome metric from database inception to March 2023, supplemented by reference tracking and internet searches. We identified 6122 potential titles of which 53 were included: 39 original investigations, eight literature reviews and six grey literature reports. Only five system-wide p4p policies have been implemented, and the quality of evidence was low overall. About half of the studies (52%) included failed to observe improvement in outcomes, with positive findings heavily skewed towards poor quality evaluations. The exception was the Fragility Hip Fracture Best Practice Tariff (BPT) in England, where sustained improvement was observed across various evaluations. All policies had a miniscule impact on total hospital revenue. Our findings suggest the importance of simplicity and transparency in policy design, involvement of clinical community, explicit links to other quality improvement initiatives, and gradual implementation. We also propose a research agenda to lift the quality of evidence in this field.

Revision note

Ref.: **HEAP-D-23-00634**
Pay-for-performance and patient safety in acute care: a systematic review
Health Policy

We sincerely thank both reviewers for their thoughtful feedback, which has improved our paper. Below we respond to each comment, then outline other changes we have made to the manuscript.

Reviewer #2

1. *Given other reviews have focused on the effects of incentives in hospital settings and on hospital safety, it might be helpful to further elaborate on what specifically this review adds to the existing ones.*

    - We agree and have provided a stronger justification for this review in the Introduction (see p2 marked up version - the new text is pasted below). Thank you.

      "Our paper adds to the evidence for two main reasons. First, we were able to identify any reviews specific to acute care safety (the only review we found focused on a specific policy in the United States [65]). Yet safety is arguably the most fundamental aspect of good health care [1, 24]. It is pre-eminent in the Institute of Medicine's influential definition of quality, as care that is: 1. safe, 2. person-centred, 3. effective, 4. efficient, 5. timely and 6. equitable [1]. Even from a narrow perspective like a single intervention on a single patient, achieving quality is impossible without ensuring safety. Whether the same can be said about, say, efficiency can be debated. Second, the most recent review of hospital p4p we identified [17] was published before recent p4p initiatives of interest had been evaluated."

2. *In the methods section, the authors do not mention whether their search strategy was developed in consultation with a librarian or whether a librarian peer review was conducted. My understanding is that the former is typically standard practice, and the latter is recommended to ensure the search strategy is sound.*

    - The search strategy was indeed developed in consultation with a university librarian – we have added this information (see p3 marked up version). Thank you.

3. *Again in the methods section, the authors mention that they included only quantitative studies. However, given their objectives included identifying effective incentive features, and highlighting any unintended consequences, I find it surprising they did not opt for a scoping review approach that would allow for a broader range of study types. It seems to me that an analysis of themes related to incentive design and unintended consequences should include relevant qualitative research. In addition, I wonder if this would have allowed for the use of a deductive framework to guide the analysis of the incentive design and implementation.*

- We opted for quantitative studies for several reasons. First, this review is part of a larger project examining the effect of an Australian p4p policy on the incidence of HACs. This will be a quantitative analysis (Interrupted Time Series) using eight years of national data. As such, we were primarily interested in whether these initiatives have a discernible effect on patient outcomes such as adverse event rates and believe that a systematic review of quantitative studies provides the appropriate scope.
- We thought it would be of value to examine if the effective policies shared any common features. While this *is* a qualitative exercise, we see it as being adequately informed by quantitative studies with additional background research on the contextual aspects of the initiatives (which we have done). The features associated with success or failure are also discussed in the reviews and grey literature we included.
- Our protocol registered *ex ante* with PROSPERO reflects this, limiting the scope for changes at this stage.
- We have added more content to the Limitations section (see p18 marked up version):

  "Likewise, excluding qualitative studies limited our capacity of this review to categorically identify factors associated with success, only the distinguishing features of successful policies (our secondary objective). The fact that our review found just one successful policy reduces the impact of this."

- We believe that an inductive framework is appropriate to identify common features of successful policies. However, we have decided to relinquish the objective concerning unintended consequences, which we agree is beyond scope. We have amended the wording of the relevant section to reflect these changes (see p2 marked up version, also below):

  "The primary objective of our review was to identify p4p initiatives that have reduced the incidence of iatrogenic harm in acute care. Our secondary objective was to identify distinguishing features of successful policies."

- However, we have retained some reference to the findings (now marked incidental) regarding unintended consequences on p9-10 and p14 of the marked up version (also below). We are happy to be guided by the reviewers as to whether this should be retained or discarded.

  "While not an objective of the review, we saw little evidence for negative unintended consequences on patients who were classified into a minority group based on race or ethnicity, in hospitals serving these minority populations, or on non-incentivised conditions or interventions [38, 49, 55, 66, 70]. Two studies investigating the Advancing Quality Initiative in England found evidence for a *positive* spillover effect on outcomes in for non-incentivised interventions at participating hospitals as well as outcomes at non-participating hospitals [44, 45]. Positive spillover was also noted in one literature review [14]."

"We saw little evidence to suggest detrimental effects of the p4p initiatives examined such as the worsening of outcomes for non-incentivised aspects of care, and promulgating health disparities based on race, ethnicity, or socio-economic status. This, albeit incidental, observation contradicts the existing narrative about the negative impact of p4p, especially on minority populations [12, 96]. A dedicated systematic review would be needed to clarify this important issue."

- This issue is also addressed in the new section proposing a research agenda as proposed by Reviewer #3, point 5 (see p17 marked up version). Thank you.

4. *The authors also mention that they included only studies that focused on clinical endpoints. However, given pay-for-performance incentives are meant to change behaviours, should not the success of incentives focus on outcomes related to this behavioural change, rather than the endpoints that result from the behavioural change? There could be many other factors that contribute to the lack of change in clinical endpoints that are unrelated to the incentives themselves. Thus, focusing on the behavioural change (e.g., process measures related to hospital safety) seems more proximate to the effect of the incentives. Did the authors consider this and could they comment on their justification for only focusing on clinical endpoints?*

- Our pre-registered protocol states that we will focus on "incidence and severity of patient harm" and there is limited scope to change this. The decision was based on advice from the patient representative working with us on the broader project mentioned above – their view being that it is outcomes that matter to patients. We agree with this patient-centric standpoint (regrettably, we did not find any studies that used patient-reported outcomes). The primary aim of these initiatives we examined is to reduce harm and we believe that is what they should be evaluated on.
- It is certainly true that other factors unrelated to p4p can prevent improved processes translating to outcomes. But our view is that these factors should be considered in the design and implementation of a p4p initiative whether they arise as part of introducing the initiative (intentionally or not) or are already present, part of the day-to-day operations of the hospital or health service. Either way, they are an inherent part of the policy's design and implementation. To put it another way, the policy's success/failure is partly of function of it: a) including new processes and practices that facilitate behaviours translating to outcomes, or b) giving (in)sufficient consideration to pre-existing factors that may influence the right behaviours.
- We have also added this in the limitations (p17-18 marked up version):

  "Excluding studies that did not measure clinical endpoints may have dampened our findings. However, we were interested in outcomes, which matter most to patients, clinicians, and policy makers (notably, we identified no studies that used patient-reported measures). While structural and process measures can have instrumental value, they do not appear to manifest in health outcomes such as fewer deaths or adverse events [89]."

- Our search and screening process captured seven studies using only process measures. We kept a record of these. Only one observed a positive effect (in the

HQID program, a modest improvement in composite quality process measures for acute myocardial infarction, congestive heart failure, and pneumonia care). We have added new text to inform readers that studies using process metrics reflect our outcomes-based findings (see p6 and p17 marked up version – also below).

"Our initial search identified seven studies using only process measures. These were excluded, but it is worth noting that only one of them observed an effect – a modest improvement in quality processes at hospitals participating in the HQID project [101]."

"Of the seven studies we excluded for using only process metrics only one observed an effect [101]."

- We must also point out that we excluded only studies – not initiatives – that relied just on process measures. In fact, the only 'successful' initiative our review identified – hip fracture Best Practice Tariffs (BPTs) – incentivises adherence to six *processes* that constitute best practice (e.g. surgery within a specified timeframe) and not outcomes. However, we only included evaluations that used outcomes (e.g. mortality, mobility status).
- This issue is also addressed in the new section proposing a research agenda as proposed by Reviewer #3, point 5 (see p17 marked up version). Thank you.

5. *I think it is important that the actual sources for grey literature be listed. What institutional websites were selected? How were they selected? Given Google searches are practically infinite, what was the stopping criteria? How was the search strategy implemented in the Google? Was an Advanced Google Search employed?*

   - We have added content in the manuscript to address this (see p3 marked up version – also below). Thank you.

     "To identify grey literature, we conducted a Google search using the free text terms provided in the supplementary material, stopping the search on the fifth page, which yielded no relevant results. , We also searched the websites of the following national and international agencies: WHO, OECD, the European Commission (EC), NICE, the Australian Commission on Safety and Quality in Health Care (ACSQHC), Institute for Healthcare Improvement (IHI), Agency for Healthcare Research and Quality (AHRQ), RAND Corporation, Centers for Medicare and Medicaid Services (CMS), and the Health Quality and Safety Commission New Zealand. These were selected based on the authors' knowledge of the field as well as input from two academic experts in quality and safety."

6. *Did the authors consider conducting data extraction in duplicate to ensure reliability and accuracy?*

   - We did consider this but unfortunately faced limited resources. Thank you.

7. *On page 3, line 37, the authors state that they considered non-statistically significant results as zero. I disagree with this approach. A result not being statistically significant*

does not mean it has a zero effect. In addition, the use of the alpha = 0.05 threshold should be justified. If a relatively low-cost incentive had a large effect (in terms of magnitude) with a p-value of 0.051, should this not be discussed? I wonder if statistical significance should be the benchmark for discussing the importance of a study's outcome. Perhaps some other considerations, such as the magnitude of the effect size, should also be taken into account.

- We take this point. After all, many studies report results when a p-value is barely *under* the threshold but not when it is barely over. While we would stop short of saying that effect size is what matters irrespective of how wide the confidence interval may be, we agree that readers would benefit from being made aware of results such as the one you illustrate (large effect but just over the *alpha*). We have re-examined all included studies for results that show improvement of >15% with a p-value between 0.05 and 0.10. We found one (Rude et al 2018) and have flagged this in the methods and results sections (see p4 and p6 marked up version) and table 1.

8. *Page 2, line 22: I recommend defining the "6-domain definition of quality."*

- Agreed (see p2 marked up version and point #1 above). Thank you.

9. On page 3, line 30 the authors state *"For studies that presented observations in percentage-points change in rates, we calculated the relative change from the figures provided. We categorised effect magnitude as: zero (0), low (1-14% relative improvement), moderate (15-49% relative improvement), and high (50% or greater relative improvement)." I am curious how these ranges were determined. Are they based on widely accepted ranges? Or are they somewhat arbitrary?*

- We developed these based on the distribution of the results, which was left-skewed (clustered around zero) with few results >50%, and what would seem to be sensible, intuitive categories. We have inserted a sentence to reflect this (see p4 marked up version – also pasted below). Thank you.

  "We categorised effect magnitude as: zero (0), low (1-14% relative improvement), moderate (15-49% relative improvement), and high (50% or greater relative improvement). These cutoffs were chosen based on the left-skewed distribution of the results (the median effect magnitude was zero)."

10. *Page 9, line 30: The authors mention that they excluded studies that did not use outcome measures as an independent variable. However, I believe this was mentioned as an inclusion criterion. In fact, the authors do not explicitly mention any exclusion criteria, which should be included. Alternatively, if no exclusion criteria were used, this should also be mentioned.*

- We have amended the manuscript to make inclusion/exclusion criteria more explicit (see p3 marked up version). Thank you.

Reviewer #3

1. *The authors should offer a stronger justification for the relevance of their systematic literature review by explaining its place within the existing body of literature. Within their literature review, the authors incorporate seven other literature reviews. It's crucial for them to clarify how their analysis differs from these previous reviews. I appreciate that in the Discussion section the authors emphasize that "this is the first review of p4p focusing on hospital safety and our conclusions align with literature reviews covering the topic more broadly". However, the contribution and relevance of the work presented should be better justified and presented preferably in the introduction section.*

    - We agree and have provided a stronger justification for this review in the Introduction (see p2 marked up version - the relevant paragraph is pasted below). Thank you.

      "Our paper adds to the evidence for two main reasons. First, we were able to identify any reviews specific to acute care safety (the only review we found focused on a specific policy in the United States [65]). Yet safety is arguably the most fundamental aspect of good health care [1, 24]. It is pre-eminent in the Institute of Medicine's influential definition of quality, as care that is: 1. safe, 2. person-centred, 3. effective, 4. efficient, 5. timely and 6. equitable [1]. Even from a narrow perspective like a single intervention on a single patient, achieving quality is impossible without ensuring safety. Whether the same can be said about, say, efficiency can be debated. Second, the most recent review of hospital p4p we identified [17] was published before recent p4p initiatives of interest had been evaluated."

2. *The authors also claim that, regarding the quality of the studies, 19 out of the 40 original investigations were assessed as 'poor'. However, it is not very clear how the quality of the papers evaluated impacted the conclusions reached by the authors of the manuscript. Were the findings of the 'poor' studies considered in the same way as the findings of the 'good' studies? This should be clarified in the manuscript.*

    - A good point. We do this in a qualitative fashion and refer to it in the manuscript – but have added some content to emphasis this in various places throughout, including the point that the generally poor-quality evidence undermines the null findings as well as the higher magnitude effects (see several examples below underlined). Thank you.

      "Two investigators independently assessed study quality using the National Institute for Health and Care Excellence (NICE) Quality Appraisal Checklist for quantitative intervention studies, assigning each paper a combined rating of 'poor', 'fair' or 'good' based on our assessment of internal and external validity [26]. […] We considered the general quality of the evidence as well as the quality of indidual studies evaluating smaller, single-site initatives when forming our conclusions." (p5)

      "Eight studies (15%) observed low magnitude effect [36, 43-46, 62, 67, 68], thirteen (24%) a moderate effect [33, 35, 41, 47, 48, 53, 54, 56, 59, 60, 62, 69, 76], and five (7%) a 'high' effect [34, 36, 52, 58, 75]. None of the latter came from 'good' quality studies (three were assessed

as ‘poor’, two as ‘fair’). […] with 11 of these (69%) examining non-mortality outcomes derived from claims data found a moderate or high effect. All but one of the five studies observing an effect of 50% or greater relied on claims data. Papers of lower quality were more likely to observe an effect, and the effect was typically of greater magnitude than in studies of higher quality. No studies rated as ‘good’ observed a result of 15% or greater, while two fifths of studies rated ‘fair’ and ‘poor’ did.” (p6-7)

“Concerning the three main p4p policies in the United States (see Supplementary material), 11 of the 17 HAC-POA studies observed a favourable effect [33-36, 52, 53, 54, 56, 58, 59, 76], with eight of the papers producing these studies were rated ‘poor’ and the remaining five as ‘fair’. The HACRP initiative was observed to have an effect in just two of the 10 studies [47, 48] […] Of the eight papers producing these results, six were rated ‘fair’ and two ‘poor’.” (p7)

“The rest were single-site studies observing mortality reductions ranging from 8% to 71% for cases satisfying the BPT criteria compared to cases that did not (the latter result stemming from a ‘poor’ quality paper) [62, 67, 68, 75].” (p8)

“Calicoglu et al (2012) examined a Maryland-specific p4p initiative based on the nation-wide HACRP policy, observing a considerably greater effect than the studies examining the other United States p4p initiatives (using claims data), and suggesting more effort on change management and acceptance of the policy among providers as potential reasons for the comparative success [41].” (p10)

“The quality of studies appeared to be associated with the result.” (p12)

“The quality of the evidence undermines our findings. While we do note an inverse association between quality and observed effect, it is possible that better quality may have observed an effect, or that it may not have dampened the effects observed in some of the studies included here. We therefore propose a research agenda for this policy area towards the end of this paper.” (p13)

“Again, better quality evidence is needed to answer these questions.” (p14)

3. *When discussing the Results, in my opinion, the authors should provide further details regarding Figure 1. For example, in the manuscript, the authors mention that 6122 titles and abstracts were identified; however, Figure 1 seems to suggest that 7340 studies were identified. Then, it is mentioned that 1231 were removed, but no explanation is offered for why these references were removed. Furthermore, it is indicated that 6121 studies were screened, not 6122 as the authors seem to suggest. Again, of these, 5975 were excluded, but no indication is offered as to why. The figure also suggests that 146 studies were assessed for eligibility, not 147 as indicated in the manuscript.*

- We apologise for these typographical errors. They have been fixed. The reason for the initial removal of 1231 papers was duplicates and is now provided in the figure. Thank you.

4. *The authors also reference Figure 2, but in fact, Figure 2 comprises a textual description of the main p4p policies. In my opinion, the authors should attempt to systematize this information in a figure and present the textual description of the p4P policies as an Appendix or Supplementary Material*

    - Indeed. We have resubmitted this as supplementary material. Thank you.

5. *I agree with the authors that due to some of the limitations of the previous studies, no strong evidence exists to inform p4p policy design recommendations. However, I believe that that the authors could have explored in more detail the opportunities that these limitations offer for further research in the area. In particular, I believe the literature review would benefit if the authors proposed a research agenda on this topic.*

    - Good suggestion – we have added a short section towards the end of the manuscript (see p18 marked up version – also below). Thank you.

      "A research agenda for p4p in acute care

      Our findings indicate that further experimentation and research on p4p in acute care is warranted, as others have suggested [99]. We therefore propose a research agenda base on six themes:

      1. Promoting collaboration between policy makers and academic institutions in rolling out initiatives in an evaluation-friendly fashion. For example, an international collaborative or 'community of practice' to share ideas, knowledge and best practice on all aspects of p4p from design, implementation to evaluation.
      2. Implementing policies in an incremental way, and attempting to adopt an experimental approach, wherever possible, particularly randomisation of participants, and stratifying design features such as incentive size, direction, and target.
      3. Using a blend of process and outcome metrics in the evaluation studies. This (along with the next point) can shed light on what factors prevent/enable translating behaviours and practices into actual outcomes.
      4. Using mixed methods where possible to generate important contextual information from the 'coal face' in a consistent, reproducible manner with the aim of:
          - Identifying key barriers to, and enablers of, success.
          - In cases where processes but not outcomes improve, identifying potential intrinsic or extrinsic factors that influence this translation.
      5. Incorporating patient-reported outcomes into the design and evaluation of policies.
      6. More concerted efforts to identify the positive and negative spillovers/consequences of p4p – original investigations as well as systematic reviews."

Additional changes

1. Because we have relinquished unintended consequences as an objective, the Results and Discussion sections no longer have sub-headings addressing this topic. However, we have retained edited version of these findings (marked as incidental) as we feel that the information is useful given it contradicts the prevailing narrative on this issue. Examining this issue as a primary objective is proposed as part of the ongoing research agenda (see Reviewer 3, point 5).

2. We have decided to remove the Epstein et al (2014) paper. The focus was on racial disparities under p4p, because of changes to our objectives (as suggested by reviewer #2). We feel that including this result no longer add value to the manuscript – although we cite the paper in our (now incidental) findings regarding racial disparities. This reduces the number of original investigations to 39.

3. We have added a literature review by Milstein and Schreyoegg (2016) to the review. This was erroneously omitted in the first iteration. It increases the number of reviews to eight.

4. Incidental corrections and improvements throughout the manuscript can be found in the marked-up version.

Highlights

- Pay-for-performance (p4p) is used to improve the safety and quality of care.
- Five system-wide p4p initiatives targeting hospital patient safety were identified but the quality of the evidence was low.
- Only one initiative was observed to improve patient outcomes.
- Initiatives should be simple, transparent, and implemented gradually and inclusively.

Highlights

- Pay-for-performance (p4p) is used to improve the safety and quality of care.
- Five system-wide p4p initiatives targeting hospital patient safety were identified but the quality of the evidence was low.
- Only one initiative was observed to improve patient outcomes.
- Initiatives should be simple, transparent, and implemented gradually and inclusively.

Pay-for-performance and patient safety in acute care: a systematic review

Abstract

Pay-for-performance (p4p) has been tried across all healthcare settings to address ongoing deficiencies in the quality and outcomes of care. The evidence for the effect of these policies has been inconclusive, especially in acute care. This systematic review focused on patient safety p4p in the hospital setting. Using the PRISMA guidelines, we searched five biomedical databases for studies using at least one outcome metric from database inception to March 2023, supplemented by reference tracking and internet searches. We identified 6122 potential titles of which 53 were included: 39 original investigations, eight literature reviews and six grey literature reports. Only five system-wide p4p policies have been implemented, and the quality of evidence was low overall. About half of the studies (52%) included failed to observe improvement in outcomes, with positive findings heavily skewed towards poor quality evaluations. The exception was the Fragility Hip Fracture Best Practice Tariff (BPT) in England, where sustained improvement was observed across various evaluations. All policies had a miniscule impact on total hospital revenue. Our findings suggest the importance of simplicity and transparency in policy design, involvement of clinical community, explicit links to other quality improvement initiatives, and gradual implementation. We also propose a research agenda to lift the quality of evidence in this field.

Introduction

Pay-for-performance (p4p) has proliferated in recent decades as part of the policy response to deficiencies in the quality and value of health care [1-8]. Most commonly, p4p aims to improve care quality by financially rewarding desirable structures, practices and outcomes, and/or penalising undesirable ones. P4p draws on rational choice theory of neoclassical economics [9], linking provider utility with remuneration for desired policy outcomes more overtly than existing approaches, including output-based funding models such as fee-for-service. The approach seeks to align clinical risk with corporate risk, especially in larger and more complex organisations where health professionals and management may have fewer common objectives and incentives than smaller practices. Demonstrating clear cut success has proven elusive, however, and over two decades of empirical research on the subject has not made a compelling case for p4p as an effective policy lever [10-12].

P4p has been used in the primary/ambulatory care setting for some time and in many countries, attempting to optimise activities ranging from vaccination to care coordination and adherence to best practice [11, 13-16]. Its deployment in the hospital setting is more recent, taking place mostly in the United States and the United Kingdom [17]. For example, a trio of initiatives introduced by the Centers for Medicare and Medicaid (CMS) in 2012 use a combination of financial rewards and penalties based on a series of process and outcome measures [18-20]. Advancing Quality, introduced in 2008 in northwestern England, incentivised performance during admissions for six hospital interventions based on 28 quality metrics [16]. Best Practice Tariffs (BPTs) were introduced in 2010 across English hospitals to incentivise adherence to best-practice care protocols for a range of conditions and interventions [16]. Elsewhere, France introduced a hospital p4p initiative based on structural and process measures in 2016 [21, 22]. A policy that withholds funding for the incremental cost of 16 hospital-acquired complications was introduced in Australian public hospitals in 2018 [23].

We present the findings of a systematic review of the literature evaluating p4p initiatives targeting patient safety in the hospital setting. Our paper adds to the evidence for two main reasons. First, we were able to identify any reviews specific to acute care safety (the only review we found focused on a specific policy in the United States [65]). Yet safety is arguably the most fundamental aspect of good health care [1, 24]. It is pre-eminent in the Institute of Medicine's influential definition of quality, as care that is: 1. safe, 2. person-centred, 3. effective, 4. efficient, 5. timely and 6. equitable [1]. Even from a narrow perspective like a single intervention on a single patient, achieving quality is impossible without ensuring safety. Whether the same can be said about, say, efficiency can be debated. Second, the most recent review of hospital p4p we identified [17] was published before recent p4p initiatives of interest had been evaluated.

Our primary objective was to identify p4p initiatives that reduced the incidence of iatrogenic harm in acute care. Our secondary objective was to identify any distinguishing features of successful policies.

Methods

Search protocol and sources

The reporting of this systematic review was guided by the standards of the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) Statement [25]. The protocol was registered with PROSPERO (CRD42023400087). We searched three

biomedical databases (Medline, Embase and Web of Science), plus the Centre for Reviews and Dissemination (CRD) Database of Abstracts of Reviews of Effects (DARE), and the Cochrane Database, from database inception to March 2023. We included papers in any language examining p4p initiatives aimed at reducing iatrogenic harm in the hospital setting. The search strategy was developed in consultation with a University of Tasmania librarian. Search terms are provided as a <Supplementary file>.

Papers evaluating p4p initiatives aimed at care quality were included if the initiative (and the study evaluating it) contained at least one safety component. We included only quantitative studies that used as a dependent variable at least one outcome metric such as mortality, adverse event rates, hospital-acquired condition or complication rates, and patient-reported incident- or outcome measures. We excluded qualitative studies and studies that measured processes only. Additional papers were identified through reference lists of included articles, reviews, editorials, and expert recommendations.

To identify grey literature, we conducted a Google search using the free text terms provided in the supplementary material, stopping the search on the fifth page (which yielded no relevant results). We also searched the websites of the following national and international agencies: WHO, OECD, the European Commission (EC), NICE, the Australian Commission on Safety and Quality in Health Care (ACSQHC), Institute for Healthcare Improvement (IHI), Agency for Healthcare Research and Quality (AHRQ), RAND Corporation, Centers for Medicare and Medicaid Services (CMS), and the Health Quality and Safety Commission New Zealand – selected based on the authors' knowledge of the field and input from two academic experts.

Extraction and analysis

Search results were exported to Covidence software to assist with managing eligible papers. Two investigators (LS and BdG) independently screened each title and abstract against the inclusion criteria, and then completed a full text review of the shortlisted papers.

One investigator (LS) abstracted data from each included paper. Results were qualitatively synthesized by location, health system context, specific p4p policy, data source and sample size, outcome metric(s), study design, follow-up period, incentive type (reward/penalty), size and target, and magnitude of effect. For effect magnitude we used relative, not absolute, change (i.e. the difference between the incidence of the respective outcome before and after the policy was introduced, expressed as a percentage). For studies that presented observations in percentage-points change in rates, we calculated the relative change from the figures provided. We categorised effect magnitude as: nil (0), low (up to14% relative improvement), moderate (from 15 to 49% relative improvement), and high (50% or greater

relative improvement). These cutoffs were chosen based on the left-skewed distribution of the results (the median effect magnitude was zero). All results that were not statistically significant at an alpha value of 0.05 were assigned an effect of zero. We chose this threshold because it is the most used in the literature we included. To not categorially exclude results that are marginally over the chosen alpha (while those marginally *under* it are included), we flag those showing an effect of 15% or greater (i.e. Moderate and High) with a p-value between 0.05 and 0.10. We did not perform meta-analysis because all studies were observational and due to the heterogeneity of methods and metrics used.

Quality assessment

Two investigators (LS and BdG) independently assessed study quality using the National Institute for Health and Care Excellence (NICE) Quality Appraisal Checklist for quantitative intervention studies, assigning each paper a combined rating of 'poor', 'fair' or 'good' based on our assessment of internal and external validity [26]. Literature reviews were assessed using the AMSTAR 2 critical appraisal tool for systematic reviews [27]. We considered the general quality of the evidence as well as the quality of indidual studies evaluating smaller, single-site initatives when forming our conclusions.

Results

After removing duplicates, we screened 6,122 titles and abstracts from which 147 potentially eligible full-text titles were assessed for eligibility. We ultimately included 53 titles (Figure 1) comprising 39 original peer-reviewed investigations (Table 1), eight literature reviews (Table 2), and six grey literature reports [13, 28-32]. The earliest study was published in 2006, the latest in 2023. All but two papers were published since 2012, with 17 published in the last five years (14 original investigations, one literature review and two grey literature titles).

Figure 1 here

Study design, methods and quality

Among the original investigations, the two most common study designs were Difference-in-Difference (n=14) [33-46], and Interrupted Time Series (n=13) analysis [47-59]. One paper used both [60]. The mean (median) follow-up time was 3.6 (3.2) years. The longest follow-up was eight years [36], the shortest eight months [61]. We saw no evidence of an association between the length of follow-up and observed effect. Most investigations used claims data (n=23, 59%), with the remainder using infection surveillance and registry data. One used both claims and surveillance data [52].

The most common outcome measures were mortality, and adverse event rates or hospital-acquired condition (HAC) rates. Thirteen original investigations analysed more than one health outcome, or the same outcome over multiple follow-up periods. The 39 original investigations thus yielded 54 results (henceforth referred to as 'studies'). One study used patients' mobility status and residential status in addition to mortality [62]. We found no studies using patient-reported measures. The most common comparison groups were either hospitals not participating in the p4p policy, or diagnoses, interventions or patient groups that were not incentivised under the policy but were managed in participating hospitals.

Nineteen (49%) of the original investigations were assessed as 'poor' quality, 17 (44%) as 'fair', and three (8%) as 'good'. The most common issues were: lack of an appropriate comparator, insufficient observation time before and/or after introduction of the policy, lack of record-level risk adjustment, and the use of claims data, which are not considered as reliable as other data sources for deriving outcomes other than mortality [63, 64]. The seven literature reviews comprised one Cochrane, one integrative, three literature, and three systematic reviews. We assessed the quality of all but one of these as 'good' (Table 2). Three reviews focused on the hospital setting. One of these focused on a single p4p policy targeting safety (the CMS' HAC-POA initiative) while the other two included p4p targeting quality of care more broadly [11, 12, 14-17, 65, 84].

Location of studies and initiatives

Twenty-eight (52%) of the original investigations were conducted in the United States, with all but three papers examining the three main p4p initiatives of the Centers for Medicare and Medicaid (CMS): 1. Premier Hospital Quality Improvement Demonstration (HQID) project; 2. HAC Present on Admission (HAC-POA) policy [18]); and 3. HAC Reduction Program (HACRP) [20]. Three papers examined the effects of local initiatives in California, Hawaii, and Maryland [41, 42, 66]. Ten papers were from England, focusing on the Advancing Quality initiative [43-45], and the Best Practice Tariff (BPT) policy for hip fracture care [45, 46, 60, 62, 67, 68] and chronic obstructive pulmonary disease (COPD) [61]. The remaining study was conducted in Taiwan [69]. Descriptions of the main English and US initiatives are provided in the <Supplementary material>.

Overall findings

Key findings from the original investigations are summarised in Table 1. Twenty-eight studies (52%) failed to observe a statistically significant effect [36-40, 42, 44, 45, 48-53, 55, 57, 61, 62, 69, 71-75] (one of these observed a moderate effect with a p-value of 0.077 [72].) Eight studies (15%) observed a low magnitude effect [36, 43-46, 62, 67, 68], thirteen (24%) a moderate effect [33, 35, 41, 47, 48, 53, 54, 56, 59, 60, 62, 69, 76], and five (9%) a high

effect [34, 36, 52, 58, 75]. None of the latter came from 'good' quality studies (three were assessed as 'poor', two as 'fair'). None of the included studies observed a deleterious effect on incentivised or non-incentivised outcomes. Of the 22 studies using mortality as an outcome, nine (41%) observed an effect and all were evaluations of English initiatives [43-46, 60, 62, 67, 68, 75]. Sixteen (50%) of the remaining studies using non-mortality outcomes observed an effect [33-36, 41, 47, 48, 52, 53, 54, 56, 58, 59, 62, 69, 76], with 11 of these (69%) using claims data observed a moderate or high effect. All but one of the five studies observing an effect of 50% or greater relied on claims data. Papers of lower quality were more likely to observe an effect, and the effect was typically of greater magnitude than in studies of higher quality. No studies rated as 'good' observed a result of moderate or high magnitude, while two fifths of studies rated 'fair' and 'poor' did.

Our initial search identified seven studies using only process measures. These were excluded, but it is worth noting that only one of them observed an effect – a modest improvement in quality processes at hospitals participating in the HQID project [101].

Table 1 here

United States initiatives

Of the studies based in the United States, 21 (58%) did not observe an effect, one (3%) observed a low, ten (28%) a moderate, and four (11%) a high magnitude effect. One study using mortality as the outcome observed a low magnitude effect [42]. Concerning the three main p4p policies in the United States (see Supplementary material), 11 of the 17 HAC-POA studies observed an effect [33-36, 52, 53, 54, 56, 58, 59, 76], with eight of the papers producing these studies rated as 'poor' and the remaining five as 'fair'. The HACRP initiative was observed to have an effect in just two of the 10 studies [47, 48] although one observed a moderate effect with a p-value of 0.077. Of the eight papers producing these results, six were rated 'fair' and two 'poor'. None of the five studies evaluating the patient safety component of the HQID project observed an effect [37-40], although all but one of these used risk-adjusted mortality as the outcome, compared to only a minority of studies evaluating the HAC-POA and HACRP policies. Of the three studies evaluating other, smaller initiatives in the United States, the only one to observe an effect was Calikoglu et al (2012) who, using claims data over a relatively short, 2-year follow-up period, observed an average 15% reduction in HACs included, compared to HACs not included, in a Maryland p4p policy [41].

The grey literature included a patient safety 'scorecard' for admissions under Medicare from the Agency for Healthcare Research and Quality (AHRQ), which reported a 17% reduction in the crude rates of 28 hospital-acquired conditions between 2010 and 2014 and a further

13% between 2014 and 2019 [32]. Although these results cover the period spanning the HAC-POA and the HACRP initiatives and are reflected in some of the original investigations we included [77], the AHRQ scorecard report is not intended to serve as a formal evaluation of these policies. *Inter alia,* it lacks comparators nor does it attempt to separate the p4p component from other policies and activities aimed at reducing hospital-acquired conditions over this period [6].

Initiatives in England

The papers examining English initiatives were generally of higher quality than the others. All used mortality as a dependent variable. Two papers on the Advancing Quality initiative assessed effects over distinct time periods, observing small but significant effect in the short term (up to 18 months post-intervention) but not thereafter (42 months) [44, 45]. One of these observed evidence for a possible positive spillover effect to conditions not covered by this initiative [44].

Six papers examined the hip fracture BPT policy. Two were nation-wide studies. Metcalfe et al (2019) compared English hospitals (included in the scheme) with Scottish counterparts (not included) observing a 6% relative mortality reduction during a 7-year follow-up, as well as reductions in readmissions and average length of stay [46]. Zogg et al (2022) compared England with the United States, observing a 27% relative reduction in mortality during a 6-year follow-up [60]. The rest were single-site studies observing mortality reductions ranging from 8% to 71% for cases satisfying the BPT criteria compared to cases that did not (the latter result stemming from a ‘poor’ quality paper) [62, 67, 68, 75].

Stone et al (2019) examined the COPD BPT using a nation-wide cohort but observing no effect on either mortality or readmissions [61]. We failed to identify peer-reviewed evaluations of other BPTs relevant to this review, such as stroke care, which appeared only in two grey literature reports [30, 31]. These were published within two years of its implementation and suggest that the policy had no impact on outcomes beyond what was achieved by other national stroke care initiatives, with the complexity of that specific BPT variant was cited as a key reason [31].

Taiwan

The Taiwanese study evaluated a nation-wide p4p initiative targeting diabetes care, examining surgical-site infections and revisions following total knee arthroplasty in diabetic patients. It failed to observe any effect on infection rates in patients included in the initiative but observed a significant reduction in the risk of surgical revision [69]. Evidence cited in the

grey literature suggests that this initiative was prone to excluding more complex patients [78].

Reviews and grey literature

Findings from the literature reviews are summarised in Table 2. All failed to find conclusive evidence that financial incentives aimed at hospital care have a sustained, positive effect on patient safety outcomes (more success is reported non-acute care, however). Effects were small at best, with little evidence that any gains were sustained over time. Low quality of the evidence was consistently remarked upon. The grey literature drew similar conclusions, finding the evidence to be of low quality and any effect on outcomes small, unsustained and difficult to separate from other, concurrent quality improvement initiatives [13, 28, 29]. Both the reviews and the grey literature highlighted the importance of context, design and especially implementation in the success of hospital p4p initiatives tested so far with several commenting that the way policies were introduced has created methodological challenges to studying their effects [11, 12, 14-17, 65].

Table 2 here

P4p policy design features

The evidence of an association between effect and design features such as incentive size was inconclusive, mostly due to an insufficient number of high-quality, comparable evaluations of policies that share a similar design but different designs, let alone head-to-head comparisons of different p4p designs.

The impact of the main p4p policies on the total revenue of participating hospitals was:

- HQID (United States): less than 0.01% [39, 79, 80]
- HAC-POA (United States): less than 0.01% [81, 82]
- HACRP (United States): 1% [20, 83]
- Advancing Quality (England): 0.1% [84]
- BPT (England): [85, 86]
    - Hip fracture care: 1% [31, 87]
    - Acute stroke care: 0.2% [31, 87]
    - Acute COPD exacerbation: 1.2% [61, 87, 88]

We did, however, see evidence for an association between incentive size and effect in hospitals stratified according to financial strength. Calderwood et al (2016) observed a decline in HAC rates at hospitals operating at a financial loss, but not in hospitals with positive operating margins, following introduction of the HAC-POA policy [53].

We did not see any consistent associations between effect and the direction (penalty vs. reward) of the incentive. Hospital management and administration was the reported target in all our included studies, and it was unclear if the incentives were 'passed down' to the departments, teams, or individuals responsible for attracting the reward or penalty. We are therefore unable to draw any conclusions about this.

While not an objective of the review, we saw little evidence for negative unintended consequences on patients who were classified into a minority group based on race or ethnicity, in hospitals serving these minority populations, or on non-incentivised conditions or interventions [38, 49, 55, 66, 70]. Two studies investigating the Advancing Quality Initiative in England found evidence for a *positive* spillover effect on outcomes for non-incentivised interventions at participating hospitals as well as outcomes at non-participating hospitals [44, 45]. Positive spillover was also noted in one literature review [14].

Context and implementation

The importance of local context and policy implementation was a common theme throughout the literature we reviewed, in particular: a) ensuring stakeholder acceptance and buy-in, b) providing support and resourcing where needed, and c) tailoring the design to local context [15, 16, 41, 45, 47]. For example, Alrawashdeh et al (2021), who observed a 6% decline in hospital-acquired *clostridium difficile* infection rates associated with the HACRP initiative, argued that the decline "may be due to greater opportunities for improvement in diagnostic and antimicrobial stewardship, which also align with the value these programs intend to achieve" [47]. Calicoglu et al (2012) examined a Maryland-specific p4p initiative based on the nation-wide HACRP policy, observing a considerably greater effect than the studies examining the other United States p4p initiatives (although using claims data), and suggesting more effort on change management and acceptance of the policy among providers as potential reasons for the comparative success [41]. In some cases, improvements in safety began *before* the introduction of a p4p initiative, lending weight to the importance of stakeholder communication and change management in policy implementation [76].

The claims data effect

Studies using HAC or adverse event rates as their outcome were twice as likely to observe an effect if the rates were derived from claims data compared to other data sources, with a four-fold greater effect (22% versus 5%) observed, on average. While the limitations of these data in the research context are known [63, 64], this finding suggests a potential association between p4p and changes in administrative coding. The most striking example is Calderwood et al (2014) who used both Medicare billing and National Healthcare Safety

Network (NHSN) infection surveillance data on 638,761 procedures at 1,234 hospitals to examine the effect of the HAC-POA policy on mediastinitis following coronary artery bypass grafting (CABG). They observed a sudden, 64% reduction in billing for mediastinitis at the introduction of the policy but no corresponding change in rates calculated from NHSN data [52].

Discussion

We present findings from a comprehensive systematic review of p4p targeting iatrogenic harm in the hospital setting. Overall, we found little conclusive evidence for a sustained positive effect of such policies on patient safety. To our knowledge, this is the first review of p4p policies and safety in acute care. Our conclusions generally align with reviews that examined p4p more broadly [11, 12, 14-17, 65, 84]. However, there was one exception: the fragility hip fracture BPT, which was associated with a significant reduction in risk-adjusted mortality, as well as reductions in length of stay, patients' mobility status, and readmissions up to seven years post-implementation [46, 60, 62, 67, 68, 75]. Three features of this policy set it apart from others: it is straightforward, it was not imposed, with key aspects of the development led by stakeholders, and it was implemented gradually. The latter is exemplified through the explicit link to the National Hip Fracture Database (NHFD), a public registry that publishes hospital-level results and was introduced following a clinician-led audit in 2007 – three years before the p4p component began to be applied [90].

Study quality

The quality of studies appeared to be associated with the result. Observed effects varied considerably, even between studies examining the same p4p policies. Studies using mortality as an outcome variable were less likely to observe an effect than those using adverse event or HAC rates. This may reflect differences in the sensitivity of these metrics, although while one may expect a lower effect magnitude, the large datasets in most of the included studies should aid the detection of small effects. Among the studies using adverse event or complication rates, those that drew on claims data were twice as likely to observe an effect than those that used other data sources, averaging a four-fold greater effect magnitude. This was especially pronounced where the study used the same data as those that were used to assess performance by the policy [33-35, 52, 58, 91, 92].

This is perhaps unsurprising. Claims data have low sensitivity and do not correlate with clinically meaningful outcomes [52]. They are also less likely to capture actual occurrence of events and may be subject to reporting bias to improve performance scores [93]. Kawai et al

(2015) examined billing patterns at 569 hospitals before and after the introduction of the HAC-POA policy, focusing on two of the incentivised conditions (central line associated blood stream infection (CLABSI) and catheter-associated urinary tract infections (CAUTI)) and non-incentivised surgical site infections (SSI). For the incentivised outcomes, after a steady rise before the policy was introduced (on average by 17% and 19% per quarter respectively), billing rates then dropped by 25% upon introduction followed by a slightly downward trend for CLABSI and a levelling off for CAUTI during the follow-up period. No discernible change in SSI rates was observed before or after the policy's introduction [91]. Rhee et al (2019) applied the same method to the subsequent introduction of the HAC-POA policy for *Medicaid* patients. They found little impact on billing rates at 546 hospitals for incentivised *and* non-incentivised conditions following its introduction compared to the pre-policy trajectory. CAUTI billing did not change, while CLABSI billing reduced significantly during the pre- and post-introduction periods [92]. We would therefore question not only the validity of studies that rely on these data, but also the use of claims data in p4p policies.

Nevertheless, the low quality of the evidence undermines our findings. While we note an inverse association between quality and observed effect, it is difficult to say if better studies may have observed an effect, or that better quality may not have dampened the effects observed in some of the studies included here. We therefore propose a research agenda for this policy area at the end of the paper.

Design characteristics

We found little evidence to inform p4p policy design recommendations and are unable to recommend a 'carrot' or 'stick' approach due to the lack of comparable policies that differ only in this aspect, as well as the variability of the studies. Based on the relative success of the hip fracture BPT policy (England), we can speculate that 1. a 'withhold' model based on differential pricing may be optimal, 2. simplicity is preferred to more convoluted approaches that involve ranking, calculating indices and risk-adjustment; and 3. a scheme incentivising only processes of care (best practice) can achieve desired outcomes.

Regarding incentive size, we were struck by the small impact of incentives on total hospital revenue. This ranged from 0.01% HAC-POA to approximately 1.2% under the COPD BPT. The 'headline' incentive size generally overstated the overall impact because the penalties or rewards only applied to admissions for selected interventions. For example, the 4% payment adjustment under Advancing Quality (England) is twice that of the HQID project (United States) on which it is based. But because the adjustment is applied only to six interventions, the aggregate financial impact was 0.1%, while the impact of HQID was effectively zero [39, 84].

These miserly amounts not only render irrelevant discussion about the size of financial incentives (a small number multiplied by two or three is still a small number), they cast doubt over the supposed mechanism(s) that these policies are designed to elicit from hospital staff. If p4p aims to align the corporate and clinical interests within complex healthcare organisations, it is difficult to see how placing rounding-error levels of revenue at risk would drive the organisational changes needed to improve patient safety – changes that may need investments several orders of magnitude greater than those distributed under the p4p initiatives. Again, better quality evidence is needed to answer these types of questions.

Incentive size may be better seen in relative, not absolute. We saw evidence suggesting that hospitals with lower profits respond better to financial incentives [53]. This makes sense. At the economic margin, a 1% revenue reduction would mean more to a hospital operating at a loss than to one with a surplus. Conversely, Werner et al (2011) observed, albeit using only process measures, that the HQID project had a larger impact in hospitals that were in *better* financial shape and those situated in less competitive markets. They suggest that a less financially successful hospital may lack the resources to invest in quality improvement, and that without p4p, hospitals that face less competition do not have the additional incentive to improve than do those in more competitive markets [94]. This also illustrates the importance of context in designing and implementing these policies.

It was difficult to tell if rewards or penalties under the various initiatives were passed 'down' to the clinical level. Some insight on this is provided by Kristensen et al (2016 - not included in our review), who examined a Danish p4p policy that incentivised hospitals to assign case managers to their chronic care patients estimating that targeting the department rather than the hospital results in a five percentage-point improvement in performance measured by processes [95].

The evidence for design aspects is more conclusive in non-acute settings [14-16]. While these are increasingly complex and team-based, administratively they remain more fragmented and therefore better suited to enable more precise targeting of practices and even individual practitioners or practices with incentives. In fact, one of the literature reviews we included cites evidence of a negative association between practice size and response to financial incentives [12].

We saw little evidence to suggest detrimental effects of the p4p initiatives examined such as the worsening of outcomes for non-incentivised aspects of care, and promulgating health disparities based on race, ethnicity, or socio-economic status. This incidental observation contradicts the existing narrative about the negative impact of p4p, especially on minority populations [12, 96]. A dedicated review would be needed to clarify this important issue.

Implementation matters

Our findings suggests that contextual factors, co-design, and implementation of p4p are important. The literature emphasises the importance of: accounting for local context (such as financial position of the hospitals), providing adequate support and resourcing (separate to any bonuses distributed), and complementing the policy with other quality improvement initiatives (and vice versa). The most common theme, however, was careful implementation that ensures stakeholder buy-in and support from the beginning (the design phase). It may be easier for a sole practitioner or practice to change their behaviour in response to financial stimuli. It's quite another to expect complex, adaptive organisations to do so without concurrent efforts to change institutional habits and beliefs.

Several examples from our review illustrate the importance of policy consensus and ownership. Advancing Quality was associated with short-term mortality reduction [43], was assessed as cost-effective [97], and generated positive spill-over to non-incentivised conditions, patients and hospitals [44]. The literature suggests efforts were made to maximise provider engagement. For example, the initiative was altered to penalise *all* hospitals if all did not meet targets, promoting shared ownership and cooperation across organisations [16, 45]. We did not find much to suggests its 'sister' scheme HQID (under which not even a short-term effect on outcomes was observed) was implemented with such considerations [37, 70]. It is difficult to say if 'soft' mechanisms like these, as opposed to the very small financial gain, drove Advancing Quality's (modest) results, and McDonald et al (2015) found no evidence that changes in care resulting from the initiative being institutionalised, with practice "still largely reliant on prompting by particular individuals" [45]. Further research is needed.

The hip fracture BPT policy, which was the only p4p initiative for which a sustained effect on mortality and other outcomes was observed [46, 60, 62], illustrates effective implementation. Using an approach that appears to be partially modelled on the Kotter change management framework [102], the policy was introduced gradually and inclusively, with emphasis on building consensus from two key professional groups: orthopaedic surgeons and geriatricians. The stepwise implementation began with a national clinician-led audit in 2007, which led to the 'best practice' criteria and establishment of the NHFD and was associated with modest but non-significant mortality reductions. Three years after this preparatory work had been instituted the financial incentive component was introduced. This is when improvements in outcomes were observed [46]. Other studies included here highlight the importance of implementation [41, 47]. For example, Tedesco et al (2021) observed that

improvement preceded the introduction of p4p underscoring the value of good communication leading up to a policy intervention [76].

The role of change management in the success of p4p could be interpreted as further evidence for its flimsy theoretical basis [98]. Do we really think that the marginal utility of a salaried clinician working in a highly complex, team-based environment would be influenced by a very small adjustment in the revenue of their employer? Would this result in predictable behaviour change everywhere? If so, is the institutional transformation required to achieve policy objectives merely the sum of these individual behavioural changes *ceteris paribus*? Are the miserly amounts of money at stake likely to influence the behaviour of management? Given the importance of institutional culture recognised as a driver of quality improvement this should not be surprising [89, 90].

But what, then, of the p4p policies that have 'worked'? One hypothesis is that the financial aspect is a red herring, and that other mechanisms are at play. For example, p4p might motivate improvement by encouraging clinical teams to examine their own performance against their peers. Attaching a small amount of money helps by drawing attention to data. Management plays a critical role by, for example, feeding data back to clinical teams in a timely and digestible manner. The additional factors involved in implementation may thus be what determines success. After all, the evaluations of HQID failed to observe differences between public reporting plus p4p, and only public reporting. [37]. A third 'business as usual' group may have yielded observable results. Again, more targeted, high-quality research is needed to shed light on these questions.

Limitations

Our review has several limitations, the main one being the poor quality of the evidence. This is not a criticism of the included studies' authors, who in most cases did the best they could with the available data. While all reviews we included make this observation, the quality of studies examining policies aimed at hospital care is generally worse than other settings [17]. A key problem was that none of the policies relevant to our review were introduced with objective, unbiased evaluation in mind. Many were part of a suite of quality initiatives, which meant isolating the pure effect of p4p was inherently difficult. While truly randomised, experimental introduction of policies may be too much to expect [12], decision makers should aim to introduce future policy interventions in an evaluation-friendly fashion. Second, the evidence base for hospital p4p is much thinner than for non-acute settings [11, 14-16, 28, 100]. Policies targeting acute care are newer and comprise five key initiatives. More successful approaches may not have been tested yet. Third, excluding studies that did not measure clinical endpoints may have dampened our findings. However, we were interested

in outcomes, which matter most to patients, clinicians, and policy makers (notably, we identified no studies that used patient-reported measures). While structural and process measures can have instrumental value, they do not appear to manifest in health outcomes such as fewer deaths or adverse events [89]. Of the seven studies we excluded for using only process metrics only one observed an effect [101]. Fourth, excluding qualitative studies limited our capacity of this review to categorically identify factors associated with success, only the distinguishing features of successful policies (our secondary objective), although the fact that our review identified just one successful policy reduces the impact of this. Finally, publication bias may affect the evidence, with authors reluctant to submit null findings for publication, journals reluctant to publish these, or both.

A research agenda for p4p in acute care

Our findings indicate that further experimentation and research on p4p in acute care is warranted, as others have suggested [99]. We therefore propose a research agenda based on the following themes:

1. Promoting collaboration between policy makers and academic institutions in rolling out initiatives in an evaluation-friendly fashion. For example, an international collaborative or 'community of practice' to share ideas, knowledge and best practice on all aspects of p4p from design, implementation to evaluation.
2. Implementing policies in an incremental way, and attempting to adopt an experimental approach, wherever possible, particularly randomisation of participants, and stratifying design features such as incentive size, direction, and target.
3. Using a blend of process and outcome metrics in the evaluation studies. This (along with the next point) can shed light on what factors prevent/enable translating behaviours and practices into actual outcomes.
4. Using mixed methods where possible to generate important contextual information from the 'coal face' in a consistent, reproducible manner with the aim of:
    - Identifying key barriers to, and enablers of, success.
    - In cases where processes but not outcomes improve, identifying potential intrinsic or extrinsic factors that influence this translation.
5. Incorporating patient-reported outcomes into the design and evaluation of policies.
6. More concerted efforts to identify the positive and negative spillovers/consequences of p4p – original investigations as well as systematic reviews.

Conclusion

Much hope as well as hype have been placed in p4p to help overcome ongoing health policy problems. While this systematic review of p4p targeting hospital safety was inconclusive, we are hesitant to sound its death knell yet. Few system-wide hospital policies have been tried so far. Those that have contain several flaws including design complexity and a negligible impact on hospital finances. Most have not been implemented in a way that promotes ownership and engagement, embedding better practice across entire organisations. While the quality of evaluations is poor overall, the comparative success of the hip fracture BPT suggests that further experimentation and research may be warranted. Rather than a stand-alone solution, p4p in in the hospital setting may better seen as part of a suite of policies to improve patient outcomes. Future work should focus on simple and transparent design, clinician- and consumer-led development, material financial risk, and incremental implementation focusing on change management and evaluation. A research agenda is proposed to generate better quality evidence on this subject.

References

1. Institute of Medicine, *Crossing the Quality Chasm: A New Health System for the 21st Century*. 2001, Washington (DC): National Academies Press (US).

2. Padula, W.V. and P.J. Pronovost, *Improvements in Hospital Adverse Event Rates.* JAMA, 2022. **328**(2): p. 148.

3. Shrank, W.H., T.L. Rogstad, and N. Parekh, *Waste in the US Health Care System.* JAMA, 2019. **322**(15): p. 1501.

4. Runciman, W.B., et al., *CareTrack: assessing the appropriateness of health care delivery in Australia.* Med J Aust, 2012. **197**(2): p. 100-5.

5. Auraaen, A., L. Slawomirski, and N.S. Klazinga, *The economics of patient safety in primary and ambulatory care*, in *OECD Health Working Papers*. 2018, OECD: Paris.

6. Slawomirski, L. and N.S. Klazinga, *The economics of patient safety: From analysis to action*, in *OECD Health Working Papers*. 2022, OECD: Paris.

7. Braithwaite, J., et al., *Quality of Health Care for Children in Australia, 2012-2013.* Jama, 2018. **319**(11): p. 1113-1124.

8. Bates, D.W. and H. Singh, *Two Decades Since To Err Is Human: An Assessment Of Progress And Emerging Priorities In Patient Safety.* Health Affairs, 2018. **37**(11): p. 1736-1743.

9. Boudon, R., *Limitations of Rational Choice Theory.* American Journal of Sociology, 1998. **104**(3): p. 817-828.

10. Frakt, A. and A.K. Jha, *Face the Facts: We Need to Change the Way We Do Pay for Performance.* Annals of Internal Medicine, 2018. **168**(4): p. 291-292.

11. Mendelson, A., et al., *The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review.* Ann Intern Med, 2017. **166**(5): p. 341-353.

12. Markovitz, A.A. and A.M. Ryan, *Pay-for-performance: disappointing results or masked heterogeneity?* Medical Care Research and Review, 2017. **74**(1): p. 3-78.

13. OECD, *Better Ways to Pay for Health Care*. 2016, OECD Publishing: Paris.

14. Eijkenaar, F., et al., *Effects of pay for performance in health care: a systematic review of systematic reviews.* Health Policy, 2013. **110**(2-3): p. 115-30.

15. Van Herck, P., et al., *Effects, design choices, and context of P4P in health care: systematic review*.* BMC health services research, 2010. **10**(1): p. 1-13.

16. Meacock, R., S. Kristensen, and M. Sutton, *Paying for improvement: recent experience in the National Health Service.* Nordic Journal of Health Economics, 2014. **2**(1).

17. Mathes, T., et al., *Pay for performance for hospitals.* Cochrane Database of Systematic Reviews, 2019. **7**: p. CD011156.

18. Centers for Medicare and Medicaid Services, *Medicare program; changes to the hospital inpatient prospective payment systems and fiscal year 2008 rates.* Federal Register, 2007. **72**(162): p. 47129-8175.

19. Centers for Medicare and Medicaid Services. *Hospital Readmissions Reduction Program (HRRP)*. 2022 accessed 15 July 2023]; Available from: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/value-based-programs/hrrp/hospital-readmission-reduction-program

20. Centers for Medicare and Medicaid Services. *Hospital-Acquired Condition Reduction Program*. 2022 [accessed 15 July 2023]; Available from: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/hac-reduction-program

21. Girault, A., et al., *Implementing hospital pay-for-performance: Lessons learned from the French pilot program.* Health Policy, 2017. **121**(4): p. 407-417.

22. Lalloué, B., et al., *Evaluation of the effects of the French pay-for-performance program—IFAQ pilot study.* International Journal for quality in Health care, 2017. **29**(6): p. 833-837.

23. IHACPA. *Safety and quality*. 2022 [accessed 1 July 2023]; Available from: https://www.ihpa.gov.au/what-we-do/safety-and-quality

24. Runciman, B., A. Merry, and M. Walton, *Safety and Ethics in Healthcare: A Guide to Getting it Right*. 2007, London, UK: CRC Press.

25. Page, M.J., et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.* BMJ, 2021. **372**: p. n71.

26. NICE. *Appendix F Quality appraisal checklist – quantitative intervention studies*. Methods for the development of NICE public health guidance (third edition) 2023 [accessed 25 May2023]; Available from: https://www.nice.org.uk/process/pmg4/chapter/appendix-f-quality-appraisal-checklist-quantitative-intervention-studies

27. Shea, B.J., et al., *AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both.* BMJ, 2017. **358**: p. j4008.

28. Cashin, C., et al., *Paying for performance in health care: implications for health system performance and accountability*. 2014.

29. Eagar, K., J. Sansoni, and C. Loggie, *A Literature Review on Integrating Quality and Safety into Hospital Pricing Systems*. 2013, Centre for Health Service Development.

30. McDonald, R., et al., *A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs. An evaluation report commissioned by the Department of Health*. 2012.

31. Audit Commission, *Best practice tariffs and their impact*. 2012, Audit Commission for Local Authorities and the National Health Service in England.

32. AHRQ. *Healthcare Acquired Conditions (HAC) Annual Report*. 2019 [cited 2023 16 June]; Available from: https://www.ahrq.gov/hai/pfp/hacreport.html

33. Thirukumaran, C.P., et al., *Impact of Medicare's Nonpayment Program on Venous Thromboembolism Following Hip and Knee Replacements.* Health Services Research, 2018. **53(6)**: p. 4381-4402.

34. Thirukumaran, C.P., et al., *Impact of Medicare's Nonpayment Program on Hospital-acquired Conditions.* Medical Care, 2017. **55(5)**: p. 447-455.

35. Gidwani, R. and J. Bhattacharya, *CMS Reimbursement Reform and the Incidence of Hospital-Acquired Pulmonary Embolism or Deep Vein Thrombosis.* Journal of General Internal Medicine, 2015. **30**(5): p. 588-596.

36. Kim, K.M., et al., *Evaluation of Clinical and Economic Outcomes Following Implementation of a Medicare Pay-for-Performance Program for Surgical Procedures.* JAMA Network Open, 2021. **4**(8): p. e2121115.

37. Jha, A.K., et al., *The long-term effect of the Premier Hospital Quality Improvement Demonstration on patient outcomes.* New England Journal of Medicine, 2012. **366**(17): p. 1606-1615.

38. Glickman, S.W., et al., *Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction.* JAMA, 2007. **297**(21): p. 2373.

39. Ryan, A.M., *Effects of the Premier Hospital Quality Incentive Demonstration on Medicare Patient Mortality and Cost.* Health Serv Res, 2009. **44**(3): p. 821-42.

40. Shih, T., et al., *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): p. 677-681.

41. Calikoglu, S., R. Murray, and D. Feeney, *Hospital pay-for-performance programs in Maryland produced strong results, including reduced hospital-acquired conditions.* Health Affairs, 2012. **31**(12): p. 2649-58.

42. Keller, M.S., et al., *Evaluating inpatient adverse outcomes under California's Delivery System Reform Incentive Payment Program.* Health Services Research, 2021. **56**(1): p. 36-48.

43. Sutton, M., et al., *Reduced mortality with hospital pay for performance in England.* The New England journal of medicine, 2012. **367**(19): p. 1821-1828.

44. Kristensen, S.R., et al., *Long-Term Effect of Hospital Pay for Performance on Mortality in England.* New England Journal of Medicine, 2014. **371**(6): p. 540-548.

45. McDonald, R., et al., *A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West.* Health Services and Delivery Research, 2015. **3**(23): p. 1-104.

46. Metcalfe, D., et al., *Pay for performance and hip fracture outcomes: an interrupted time series and difference-in-differences analysis in England and Scotland.* The bone & joint journal, 2019. **101**(8): p. 1015-1023.

47. Alrawashdeh, M., et al., *Assessment of Federal Value-Based Incentive Programs and In-Hospital Clostridioides difficile Infection Rates.* JAMA Network Open, 2021. **4**(10): p. e2132114.

48. Arntson, E., et al., *Changes in hospital acquired conditions and mortality associated with the hospital acquired condition reduction program.* Journal of General Internal Medicine, 2019. **34(2 Supplement)**: p. S159.

49. Hsu, H.E., 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): p. 13.

50. Hsu, H.E., et al., *Association between value-based incentive programs and catheter-associated urinary tract infection rates in the critical care setting.* JAMA - Journal of the American Medical Association, 2019. **321(5)**: p. 509-511.

51. Waters, T.M., 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): p. 958.

52. Calderwood, M.S., et al., *Impact of Medicare's Payment Policy on Mediastinitis Following Coronary Artery Bypass Graft Surgery in US Hospitals.* Infection Control & Hospital Epidemiology, 2014. **35**(2): p. 144-151.

53. Calderwood, M.S., 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)**: p. 100-103.

54. He, J., et al., *Unit-level time trends and seasonality in the rate of hospital-acquired pressure ulcers in US acute care hospitals.* Research in Nursing & Health, 2013. **36**(2): p. 171-180.

55. Vaz, L.E., 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): p. 649-655.

56. Zielinski, M.D., et al., *Is the Centers for Medicare and Medicaid Service's lack of reimbursement for postoperative urinary tract infections in elderly emergency surgery patients justified?* Surgery (United States), 2014. **156(4)**: p. 1009-1017.

57. Lee, G.M., et al., *Effect of Nonpayment for Preventable Infections in U.S. Hospitals.* New England Journal of Medicine, 2012. **367**(15): p. 1428-1437.

58. Padula, W.V., et al., *Hospital-Acquired Pressure Ulcers at Academic Medical Centers in the United States, 2008-2012: Tracking Changes Since the CMS Nonpayment Policy.* Joint Commission Journal on Quality & Patient Safety, 2015. **41**(6): p. 257-63.

59. Padula, W.V., et al., *Adverse Effects of the Medicare PSI-90 Hospital Penalty System on Revenue-Neutral Hospital-Acquired Conditions.* Journal of patient safety, 2020. **16(2)**: p. e97-e102.

60. Zogg, C.K., et al., *Learning From England's Best Practice Tariff: Process Measure Pay-for-Performance Can Improve Hip Fracture Outcomes.* Ann Surg, 2022. **275**(3): p. 506-514.

61. Stone, P., et al., *Does pay-for-performance improve patient outcomes in acute exacerbation of COPD admissions?* Thorax, 2022. **77**: p. 239-246.

62. Whitaker, S.R., et al., *Does achieving the 'Best Practice Tariff' criteria for fractured neck of femur patients improve one year outcomes?* Injury, 2019. **50**(7): p. 1358-1363.

63. Stein, J.D., et al., *Use of health care claims data to study patients with ophthalmologic conditions.* Ophthalmology, 2014. **121**(5): p. 1134-41.

64. Ferver, K., B. Burton, and P. Jesilow, *The Use of Claims Data in Healthcare Research.* The Open Public Health Journal, 2009. **2**(1): p. 11-24.

65. Bae, S.H., *CMS Nonpayment Policy, Quality Improvement, and Hospital-Acquired Conditions: An Integrative Review.* Journal of Nursing Care Quality, 2017. **32**(1): p. 55-61.

66. Razi, S.S., et al., *Pay-for-performance programs reduce hospital costs without compromising care in surgical patients.* Journal of the American College of Surgeons, 2011. **1)**: p. S112.

67. Mubark, I., et al., *Mortality Following Distal Femur Fractures Versus Proximal Femur Fractures in Elderly Population: The Impact of Best Practice Tariff.* Cureus, 2020. **12**(9): p. e10744.

68. Wong, E., et al., *The impact of best practice tariff attainment on 30-day mortality and length of stay in hip fracture patients: an observational cohort study at a major trauma centre.* Orthopaedic Proceedings, 2022. **104-B**(SUPP_4): p. 13-13.

69. Tsai, Y.S., et al., *Effects of pay for performance on risk incidence of infection and of revision after total knee arthroplasty in type 2 diabetic patients: A nationwide matched cohort study.* Plos One, 2018. **13**(11): p. 18.

70. Epstein, A., A. Jha, and E. Oray, *The Impact of Pay-for-Performance on Quality of Care for Minority Patients.* The American journal of managed care, 2014. **20**: p. e479-86.

71. Rodriguez, M., et al., *Incidence of C. difficile infection Cases by CDI-Standardized Infection Ratio among Centers for Medicare and Medicaid Services-Participating Short-Term Acute Care Hospitals Post Implementation of the Hospital-Acquired Condition Reduction Program (HACRP).* Open Forum Infectious Diseases, 2021. **8(SUPPL 1)**: p. S480-S481.

72. Rude, T.L., 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**: p. 79-84.

73. Figueroa, J.F., et al., *Association between the value-based purchasing pay for performance program and patient mortality in US hospitals: observational study.* bmj, 2016. **353**.

74. Berthiaume, J.T., et al., *Aligning financial incentives with quality of care in the hospital setting.* Journal for Healthcare Quality, 2006. **28**(2): p. 36-44, 51.

75. Oakley, B., et al., *Does achieving the best practice tariff improve outcomes in hip fracture patients? An observational cohort study.* BMJ Open, 2017. **7**(2): p. e014190.

76. Tedesco, D., et al., *Improvement in Patient Safety May Precede Policy Changes: Trends in Patient Safety Indicators in the United States, 2000-2013.* Journal of patient safety, 2021. **17**(4): p. e327-e334.

77. Eldridge, N., et al., *Trends in Adverse Event Rates in Hospitalized Patients, 2010-2019.* JAMA, 2022. **328**(2): p. 173.

78. Eagar, K., et al., *A literature review on integrating quality and safety into hospital pricing systems*. 2013, Australian Health Services Research Institute: Sydney.

79. Kruse, G.B., et al., *The impact of hospital pay-for-performance on hospital and Medicare costs.* Health services research, 2012. **47**(6): p. 2118-2136.

80. Centers for Medicare and Medicaid Services. *Premier Hospital Quality Incentive Demonstration*. 2021 [accessed 17 July 2023]; Available from: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalpremier

81. McNair, P.D., H.S. Luft, and A.B. Bindman, *Medicare's policy not to pay for treating hospital-acquired conditions: the impact.* Health Affairs, 2009. **28**(5): p. 1485-93.

82. Centers for Medicare and Medicaid Services. *Hospital-Acquired Conditions (Present on Admission Indicator)*. 2021 [accessed 12 July2023]; Available from: https://www.cms.gov/medicare/medicare-fee-for-service-payment/hospitalacqcond

83. Medicare Learning Network. *Medicare Payment Systems*. 2023 [accessed 12 July2023]; Available from: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/html/medicare-payment-systems.html#Acute

84. Milstein, R. and J. Schreyoegg, *Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries.* Health Policy, 2016. **120**(10): p. 1125-1140.

85. Department of Health and Social Care. *Payment by Results 2013-14*. 2013 [accessed 12 July 2023]; Available from: https://www.gov.uk/government/collections/payment-by-results-2013-14

86. Marshall, L., A. Charlesworth, and J. Hurst, *The NHS payment system: evolving policy and emerging evidence*. 2014, Nuffield Trust.

87. NHS Digital. *Hospital Admitted Patient Care Activity*. 2023 [accessed 12 July2023]; Available from: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity

88. Digital, N. *SUS+ PbR Reference Manual*. 2019 [accessed 12 July2023]; Available from: https://digital.nhs.uk/services/secondary-uses-service-sus/payment-by-results-guidance/sus-pbr-reference-manual

89. Spaulding, A., M. Zhao, and D.R. Haley, *Value-based purchasing and hospital acquired conditions: Are we seeing improvement?* Health Policy, 2014. **118**(3): p. 413-421.

90. Royal College of Physicians. *National Hip Fracture Database*. 2023 [cited 2023 12 July]; Available from: https://www.nhfd.co.uk/20/NHFDcharts.nsf/vwCharts/BestPractice.

91. Kawai, A.T., et al., *Impact of the Centers for Medicare and Medicaid Services Hospital-Acquired Conditions Policy on Billing Rates for 2 Targeted Healthcare-Associated Infections.* Infection control and hospital epidemiology, 2015. **36(8)**: p. 871-877.

92. Rhee, C., et al., *Impact of the 2012 medicaid health care-acquired conditions policy on catheter-associated urinary tract infection and vascular catheter-associated infection billing rates.* Open Forum Infectious Diseases, 2018. **5(9)**.

93. Harris, J.M., 2nd, et al., *Comparison of Administrative Data Versus Infection Control Data in Identifying Central Line-Associated Bloodstream Infections in Children's Hospitals.* Hosp Pediatr, 2013. **3**(4): p. 307-13.

94. Werner, R.M., et al., *The effect of pay-for-performance in hospitals: lessons for quality improvement.* Health affairs, 2011. **30**(4): p. 690-698.

95. Kristensen, S.R., M. Bech, and J.T. Lauridsen, *Who to pay for performance? The choice of organisational level for hospital performance incentives.* The European Journal of Health Economics, 2016. **17**: p. 435-442.

96. Roberts, E.T., A.M. Zaslavsky, and J.M. McWilliams, *The Value-Based Payment Modifier: Program Outcomes and Implications for Disparities.* Annals of Internal Medicine, 2018. **168**(4): p. 255.

97. Meacock, R., S.R. Kristensen, and M. Sutton, *The cost-effectiveness of using financial incentives to improve provider quality: framework and application.* Health Economics, 2014. **23**(1): p. 1-13.

98. Sen, A.K., *Rational Fools: A Critique of the Behavioral Foundations of Economic Theory.* Philosophy & Public Affairs, 1977. **6**(4): p. 317-344.

99. Khullar, D., D. Wolfson, and L.P. Casalino, *Professionalism, Performance, and the Future of Physician Incentives.* JAMA, 2018. **320**(23): p. 2419.

100. Ilic, M. and L. Markovic-Denic, *Repeated prevalence studies of nosocomial infections in one university hospital in Serbia.* Turkish Journal of Medical Sciences, 2017. **47(2)**: p. 563-569.

101. Lindenauer, P.K., Remus, D., Roman, S., Rothberg, M.B., Benjamin, E.M., Ma, A., Bratzler, D.W. *Public Reporting and Pay for Performance in Hospital Quality Improvement.* New England Journal of Medicine, 2007 Vol. 356 Issue 5 Pages 486-496

102. Campbell, R.J. *Change management in health care*. The Health Care Manager, 2008 Vol. 27 Issue 1 Pages 23-39

Pay-for-performance and patient safety in acute care: a systematic review

Abstract

Pay-for-performance (p4p) has been tried across all healthcare settings to address ongoing deficiencies in the quality and outcomes of care. The evidence for the effect of these policies has been inconclusive, especially in acute care. This systematic review focused on patient safety p4p in the hospital setting. Using the PRISMA guidelines, we searched five biomedical databases for studies using at least one outcome metric from database inception to March 2023, supplemented by reference tracking and internet searches. We identified 6122 potential titles of which 53 were included: 39 original investigations, eight literature reviews and six grey literature reports. Only five system-wide p4p policies have been implemented, and the quality of evidence was low overall. About half of the studies (52%) included failed to observe improvement in outcomes, with positive findings heavily skewed towards poor quality evaluations. The exception was the Fragility Hip Fracture Best Practice Tariff (BPT) in England, where sustained improvement was observed across various evaluations. All policies had a miniscule impact on total hospital revenue. Our findings suggest the importance of simplicity and transparency in policy design, involvement of clinical community, explicit stakeholders, linked to other quality improvement initiatives, and gradual implementation. We also propose a research agenda to lift the quality of evidence in this field.

Introduction

Pay-for-performance (p4p) has proliferated in recent decades as part of the policy response to deficiencies in the quality and value of health care [1-8]. Most commonly, p4p aims to improve care quality by financially rewarding desirable structures, practices and outcomes, and/or penalising undesirable ones. P4p draws on the rational choice theory of neoclassical economics [9], linking provider utility with remuneration for desired policy outcomes more overtly than existing policy approaches, including output-based funding

Field Code Changed

models such as fee-for-service., ~~which are seen as lacking in the necessary intrinsic incentives to improve quality.~~ The approach ~~. P4p~~ seeks to align clinical risk with corporate risk, especially in larger and more complex organisations where health professionals and management may ~~have few share~~have fewer common~~d~~ objectives and incentives than smaller practices. ~~Translating the theory into real-world practice~~Demonstrating clear cut success has proven ~~difficult~~elusive, however, and over two decades of empirical research on the subject has not made a compelling case for p4p as an effective policy lever [10-12].

P4p has been used in the primary/ambulatory care setting for some time and in many countries, attempting to optimise activities ranging from vaccination to care coordination and adherence to best practice [11, 13-16]. Its deployment in the hospital setting is more recent, taking place mostly in the United States and the United Kingdom [17]. For example, a trio of initiatives introduced by the Centers for Medicare and Medicaid (CMS) in 2012 use a combination of financial rewards and penalties based on a series of process and outcome measures [18-20]. Advancing Quality, introduced in 2008 in northwestern England, incentivised performance during admissions for six hospital interventions based on 28 quality metrics [16]. ~~, and~~ Best Practice Tariffs (BPTs) were introduced in 2010 across English hospitals to ~~improve~~ incentivise adherence to best-practice care protocols for a range of conditions and interventions [16]. Elsewhere, France introduced a hospital p4p initiative based on structural and process measures in 2016 [21, 22]., ~~and a~~ A policy that withholds funding for the incremental cost of 16 ‘hospital-acquired complications’ was ~~rolled out across~~ introduced in Australian public hospitals in 2018 [23].

We present the findings of a systematic review of the literature evaluating p4p initiatives targeting patient safety in the hospital setting. Our paper adds to the evidence for two main reasons. First, we were able to identify any reviews specific to acute care safety (the only review we found focused on a specific policy in the United States [65]). Yet safety is arguably the most fundamental aspect of good health care [1, 24]. It is pre-eminent in the Institute of Medicine’s influential definition of quality, as care that is: 1. safe, 2. person-centred, 3. effective, 4. efficient, 5. timely and 6. equitable [1]. Even from a narrow perspective like a single intervention on a single patient, achieving quality is impossible without ensuring safety. Whether the same can be said about, say, efficiency can be debated. Second, the most recent review of hospital p4p we identified [17] was published before recent p4p initiatives of interest had been evaluated.

~~We focused on safety for two reasons. First, safety is arguably the most fundamental of the commonly used 6-domain definition of quality [1, 24]. Even in the narrowest context – a single intervention on a single patient – achieving quality is impossible without ensuring~~

Field Code Changed

~~safety. Whether the same can be said about other quality domains such as efficiency can be debated. Second, few systematic reviews have, to our knowledge, specifically focused on safety and the impact of p4p on this domain alone can therefore be difficult to parse. The~~Our primary ~~questions~~objective ~~of our review were: a) to what extent these~~ was to identify p4p initiatives that ~~have been effective in~~ reduc~~ed~~ing the incidence of ~~inpatient~~ iatrogenic harm in acute care.~~; and b)~~ Our secondary objective was to identify any distinguishing features of successful policies. ~~?~~ ~~Our secondary question of was whether these initiatives created unintended consequences.~~

Methods

Search protocol and sources

The reporting of this systematic review was guided by the standards of the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) Statement [25]. The protocol was registered with PROSPERO (CRD42023400087). We searched three biomedical databases (Medline, Embase and Web of Science), plus the Centre for Reviews and Dissemination (CRD) Database of Abstracts of Reviews of Effects (DARE), and the Cochrane Database, from database inception to March 2023. We included papers in any language examining p4p initiatives aimed at reducing iatrogenic harm in the hospital setting. The search strategy was developed in consultation with a University of Tasmania librarian~~Searches were conducted from database inception to March 2023~~. Search terms are provided as a <Supplementary file>.

~~We included papers in any language examining p4p initiatives aimed at safety (iatrogenic harm) in the hospital setting using any incentive structure (penalties, bonuses, or a combination).~~ Papers evaluating p4p initiatives aimed at ~~to improve~~ care quality ~~more generally~~ were included if the~~y~~ initiative (and the study evaluating it) contained at least one safety component. We included only quantitative studies that used as a dependent variable at least one outcome metric such as ~~(~~mortality, adverse event rate~~s~~, hospital-acquired condition or complication rates, and patient-reported incident- or outcome measures~~) as a dependent variable~~. ~~-~~We excluded qualitative studies and studies that measured processes only. Additional papers were identified through reference lists of included articles, reviews, editorials, and expert recommendations.

To identify grey literature, we conducted a Google search using the free text terms provided in the supplementary material, stopping the search on the fifth page (which yielded no relevant results). ~~We searched Google, and~~ We also searched ~~key~~ the websites of the

Field Code Changed

following national and international ~~institutions'~~agencies ~~websites~~; WHO, OECD, the European Commission (EC), NICE, the Australian Commission on Safety and Quality in Health Care (ACSQHC), Institute for Healthcare Improvement (IHI), Agency for Healthcare Research and Quality (AHRQ), RAND Corporation, Centers for Medicare and Medicaid Services (CMS), and the Health Quality and Safety Commission New Zealand – selected based on the authors' knowledge of the field and input from two academic experts. ~~, to identify grey literature.~~

Extraction and analysis

Search results were exported to Covidence software to assist with managing ~~, screening and selecting~~ eligible papers. Two investigators (LS and BdG) independently screened each title and abstract against the inclusion criteria, and then completed a full text review of the shortlisted papers.

One investigator (LS) abstracted data from each included paper. Results were qualitatively synthesized by location, health system context, specific p4p policy, data source and sample size, outcome metric(s), study design, follow-up period, incentive type (reward/penalty), size and target, and magnitude of effect. For effect magnitude we used relative, not absolute, change (i.e. the difference between the incidence of the respective outcome before and after the policy was introduced, expressed as a percentage). For studies that presented observations in percentage-points change in rates, we calculated the relative change from the figures provided. We categorised effect magnitude as: ~~zero~~ nil (0), low (~~1~~ up to14% relative improvement), moderate (from 15 to ~~-~~49% relative improvement), and high (50% or greater relative improvement). These cutoffs were chosen based on the left-skewed distribution of the results (the median effect magnitude was zero). All results that were not statistically significant at ~~the~~ an ~~0.05~~ alpha value ~~level~~of ~~-~~0.05 were assigned an effect of zero. We chose this threshold because it is the most used in the literature we included. To not categorially exclude results that are marginally over the chosen alpha (while those marginally *under* it are included), we flag those showing an effect of 15% or greater (i.e. Moderate and High) with a p-value between 0.05 and 0.10. We did not perform meta-analysis because all studies were observational and due to the heterogeneity of methods and metrics used.

Quality assessment

Two investigators (LS and BdG) ~~-~~independently assessed study quality using the National Institute for Health and Care Excellence (NICE) Quality Appraisal Checklist for quantitative intervention studies, assigning each paper a combined rating of 'poor', 'fair' or 'good' based on our assessment of internal and external validity [26]. Literature reviews were assessed

Field Code Changed

using the AMSTAR 2 critical appraisal tool for systematic reviews [27]. We considered the general quality of the evidence as well as the quality of indidual studies evaluating smaller, single-site initatives when forming our conclusions.

Results

~~From~~ After removing duplicates, we screened 6,122 titles and abstracts from which ~~we identified~~ 147 potentially eligible full-text titles were assessed for eligibility. We ultimately included 53 titles (Figure 1) comprising 39~~40~~ original peer-reviewed investigations (Table 1), ~~seven~~ eight literature reviews (Table 2), and six grey literature reports [13, 28-32]. The earliest study was published in 2006, the latest in 2023. All but two papers were published since 2012, with 17 published ~~during~~ in the last five years (14 original investigations, one literature review and two grey literature titles).

Figure 1 here

Study design, methods and quality

Among the original investigations, the two most common study designs were Difference-in-Difference (n=14) [33-46], and Interrupted Time Series (n=13) analysis [47-59]. One paper used both [60]. The mean (median) follow-up time was 3.6 (3.2) years. The longest follow-up was eight years [36], the shortest eight months [61]. We saw no evidence of an association between the length of follow-up and observed effect. Most investigations used ~~a~~ claims~~/administrative~~ data (n=23~~4~~, 59~~60~~%), with the remainder using infection surveillance and registry data. One used both claims and surveillance data [52].

The most common outcome measures were mortality, and ~~some variation of~~ adverse event rates or hospital-acquired condition (HAC) rates. Thirteen original investigations analysed more than one health outcome, or the same outcome over ~~two~~ multiple follow-up periods. The 39 original investigations ~~40 studies~~ thus yielded 5~~4~~ ~~*de facto*~~ results (henceforth ~~also~~ referred to as 'studies'). One study used patients' mobility status and residential status in addition to mortality [62]. We found no studies using patient-reported measures. The most common comparison groups were either hospitals not participating in the p4p policy, or diagnoses, interventions or patient groups that were not incentivised under the policy but were managed in participating hospitals.

~~Regarding study quality,~~ Nineteen~~we assessed 19~~ (49%) of the original investigations were assessed as 'poor' quality, 17 (44%) as 'fair', and ~~four~~ three (8%) as 'good'. The most

Field Code Changed

common issues were: ~~the~~ lack of an appropriate comparator, insufficient observation time before and/or after introduction of the policy, lack of record-level risk adjustment, and the use of ~~administrative/~~ claims data, which are not considered as reliable as other data sources for~~in this research context~~deriving outcomes other than mortality [63, 64]. ~~Original investigations of *lower* quality were *more* likely to show an effect and that their effect was typically of greater magnitude compared to studies of higher quality, especially studies using claims data (see below).~~

The seven literature reviews comprised one Cochrane, one integrative, three~~wo~~ literature, and three systematic reviews. We assessed the quality of all but one of these as 'good' ~~quality (see~~ Table 2). Three~~wo~~ reviews focused on the hospital setting~~.~~ ~~with~~ O~~o~~ne of these focused on a single p4p policy targeting safety (the CMS' HAC-POA initiative) while the other two included p4p targeting quality of care more broadly ~~focusing on a policy targeting patient safety~~ [11, 12, 14-17, 65, 84]. ~~We assessed all but one as 'good' quality (see Table 2).~~

Location of studies and initiatives

Twenty-eight (52%) ~~Most~~ of the original investigations were conducted in the United States ~~(n=29 (80%))~~, with all but three papers examining the three main p4p initiatives of the Centers for Medicare and Medicaid (CMS) ~~p4p initiatives~~: ~~the~~1. Premier Hospital Quality Improvement Demonstration (HQID) project; 2. HAC Present on Admission (HAC-POA) policy [18]); and 3. ~~the~~ HAC Reduction Program (HACRP) [20] ~~(see Figure 2)~~. Three papers ~~studies~~ examined the effects of local initiatives in California, Hawaii, and Maryland [41, 42, 66]. Ten papers were ~~conducted~~ from~~in~~ England, focusing on the Advancing Quality initiative [43-45], and the Best Practice Tariff (BPT) policy for hip fracture care [45, 46, 60, 62, 67, 68] and chronic obstructive pulmonary disease (COPD) [61]. The remaining study was conducted in Taiwan [69]. Descriptions of the main English and US initiatives are provided in the <Supplementary material>.

~~*Figure 2 here*~~

Overall findings

Key findings from the original investigations are summarised in Table 1. ~~Thirty~~ Twenty-eight studies (5~~6~~2%) failed to observe a statistically significant effect [36-40, 42, 44, 45, 48-53, 55, 57, 61, 62, 69, 71-75] (~~)~~ one of these observed a moderate effect with a p-value of 0.077 [72].) Eight studies (15%) observed a 'low' ~~-~~magnitude effect [36, 43-46, 62, 67, 68], ~~twelve~~ thirteen (2~~4~~4%) a 'moderate' effect [33, 35, 41, 47, 48, 53, 54, 56, 59, 60, 62, 69, 76], and ~~four~~ five (9~~7~~%) a 'high' effect [34, 36, 52, 58, 75]. None of the latter came from 'good' quality

Field Code Changed

studies (three were assessed as 'poor', two as 'fair'). None of the included studies~~no~~ observed a deleterious effect on incentivised or non-incentivised outcomes. Of the 22 studies using mortality as an outcome, nine (4~~1~~0%) observed an effect and ~~— all~~ all were evaluations of English ~~policies~~ initiatives [43-46, 60, 62, 67, 68, 75]. Sixteen (50%) o~~O~~f the remaining studies using non-mortality outcomes~~, 16 (52%)~~ observed an effect [33-36, 41, 47, 48, 52, 53, 54, 56, 58, 59, 62, 69, 76], with ~~-~~11 of these (69%) using claims data observed a moderate or high effect. All but one of the five studies observing an effect of 50% or greater relied on claims data.

Papers of lower quality were more likely to observe an effect, and the effect was typically of greater magnitude than in studies of higher quality. No studies rated as 'good' observed a result of moderate or high magnitude, while two fifths of studies rated 'fair' and 'poor' did.

Our initial search identified seven studies using only process measures. These were excluded, but it is worth noting that only one of them observed an effect – a modest improvement in quality processes at hospitals participating in the HQID project [101].

Table 1 here

United States initiatives

Of the studies based in the United States~~United States studies~~, 2~~1~~3 (58~~64~~%) did not observe an effect, ~~three~~ one (3~~8~~%) observed a low~~-magnitude effect~~, ~~seven~~ ten (28~~19~~%) a moderate, and ~~three~~ four (11~~8~~%) a high magnitude effect. One~~ None of the~~ stud~~ies~~ using mortality as the outcome~~s~~ observed a low magnitude~~n~~ effect [42]. Concerning the three main p4p policies in the United States ~~((~~see Supplementary material)~~Figure 2)~~, 11~~10~~ of the 1~~6~~7 HAC-POA studies observed an ~~favourable~~ effect [33-36, 52, 53, 54, 56, 58, 59, 76], with eight of the papers producing these studies rated as 'poor' and the remaining five as 'fair'. The HACRP initiative was observed to have an effect in just two of the 10 studies [47, 48] although one observed a moderate effect with a p-value of 0.077. Of the eight papers producing these results, six were rated 'fair' and two 'poor'. None of the five ~~six~~ studies evaluating the patient safety component of the HQID project observed an effect [37-40~~, 70~~], although all but one of these used risk-adjusted mortality as the outcome, compared to only a minority of studies evaluating the HAC-POA and HACRP policies~~-~~. ~~Tedesco et al (2021) examined the longest period of the original investigations we included (2000-13), finding reductions in the rates of 13 adverse events ranging from -31% (central-line blood stream infections) to -3% (postoperative respiratory failure) [76].~~ Of the three studies evaluating other, smaller initiatives in the United States, the only one to observe an effect was Calikoglu et al (2012) who, ~~(~~using ~~administrative~~ claims data over a relatively short, 2-year follow-up period,~~)~~ ~~estimated~~ observed an average 15% reduction~~, on average,~~ in HACs included,

Field Code Changed

compared to HACs not included~~,~~ in a Maryland p4p policy ~~over a 2-year follow-up period~~ [41].

The grey literature included a patient safety 'scorecard' ~~on patient safety~~ for admissions under ~~in~~ Medicare ~~admissions~~ from~~by~~ the Agency for Healthcare Research and Quality (AHRQ), which reported a 17% reduction in the crude rates of 28 hospital-acquired conditions between 2010 and 2014 and a further 13% between 2014 and 2019 ~~in the United States~~ [32]. Although these results cover the period spanning the HAC-POA and the HACRP initiatives and are reflected in some of the original investigations we included [77], the AHRQ scorecard report is not intended to serve as a formal evaluation of these policies. *Inter alia,* it lacks comparators ~~outcomes, hospitals or patients,~~ nor does it attempt to separate the p4p component from other policies and activities aimed at reducing hospital-acquired conditions over this period [6].

Initiatives in England

The papers examining English ~~p4p~~ initiatives were generally of higher quality than the others. All used mortality as a ~~t least one outcome~~ dependent variable. Two papers on the Advancing Quality initiative assessed effects over distinct time periods, observing small but significant ~~reductions in mortality~~effect in the short term (up to 18 months post-intervention) but not thereafter (42 months) [44, 45]. One of these observed evidence for a possible positive spillover effect to conditions not covered by this initiative [44].

Six papers examined the hip fracture BPT policy. Two ~~of these~~ were nation-wide studies. Metcalfe et al (2019) compared English hospitals (included in the scheme) with Scottish counterparts (not included) observing a 6% relative mortality reduction during a 7-year follow-up, as well as reductions in readmissions and average length of stay [46]. Zogg et al (2022) compared England with the United States, observing a 27% relative reduction in ~~risk-adjusted 365-day~~ mortality during a 6-year follow-up [60]. The rest were single-site studies observing ~~risk-adjusted~~ mortality reductions ranging from 8% to 71% for cases satisfying the BPT criteria compared to cases that did not ~~(~~-the latter result stemming from a 'poor' quality paper) [62, 67, 68, 75]~~.~~

Stone et al (2019) examined the COPD BPT using a nation-wide cohort but observing no effect on either mortality or readmissions [61]. We failed to identify peer-reviewed evaluations of other BPTs relevant to this review, such as stroke care, which appeared only in two grey literature reports [30, 31]. These were published within two years of its implementation and suggest that the policy had no impact on outcomes beyond what was achieved by other national stroke care initiatives, with ~~. T~~the complexity of that specific ~~e tariff~~BPT variant was ~~identified~~ cited as a key reason [31].

Field Code Changed

Taiwan

The Taiwanese study evaluated a nation-wide p4p initiative targeting diabetes care, examining surgical-site infections and revisions following total knee arthroplasty in diabetic patients. It failed to observe any effect on infection rates in patients included in the initiative but observed a significant reduction in the risk of surgical revision [69]. Evidence cited in the grey literature suggests that this initiative wasis prone to excluding more complex patients [78].

Reviews and grey literature

Findings from the seven literature reviews are summarised in Table 2. All failed to find conclusive evidence that financial incentives aimed at hospital care have a sustained, positive effect on patient safety outcomes in acute care(more success is reported non-acute care, however). Effects were small at best, with little evidence that any gains were sustained over time. Low quality of the evidence was consistently remarked upon. The The grey literature we included drew similar conclusions, especially for hospital-based initiatives, finding the evidence to be of low quality and any effect on outcomes at best small, unsustained and difficult to separate from other, concurrent quality improvement initiatives [13, 28, 29]. The Both the literature reviews and the grey literature highlighted the importance of context, design and especially implementation in the success of hospital p4p initiatives tested so far with several. Many commenting that the way the various policies were introduced has made itcreated methodologically challenges to studying their effects [11, 12, 14-17, 65].

Table 2 here

P4p policy design features: incentive size, direction and target

The evidence of an association between effect and design features such as incentive size was inconclusive, mostly due to an insufficient number of high-quality, uniform comparable evaluations of policies that share a similar design but different incentive sizes,designs, let let alone head-to-head comparisons of different p4p incentive designsmagnitude and direction (penalty/reward).

The impact of the main p4p policies on the total revenue of participating hospitals was:

- HQID (United States): less than 0.01% [39, 79, 80]
- HAC-POA (United States): less than 0.01% [81, 82]
- HACRP (United States): 1% [20, 83]

Field Code Changed

- Advancing Quality (England): 0.1% [84]
- BPT (England): [85, 86]
    - Hip fracture care: 1% [31, 87]
    - Acute stroke care: 0.2% [31, 87]
    - Acute COPD exacerbation: 1.2% [61, 87, 88]

We did, however, see evidence for an association between incentive size and effect in ~~studies that stratified~~ hospitals stratified according to financial strength. Calderwood et al (2016) observed a decline in HAC rates at hospitals operating at a financial loss, but not in hospitals with positive operating margins, following introduction of the HAC-POA policy [53].

We did not see any consistent associations between effect and the direction (penalty vs. reward) of the incentive. ~~The h~~Hospital management and administration was the reported target in all our included studies, and it was unclear if the incentives were 'passed ~~'~~down' to the departments, teams, or individuals responsible for attracting the reward or penalty. We are therefore unable to draw any conclusions about this.

While not an objective of the review, ~~W~~we saw little evidence~~-~~ for negative unintended consequences on patients who were classified into a minority group based on race or ethnicity, in hospitals serving these minority populations, or on non-incentivised conditions or interventions [38, 49, 55, 66, 70]. ~~. Hsu et al (2020) and Vaz et al (2015) failed to observe any disparity between safety net and non-safety net hospitals in association with the HACRP program [49, 55]. Glickman et al (2007) failed to detect any evidence that the HQID project had an adverse association on care not subject to financial incentives [38]. Epstein et al (2014) observed greater reduction in mortality among black patients than white patients associated with the HQID [70]. None of the above studies observed an association between respective policy and aggregate safety outcomes.~~

Two studies investigating the Advancing Quality Initiative in England ~~observed~~found evidence for a *positive* spillover effect on~~to~~ outcomes ~~in from admissions~~ for ~~several~~ non-incentivised interventions at participating hospitals ~~(including acute renal failure and intracranial injury) and~~as well as outcomes at non-participating hospitals ~~in other (non-participating) regions~~ [44, 45]. Positive spillover was also noted in one literature review [14].

Context and implementation

The importance of local context and policy implementation was ~~stressed~~ a common theme throughout ~~all~~ the literature we reviewed, in particular: ~~. Three common themes were:~~ a) ensuring stakeholder acceptance and buy-in, b) providing support and resourcing where

Field Code Changed

needed, and c) tailoring the design to local context [15, 16, 41, 45, 47]. For example, Alrawashdeh et al (2021), who observed a 6% decline in hospital-acquired *clostridium difficile* infection rates associated with ~~CMS'~~ the HACRP initiative, ~~speculated~~ argued that the decline "may be due to greater opportunities for improvement in diagnostic and antimicrobial stewardship, which also align with the value these programs intend to achieve" [47]. Calicoglu et al (2012) examined a Maryland-specific p4p initiative based on the nation-wide HACRP policy, observing a considerably greater effect than the studies examining the other United States p4p initiatives (although using claims data), and ~~. They s~~suggesting more effort on change management and acceptance of the policy among providers as potential reasons for the comparative success [41]. In some cases, improvements in safety began *before* the introduction of a p4p initiative, lending weight to the importance of stakeholder communication and change management in policy implementation [76].

~~*Unintended consequences*~~

~~We saw little evidence for negative unintended consequences on patients who were classified into a minority group based on race or ethnicity, hospitals serving these minority populations, or on non-incentivised conditions or interventions. Hsu et al (2020) and Vaz et al (2015) failed to observe any disparity between safety net and non-safety net hospitals in association with the HACRP program [49, 55]. Glickman et al (2007) failed to detect any evidence that the HQID project had an adverse association on care not subject to financial incentives [38]. Epstein et al (2014) observed greater reduction in mortality among black patients than white patients associated with the HQID [70]. None of the above studies observed an association between respective policy and aggregate safety outcomes.~~

~~Two studies investigating the Advancing Quality Initiative in England observed a positive spillover effect to outcomes from admissions for several non-incentivised interventions at participating hospitals (including acute renal failure and intracranial injury) and outcomes at hospitals in other (non-participating) regions [44, 45]. Positive spillover was also noted in one literature review [14].~~

~~*C*~~*The c*laims data *effect* ~~*and outcomes*~~

Studies using HAC or adverse event rates as their outcome were twice as likely to observe an effect if the rates were derived from ~~administrative/~~claims data compared to other data sources, with a four-fold greater effect (22% versus 5%) observed, on average. While the limitations of these data in the research context are known [63, 64], t~~This~~his finding suggests a potential association between p4p and changes in administrative coding. The most striking example is Calderwood et al (2014) who used both Medicare billing and National Healthcare Safety Network (NHSN) infection surveillance data on 638,761 procedures at 1,234 hospitals

Field Code Changed

to examine the effect of the HAC-POA policy on mediastinitis following coronary artery bypass grafting (CABG). They observed a sudden, 64% reduction in billing for mediastinitis at the introduction of the policy but no corresponding change in rates calculated from NHSN data [52].

Discussion

We present findings from a comprehensive systematic review of p4p ~~policies~~ targeting iatrogenic harm in the hospital setting. Overall, we found little conclusive evidence for a sustained positive effect of such policies on patient safety ~~of hospital p4p policies that have been introduced and evaluated up to now~~. To our knowledge, this is the first review of ~~of~~ p4p policies and safety in acute care ~~focusing on hospital safety.~~ ~~and our~~Our conclusions generally align with ~~with literature~~ reviews that ~~covering~~ examined ~~the topic~~p4p more broadly~~, particularly those of hospital p4p~~ [11, 12, 14-17, 65, 84]. However, there was one~~. Excluding studies that did not use outcome measures as an independent variable may have dampened our findings. However, we were interested in clinical endpoints. While structural and process measures can have instrumental value, they appear not to manifest in health outcomes such as fewer deaths or adverse events [89], which, in our opinion, matter most to patients, clinicians, and policy makers (we found no studies that used patient-reported measures).~~

~~The~~ exception~~: was~~ the fragility hip fracture BPT ~~in England~~, which ~~, which~~ was associated with a significant reduction in risk-adjusted mortality, as well as ~~a~~ reductions in length of stay, patients' mobility status, and readmissions up to seven years post-implementation [46, 60, 62, 67, 68, 75]. Three features of this policy ~~that~~ set it apart from others: it is ~~simpl~~straightforward~~icity~~, it was not imposed, with ~~;~~key aspects of the ~~-~~development led by stakeholders,~~inclusive development,~~ and it was implemented gradually ~~implementation~~. The latter is~~,~~ exemplified through the explicit ~~policy's~~ link to the National Hip Fracture Database (NHFD), ~~[90] —~~a public registry that publishes ~~showing~~ hospital-level results and was ~~that was~~ introduced following a clinician-led audit in 2007 ~~– ,~~ three years before the ~~tariff~~ p4p component began to be applied [90].

Study quality ~~*Choice of data*~~

The quality of studies appeared to be associated with the result. Observed effects varied considerably, even between studies examining the same p4p policies. ~~Overall, s~~ Studies using mortality as ~~their~~ an outcome variable were ~~much~~ less likely to observe an effect than

Field Code Changed

those using adverse event or HAC rates. ~~, perhaps~~ This may reflect differences in the~~ing~~ ~~the lower~~ sensitivity of ~~mortality~~ these~~as a~~ metrics, although while one may expect a lower effect magnitude, the large datasets in most of the included studies should aid the detection of small effects. ~~However , even a~~Among the studies using adverse event or complication rates, those that drew on ~~administrative or~~ claims data were twice as likely to observe an effect than those that used other data sources, averaging a four-fold greater effect magnitude. This was especially pronounced where the study used the same data ~~metrics~~ as those that were used to assess performance ~~by~~ by the ~~p4p~~ policy [33-35, 52, 58, 91, 92].

This is perhaps unsurprising. ~~Billing~~ Claims data have low sensitivity and do not correlate ~~well~~ with clinically meaningful outcomes [52]. They are also less likely to capture actual occurrence of events and may be subject to reporting bias to improve performance scores [93]. Kawai et al (2015) examined billing patterns at 569 hospitals before and after the introduction of the HAC-POA policy, focusing on two of the incentivised conditions (central line associated blood stream infection (CLABSI) and catheter-associated urinary tract infections (CAUTI)) and non-incentivised surgical site infections (SSI). For the incentivised outcomes, after a steady rise before the policy was introduced (on average by 17% and 19% per quarter respectively), billing rates then dropped by 25% upon introduction followed by a slightly downward trend for CLABSI and a levelling off for CAUTI during the follow-up period. No discernible change in SSI rates was observed before or after the policy's introduction [91]. Rhee et al (2019) applied the same method to the subsequent introduction of the HAC-POA policy for *Medicaid* patients. They found little impact on billing rates at 546 hospitals for incentivised *and* non-incentivised conditions following its introduction compared to the pre-policy trajectory. CAUTI billing did not change, while CLABSI billing reduced significantly during the pre- and post-introduction periods [92]. We would therefore question not only the validity of studies that rely on these data, but also the use of claims data in p4p policies.

Nevertheless, the low quality of the evidence undermines our findings. While we note an inverse association between quality and observed effect, it is difficult to say if better studies may have observed an effect, or that better quality may not have dampened the effects observed in some of the studies included here. We therefore propose a research agenda for this policy area at the end of the paper.

Design characteristics

We found little evidence to inform p4p policy design recommendations and ~~. We~~ are unable to recommend a 'carrot' or 'stick' approach due to the lack of comparable policies that differ only in this aspect, as well as the variability of the studies. Based on the relative success of the hip fracture BPT policy (England), we can speculate that 1. a 'withhold' model based on

Field Code Changed

differential pricing may be optimal, 2. simplicity is preferred to more convoluted approaches that involve ranking, calculating indices and risk-adjustment; and 3. a scheme incentivising ~~recommended~~ only processes of care (best practice) can achieve desired outcomes.

Regarding incentive size, we were struck by the small ~~aggregate~~ impact of incentives ~~on total~~ on total hospital revenue. This ranged from 0.01% HAC-POA to approximately 1.2% under the COPD BPT. ~~In most cases t~~The 'headline' incentive size generally overstated the overall impact because the penalties or rewards only applied~~y~~ to admissions for selected interventions. For example, the 4% payment adjustment under Advancing Quality (England) is twice that of the HQID project (United States) on which it is based. But because the adjustment is applied only to six interventions, the aggregate financial impact was 0.1%, while ~~(~~the impact of HQID ~~impact~~ was effectively zero~~)~~ [39, 84].

These miserly amounts not only render irrelevant discussion about the size of financial incentives (a small number ~~doubled~~ multiplied by two or three is still a small number), ~~but~~ they cast doubt over the supposed ~~precise~~ mechanism(s) that these policies are ~~supposed~~ designed to elicit from hospital staff ~~management and clinical teams~~. If p4p aims to align the corporate and clinical interests within complex healthcare organisations, it is difficult to see how placing rounding-error levels of revenue at risk would drive the organisational changes needed to improve patient safety – changes that may need investments several orders of magnitude greater than those distributed under the p4p initiatives. ~~As Markovitz and Ryan (2017) observed, there is indeed "something more fundamentally amiss with p4p", [12]~~ Again, better quality evidence is needed to answer these types of questions.~~.~~

Incentive size may be better seen in relative, not absolute. We saw ~~, terms with some~~ evidence suggesting that hospitals with lower profits respond better to financial incentives [53]. This makes sense. At the economic margin, a 1% revenue reduction would mean more to a hospital operating at a loss than to one with a surplus. Conversely, Werner et al (2011) observed, albeit using only process measures, that the HQID project had a larger impact in hospitals that were in *better* financial shape and those situated~~/or~~ in less competitive markets. They suggest that a less financially successful hospital may lack the resources to invest in quality improvement, and that without p4p, hospitals that face less competition do not have the additional incentive to improve than do those in more competitive markets ~~, illustrating the importance of context in designing these policies and their implementation~~ [94]. This also illustrates the importance of context in designing and implementing these policies.

It was difficult to ~~glean~~ tell if ~~hospitals passed~~ rewards or penalties under the various initiatives were passed 'down' to the ~~organisational~~ clinical level ~~where they were earned~~.

Field Code Changed

Some insight on this is provided by Kristensen et al (2016 - not included in our review),) who examined a Danish p4p policy that incentivised hospitals to assign case managers to their chronic care patients (the paper was not included in our review), estimating that targeting the department rather than the hospital results in a five percentage-point lift improvement in performance measured by processes [95].

The evidence for design aspects such as incentive size, direction and target is more conclusive in non-acute settings [14-16]. While these Non-acute care is are increasingly complex and team-based, but administratively they these settingsremain are more fragmented and therefore better suited to enable more precise targeting of practices and even individual practitioners or practices with payments or penaltiesincentives. In fact, oOne of the literature reviews we included cites evidence of a negative association between practice size and response to financial incentives [12].

Unintended consequences

Our review uncovered We saw little evidence to suggest detrimental effects of the p4p initiatives examined such as, including the worsening of outcomes for non-incentivised conditions and proceduresaspects of care, and promulgating health disparities based on race, ethnicity, or socio-economic status. Several studies we included or screened examined negative consequences of hospital p4p. This incidental observation contradicts the existing narrative about the negative impact of p4p, especially on minority populations [12, 96]None found convincing evidence for detrimental outcomes, with some observing potential positive spillovers [38, 44, 49, 55, 66, 70]. As with all null findings, absence of evidence is not evidence of absence, and potentially reflects deficiencies in study methods, data sources and publication bias. However, our findings contradict the existing narrative about the negative impact of p4p, especially on minority populations [12, 96]. A dedicated review would be needed to clarify this important issue.

Implementation matters: the 'how' is as important as the 'what'

Our review findings suggests that contextual factors, as well as co-design, and implementation of p4p are important determinants of success. The . The following factors are emphasised in the literature emphasises the importance of: accounting for local context (such as financial position of the hospitals), providing adequate support and resourcing (separate to any bonuses distributed), and complementing the policy with other quality improvement initiatives (and vice versa). The most common theme, however,, however, was careful implementation that ensures stakeholder buy-in and support from the beginning (at the design phase) of the policy. It may be This makes sense, too. It is one thing for aeasier for a sole practitioner or small practice to change theiritsbehaviour in response to financial

Field Code Changed

stimuli. It's quite another to expect complex, adaptive organisations to do so without concurrent efforts to change institutional habits and beliefs.

Several examples from our review illustrate the importanceance of policy consensus, acceptance, and ownership of the policy. Advancing Quality was associated with short-term mortality reduction [43], was assessed as cost-effective [97], and generated positive spill-over to non-incentivised conditions, patients and hospitals in the longer term [44]. The lLiterature we saw suggests efforts were made to maximise provider engagement. For example, the initiative was altered to penalise *all* hospitals if all did not meet targets, promoting shared ownership and cooperation across organisations [16, 45]. We did not find much to suggests its 'sister' scheme HQID (under for which not even a short-term effect on outcomes was observed) was implemented with such considerations [37, 70]. IttIt is's difficult to say if 'soft' mechanisms like these, as opposed to the very small financial gain, drove Advancing Quality's (modest) results, and , although McDonald et al (2015) found no evidence that changes in care resulting from the initiative being institutionalised, with practice "still largely reliant on prompting by particular individuals" [45]. Further research is needed.

Perhaps the best illustration of good implementation was, again, tThe hip fracture BPT policy, which showed was the strongest only p4p initiative for which a and most sustained effect on- mortality and (as well as other outcomes was observed) in our review [46, 60, 62], illustrates effective implementation. Using an approach that appears to be partially modelled on the Kotter change management framework [102], t. Thei policy was introduced gradually and inclusively, with emphasis on building consensus from two key professional groups: orthopaedic surgeons and geriatricians. The stepwise implementation began with a national clinician-led audit in 2007, which led to the 'best practice' criteria and establishment ofed the NHFD and . This was associated with modest but non-significant mortality reductions. Only tThree years later, after this preparatory work had been instituted, was the financial incentive component was introduced. This is when , at which pointthe improvedimprovements in outcomes mortality dropped significantlywere observed [46]. The United StatesOther studies included here also highlight the importance of implementation [41, 47]. For example, Tedesco et al (2021) observed that improvement preceded the introduction of p4p underscoring the value of good communication leading up to a policy intervention [76].

The critical role of change management in the success of p4p could be interpreted as further evidence for its flimsy theoretical basis [98]. Do we really think that the marginal utility of a salaried clinician working in a highly complex, team-based environment would be influenced by a very small adjustment in the revenue of their employer? , and Would this result in

Field Code Changed

predictable behaviour change everywhere? If so, is the institutional transformation required to achieve policy objectives merely ~~improvement in outcomes~~ the sum of these individual behavioural changes *ceteris paribus*? Are the miserly amounts of money at stake likely to influence the behaviour of management? Given the importance ~~It's unsurprising to see~~ of institutional culture recognised as a driver of quality improvement this should not be surprising [89, 90].

But what, then, of the p4p policies that have 'worked'? One hypothesis is that the financial aspect is a red herring, and that other mechanisms are at play. For example, p4p might motivate improvement by encouraging clinical teams to examine their own performance against their peers. Attaching a small amount of money helps by drawing attention to data. Management plays a critical role by, for example, feeding data back to clinical teams in a timely and digestible manner. The additional factors involved in ~~i~~implementation may thus be what determines success ~~turns on~~. ~~(a~~After all, the evaluations of HQID failed to observe ~~a~~ differences between ~~p4p with~~ public reporting plus p4p, and only public reporting. ~~only~~ [37]~~)~~. A third 'business as usual' group may have yielded observable results. Again, more targeted, high-quality research is needed to shed light on these questions.

~~We would agree that there is scope for further experimentation and research [99].~~

Limitations

Our review has several limitations, the main~~A key limitation of our findings is~~one being the poor quality of the evidence. This is not a criticism of the included studies' authors, who in most cases did the best they could with the available data. While all reviews we included make this observation, the quality of studies examining policies aimed at hospital care is generally worse than other settings [17]. A key problem was that none of the policies relevant to our review were introduced with objective, unbiased evaluation in mind. Many were part of a suite of quality initiatives, which meant isolating the pure effect of p4p was inherently difficult. While truly randomised, experimental introduction of policies may be too much to expect [12], decision makers should aim to introduce future policy interventions in an evaluation-friendly fashion. Second, ~~that they are based on relatively few examples. P4p policies targeting acute care are newer and comprise five key initiatives. t~~The evidence base for ~~is much thinner~~hospital p4p is much thinner than ~~that~~ for non-acute settings ~~p4p~~ [11, 14-16, 28, 100]. Policies targeting acute care are newer and comprise five key initiatives. More successful approaches may not have been tested yet. Third, ~~The existing evidence is generally lacking in quality. While all literature reviews of p4p make this observation, the quality of studies examining policies aimed at hospital care is generally worse than in other settings [17]. This is not a criticism of the authors – most of whom did the best they could~~

Field Code Changed

~~with the available data. A key problem was that none of the policies relevant to our review were introduced with objective, unbiased evaluation in mind. Many were part of a suite of quality initiatives, which made studying the pure effect of p4p difficult. While truly randomised, experimental introduction of policies may be too much to expect [12], decision makers should aim to introduce future policy interventions in an evaluation-friendly fashion. Third, we only included studies using clinical endpoints. This may have dampened our findings, although we would again make the point that health outcomes are what matters most.~~ excluding studies that did not measure clinical endpoints may have dampened our findings. However, we were interested in outcomes, which matter most to patients, clinicians, and policy makers (notably, we identified no studies that used patient-reported measures). While structural and process measures can have instrumental value, they do not appear to manifest in health outcomes such as fewer deaths or adverse events [89]. Of the seven studies we excluded for using only process metrics only one observed an effect [101]. Fourth, excluding qualitative studies limited our capacity of this review to categorically identify factors associated with success, only the distinguishing features of successful policies (our secondary objective), although the fact that our review identified just one successful policy reduces the impact of this. Finally, publication bias may affect the evidence, ~~base. This may be a factor in the sparse literature covering BPTs other than hip fracture care,~~ with authors reluctant to submit null findings for publication, journals reluctant to publish these, or both. ~~If this is the case, the evidence for p4p is even weaker than our findings suggest.~~

A research agenda for p4p in acute care

Our findings indicate that further experimentation and research on p4p in acute care is warranted, as others have suggested [99]. We therefore propose a research agenda based on the following themes:

1. Promoting collaboration between policy makers and academic institutions in rolling out initiatives in an evaluation-friendly fashion. For example, an international collaborative or 'community of practice' to share ideas, knowledge and best practice on all aspects of p4p from design, implementation to evaluation.
2. Implementing policies in an incremental way, and attempting to adopt an experimental approach, wherever possible, particularly randomisation of participants, and stratifying design features such as incentive size, direction, and target.
3. Using a blend of process and outcome metrics in the evaluation studies. This (along with the next point) can shed light on what factors prevent/enable translating behaviours and practices into actual outcomes.

Field Code Changed

4. Using mixed methods where possible to generate important contextual information from the 'coal face' in a consistent, reproducible manner with the aim of:
    - Identifying key barriers to, and enablers of, success.
    - In cases where processes but not outcomes improve, identifying potential intrinsic or extrinsic factors that influence this translation.
5. Incorporating patient-reported outcomes into the design and evaluation of policies.
6. More concerted efforts to identify the positive and negative spillovers/consequences of p4p – original investigations as well as systematic reviews.

Conclusion

Much hope as well as hype ~~was~~ have been placed in p4p to help overcome ~~to solve~~ ongoing health policy ~~challenges~~problems ~~with health policy and practice~~. While this systematic review of p4p targeting hospital safety was inconclusive, we are hesitant to sound ~~the~~ its death knell yet. Few system-wide hospital policies have been tried so far. Those that have contain several flaws including design complexity and a negligible impact on hospital finances. Most have not been implemented in a way that promotes ownership and engagement, embedding better practice across entire organisations. While the quality of evaluations is poor overall, ~~T~~the comparative success of the hip fracture BPT suggests that further experimentation and research may be warranted. Rather than a stand-alone solution, p4p in in the hospital setting may~~is~~ better seen as part of a suite of policies to improve ~~quality and~~ patient outcomes ~~of care~~. Future work ~~p4p~~ should focus on simple and transparent design, clinician- and consumer-led development, material financial risk, and incremental implementation focusing on change management and evaluation. ~~A as well as~~ concerted research agenda is proposed to generate better quality evidence on this subject. ~~sound evaluation.~~

References

1. Institute of Medicine, *Crossing the Quality Chasm: A New Health System for the 21st Century*. 2001, Washington (DC): National Academies Press (US).

Field Code Changed

2. Padula, W.V. and P.J. Pronovost, *Improvements in Hospital Adverse Event Rates.* JAMA, 2022. **328**(2): p. 148.

3. Shrank, W.H., T.L. Rogstad, and N. Parekh, *Waste in the US Health Care System.* JAMA, 2019. **322**(15): p. 1501.

4. Runciman, W.B., et al., *CareTrack: assessing the appropriateness of health care delivery in Australia.* Med J Aust, 2012. **197**(2): p. 100-5.

5. Auraaen, A., L. Slawomirski, and N.S. Klazinga, *The economics of patient safety in primary and ambulatory care*, in *OECD Health Working Papers*. 2018, OECD: Paris.

6. Slawomirski, L. and N.S. Klazinga, *The economics of patient safety: From analysis to action*, in *OECD Health Working Papers*. 2022, OECD: Paris.

7. Braithwaite, J., et al., *Quality of Health Care for Children in Australia, 2012-2013.* Jama, 2018. **319**(11): p. 1113-1124.

8. Bates, D.W. and H. Singh, *Two Decades Since To Err Is Human: An Assessment Of Progress And Emerging Priorities In Patient Safety.* Health Affairs, 2018. **37**(11): p. 1736-1743.

9. Boudon, R., *Limitations of Rational Choice Theory.* American Journal of Sociology, 1998. **104**(3): p. 817-828.

10. Frakt, A. and A.K. Jha, *Face the Facts: We Need to Change the Way We Do Pay for Performance.* Annals of Internal Medicine, 2018. **168**(4): p. 291-292.

11. Mendelson, A., et al., *The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review.* Ann Intern Med, 2017. **166**(5): p. 341-353.

12. Markovitz, A.A. and A.M. Ryan, *Pay-for-performance: disappointing results or masked heterogeneity?* Medical Care Research and Review, 2017. **74**(1): p. 3-78.

13. OECD, *Better Ways to Pay for Health Care*. 2016, OECD Publishing: Paris.

14. Eijkenaar, F., et al., *Effects of pay for performance in health care: a systematic review of systematic reviews.* Health Policy, 2013. **110**(2-3): p. 115-30.

Field Code Changed

15. Van Herck, P., et al., *Effects, design choices, and context of P4P in health care: systematic review*.* BMC health services research, 2010. **10**(1): p. 1-13.

16. Meacock, R., S. Kristensen, and M. Sutton, *Paying for improvement: recent experience in the National Health Service.* Nordic Journal of Health Economics, 2014. **2**(1).

17. Mathes, T., et al., *Pay for performance for hospitals.* Cochrane Database of Systematic Reviews, 2019. **7**: p. CD011156.

18. Centers for Medicare and Medicaid Services, *Medicare program; changes to the hospital inpatient prospective payment systems and fiscal year 2008 rates.* Federal Register, 2007. **72**(162): p. 47129-8175.

19. Centers for Medicare and Medicaid Services. *Hospital Readmissions Reduction Program (HRRP)*. 2022 accessed 15 July 2023]; Available from: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/value-based-programs/hrrp/hospital-readmission-reduction-program

20. Centers for Medicare and Medicaid Services. *Hospital-Acquired Condition Reduction Program*. 2022 [accessed 15 July 2023]; Available from: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/hac-reduction-program

21. Girault, A., et al., *Implementing hospital pay-for-performance: Lessons learned from the French pilot program.* Health Policy, 2017. **121**(4): p. 407-417.

22. Lalloué, B., et al., *Evaluation of the effects of the French pay-for-performance program—IFAQ pilot study.* International Journal for quality in Health care, 2017. **29**(6): p. 833-837.

23. IHACPA. *Safety and quality*. 2022 [accessed 1 July 2023]; Available from: https://www.ihpa.gov.au/what-we-do/safety-and-quality

24. Runciman, B., A. Merry, and M. Walton, *Safety and Ethics in Healthcare: A Guide to Getting it Right*. 2007, London, UK: CRC Press.

25. Page, M.J., et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.* BMJ, 2021. **372**: p. n71.

Field Code Changed

26. NICE. *Appendix F Quality appraisal checklist – quantitative intervention studies*. Methods for the development of NICE public health guidance (third edition) 2023 [accessed 25 May2023]; Available from: https://www.nice.org.uk/process/pmg4/chapter/appendix-f-quality-appraisal-checklist-quantitative-intervention-studies

27. Shea, B.J., et al., *AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both.* BMJ, 2017. **358**: p. j4008.

28. Cashin, C., et al., *Paying for performance in health care: implications for health system performance and accountability*. 2014.

29. Eagar, K., J. Sansoni, and C. Loggie, *A Literature Review on Integrating Quality and Safety into Hospital Pricing Systems*. 2013, Centre for Health Service Development.

30. McDonald, R., et al., *A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs. An evaluation report commissioned by the Department of Health*. 2012.

31. Audit Commission, *Best practice tariffs and their impact*. 2012, Audit Commission for Local Authorities and the National Health Service in England.

32. AHRQ. *Healthcare Acquired Conditions (HAC) Annual Report*. 2019 [cited 2023 16 June]; Available from: https://www.ahrq.gov/hai/pfp/hacreport.html

33. Thirukumaran, C.P., et al., *Impact of Medicare's Nonpayment Program on Venous Thromboembolism Following Hip and Knee Replacements.* Health Services Research, 2018. **53(6)**: p. 4381-4402.

34. Thirukumaran, C.P., et al., *Impact of Medicare's Nonpayment Program on Hospital-acquired Conditions.* Medical Care, 2017. **55(5)**: p. 447-455.

35. Gidwani, R. and J. Bhattacharya, *CMS Reimbursement Reform and the Incidence of Hospital-Acquired Pulmonary Embolism or Deep Vein Thrombosis.* Journal of General Internal Medicine, 2015. **30**(5): p. 588-596.

36. Kim, K.M., et al., *Evaluation of Clinical and Economic Outcomes Following Implementation of a Medicare Pay-for-Performance Program for Surgical Procedures.* JAMA Network Open, 2021. **4**(8): p. e2121115.

Field Code Changed

37. Jha, A.K., et al., *The long-term effect of the Premier Hospital Quality Improvement Demonstration on patient outcomes.* New England Journal of Medicine, 2012. **366**(17): p. 1606-1615.

38. Glickman, S.W., et al., *Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction.* JAMA, 2007. **297**(21): p. 2373.

39. Ryan, A.M., *Effects of the Premier Hospital Quality Incentive Demonstration on Medicare Patient Mortality and Cost.* Health Serv Res, 2009. **44**(3): p. 821-42.

40. Shih, T., et al., *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): p. 677-681.

41. Calikoglu, S., R. Murray, and D. Feeney, *Hospital pay-for-performance programs in Maryland produced strong results, including reduced hospital-acquired conditions.* Health Affairs, 2012. **31**(12): p. 2649-58.

42. Keller, M.S., et al., *Evaluating inpatient adverse outcomes under California's Delivery System Reform Incentive Payment Program.* Health Services Research, 2021. **56**(1): p. 36-48.

43. Sutton, M., et al., *Reduced mortality with hospital pay for performance in England.* The New England journal of medicine, 2012. **367**(19): p. 1821-1828.

44. Kristensen, S.R., et al., *Long-Term Effect of Hospital Pay for Performance on Mortality in England.* New England Journal of Medicine, 2014. **371**(6): p. 540-548.

45. McDonald, R., et al., *A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West.* Health Services and Delivery Research, 2015. **3**(23): p. 1-104.

46. Metcalfe, D., et al., *Pay for performance and hip fracture outcomes: an interrupted time series and difference-in-differences analysis in England and Scotland.* The bone & joint journal, 2019. **101**(8): p. 1015-1023.

47. Alrawashdeh, M., et al., *Assessment of Federal Value-Based Incentive Programs and In-Hospital Clostridioides difficile Infection Rates.* JAMA Network Open, 2021. **4**(10): p. e2132114.

Field Code Changed

48. Arntson, E., et al., *Changes in hospital acquired conditions and mortality associated with the hospital acquired condition reduction program.* Journal of General Internal Medicine, 2019. **34(2 Supplement)**: p. S159.

49. Hsu, H.E., 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): p. 13.

50. Hsu, H.E., et al., *Association between value-based incentive programs and catheter-associated urinary tract infection rates in the critical care setting.* JAMA - Journal of the American Medical Association, 2019. **321(5)**: p. 509-511.

51. Waters, T.M., 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): p. 958.

52. Calderwood, M.S., et al., *Impact of Medicare's Payment Policy on Mediastinitis Following Coronary Artery Bypass Graft Surgery in US Hospitals.* Infection Control & Hospital Epidemiology, 2014. **35**(2): p. 144-151.

53. Calderwood, M.S., 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)**: p. 100-103.

54. He, J., et al., *Unit-level time trends and seasonality in the rate of hospital-acquired pressure ulcers in US acute care hospitals.* Research in Nursing & Health, 2013. **36**(2): p. 171-180.

55. Vaz, L.E., 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): p. 649-655.

56. Zielinski, M.D., et al., *Is the Centers for Medicare and Medicaid Service's lack of reimbursement for postoperative urinary tract infections in elderly emergency surgery patients justified?* Surgery (United States), 2014. **156(4)**: p. 1009-1017.

57. Lee, G.M., et al., *Effect of Nonpayment for Preventable Infections in U.S. Hospitals.* New England Journal of Medicine, 2012. **367**(15): p. 1428-1437.

Field Code Changed

58. Padula, W.V., et al., *Hospital-Acquired Pressure Ulcers at Academic Medical Centers in the United States, 2008-2012: Tracking Changes Since the CMS Nonpayment Policy.* Joint Commission Journal on Quality & Patient Safety, 2015. **41**(6): p. 257-63.

59. Padula, W.V., et al., *Adverse Effects of the Medicare PSI-90 Hospital Penalty System on Revenue-Neutral Hospital-Acquired Conditions.* Journal of patient safety, 2020. **16(2)**: p. e97-e102.

60. Zogg, C.K., et al., *Learning From England's Best Practice Tariff: Process Measure Pay-for-Performance Can Improve Hip Fracture Outcomes.* Ann Surg, 2022. **275**(3): p. 506-514.

61. Stone, P., et al., *Does pay-for-performance improve patient outcomes in acute exacerbation of COPD admissions?* Thorax, 2022. **77**: p. 239-246.

62. Whitaker, S.R., et al., *Does achieving the 'Best Practice Tariff' criteria for fractured neck of femur patients improve one year outcomes?* Injury, 2019. **50**(7): p. 1358-1363.

63. Stein, J.D., et al., *Use of health care claims data to study patients with ophthalmologic conditions.* Ophthalmology, 2014. **121**(5): p. 1134-41.

64. Ferver, K., B. Burton, and P. Jesilow, *The Use of Claims Data in Healthcare Research.* The Open Public Health Journal, 2009. **2**(1): p. 11-24.

65. Bae, S.H., *CMS Nonpayment Policy, Quality Improvement, and Hospital-Acquired Conditions: An Integrative Review.* Journal of Nursing Care Quality, 2017. **32**(1): p. 55-61.

66. Razi, S.S., et al., *Pay-for-performance programs reduce hospital costs without compromising care in surgical patients.* Journal of the American College of Surgeons, 2011. **1)**: p. S112.

67. Mubark, I., et al., *Mortality Following Distal Femur Fractures Versus Proximal Femur Fractures in Elderly Population: The Impact of Best Practice Tariff.* Cureus, 2020. **12**(9): p. e10744.

68. Wong, E., et al., *The impact of best practice tariff attainment on 30-day mortality and length of stay in hip fracture patients: an observational cohort study at a major trauma centre.* Orthopaedic Proceedings, 2022. **104-B**(SUPP_4): p. 13-13.

Field Code Changed

69. Tsai, Y.S., et al., *Effects of pay for performance on risk incidence of infection and of revision after total knee arthroplasty in type 2 diabetic patients: A nationwide matched cohort study.* Plos One, 2018. **13**(11): p. 18.

70. Epstein, A., A. Jha, and E. Oray, *The Impact of Pay-for-Performance on Quality of Care for Minority Patients.* The American journal of managed care, 2014. **20**: p. e479-86.

71. Rodriguez, M., et al., *Incidence of C. difficile infection Cases by CDI-Standardized Infection Ratio among Centers for Medicare and Medicaid Services-Participating Short-Term Acute Care Hospitals Post Implementation of the Hospital-Acquired Condition Reduction Program (HACRP).* Open Forum Infectious Diseases, 2021. **8(SUPPL 1)**: p. S480-S481.

72. Rude, T.L., 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**: p. 79-84.

73. Figueroa, J.F., et al., *Association between the value-based purchasing pay for performance program and patient mortality in US hospitals: observational study.* bmj, 2016. **353**.

74. Berthiaume, J.T., et al., *Aligning financial incentives with quality of care in the hospital setting.* Journal for Healthcare Quality, 2006. **28**(2): p. 36-44, 51.

75. Oakley, B., et al., *Does achieving the best practice tariff improve outcomes in hip fracture patients? An observational cohort study.* BMJ Open, 2017. **7**(2): p. e014190.

76. Tedesco, D., et al., *Improvement in Patient Safety May Precede Policy Changes: Trends in Patient Safety Indicators in the United States, 2000-2013.* Journal of patient safety, 2021. **17**(4): p. e327-e334.

77. Eldridge, N., et al., *Trends in Adverse Event Rates in Hospitalized Patients, 2010-2019.* JAMA, 2022. **328**(2): p. 173.

78. Eagar, K., et al., *A literature review on integrating quality and safety into hospital pricing systems*. 2013, Australian Health Services Research Institute: Sydney.

79. Kruse, G.B., et al., *The impact of hospital pay-for-performance on hospital and Medicare costs.* Health services research, 2012. **47**(6): p. 2118-2136.

Field Code Changed

80. Centers for Medicare and Medicaid Services. *Premier Hospital Quality Incentive Demonstration*. 2021 [accessed 17 July 2023]; Available from: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalpremier

81. McNair, P.D., H.S. Luft, and A.B. Bindman, *Medicare's policy not to pay for treating hospital-acquired conditions: the impact.* Health Affairs, 2009. **28**(5): p. 1485-93.

82. Centers for Medicare and Medicaid Services. *Hospital-Acquired Conditions (Present on Admission Indicator)*. 2021 [accessed 12 July2023]; Available from: https://www.cms.gov/medicare/medicare-fee-for-service-payment/hospitalacqcond

83. Medicare Learning Network. *Medicare Payment Systems*. 2023 [accessed 12 July2023]; Available from: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/html/medicare-payment-systems.html#Acute

84. Milstein, R. and J. Schreyoegg, *Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries.* Health Policy, 2016. **120**(10): p. 1125-1140.

85. Department of Health and Social Care. *Payment by Results 2013-14*. 2013 [accessed 12 July 2023]; Available from: https://www.gov.uk/government/collections/payment-by-results-2013-14

86. Marshall, L., A. Charlesworth, and J. Hurst, *The NHS payment system: evolving policy and emerging evidence*. 2014, Nuffield Trust.

87. NHS Digital. *Hospital Admitted Patient Care Activity*. 2023 [accessed 12 July2023]; Available from: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity

88. Digital, N. *SUS+ PbR Reference Manual*. 2019 [accessed 12 July2023]; Available from: https://digital.nhs.uk/services/secondary-uses-service-sus/payment-by-results-guidance/sus-pbr-reference-manual

89. Spaulding, A., M. Zhao, and D.R. Haley, *Value-based purchasing and hospital acquired conditions: Are we seeing improvement?* Health Policy, 2014. **118**(3): p. 413-421.

Field Code Changed

90. Royal College of Physicians. *National Hip Fracture Database*. 2023 [cited 2023 12 July]; Available from: https://www.nhfd.co.uk/20/NHFDcharts.nsf/vwCharts/BestPractice.

91. Kawai, A.T., et al., *Impact of the Centers for Medicare and Medicaid Services Hospital-Acquired Conditions Policy on Billing Rates for 2 Targeted Healthcare-Associated Infections.* Infection control and hospital epidemiology, 2015. **36(8)**: p. 871-877.

92. Rhee, C., et al., *Impact of the 2012 medicaid health care-acquired conditions policy on catheter-associated urinary tract infection and vascular catheter-associated infection billing rates.* Open Forum Infectious Diseases, 2018. **5(9)**.

93. Harris, J.M., 2nd, et al., *Comparison of Administrative Data Versus Infection Control Data in Identifying Central Line-Associated Bloodstream Infections in Children's Hospitals.* Hosp Pediatr, 2013. **3**(4): p. 307-13.

94. Werner, R.M., et al., *The effect of pay-for-performance in hospitals: lessons for quality improvement.* Health affairs, 2011. **30**(4): p. 690-698.

95. Kristensen, S.R., M. Bech, and J.T. Lauridsen, *Who to pay for performance? The choice of organisational level for hospital performance incentives.* The European Journal of Health Economics, 2016. **17**: p. 435-442.

96. Roberts, E.T., A.M. Zaslavsky, and J.M. McWilliams, *The Value-Based Payment Modifier: Program Outcomes and Implications for Disparities.* Annals of Internal Medicine, 2018. **168**(4): p. 255.

97. Meacock, R., S.R. Kristensen, and M. Sutton, *The cost-effectiveness of using financial incentives to improve provider quality: framework and application.* Health Economics, 2014. **23**(1): p. 1-13.

98. Sen, A.K., *Rational Fools: A Critique of the Behavioral Foundations of Economic Theory.* Philosophy & Public Affairs, 1977. **6**(4): p. 317-344.

99. Khullar, D., D. Wolfson, and L.P. Casalino, *Professionalism, Performance, and the Future of Physician Incentives.* JAMA, 2018. **320**(23): p. 2419.

Field Code Changed

100. Ilic, M. and L. Markovic-Denic, *Repeated prevalence studies of nosocomial infections in one university hospital in Serbia*. Turkish Journal of Medical Sciences, 2017. **47(2)**: p. 563-569.

101. Lindenauer, P.K., Remus, D., Roman, S., Rothberg, M.B., Benjamin, E.M., Ma, A., Bratzler, D.W. *Public Reporting and Pay for Performance in Hospital Quality Improvement*. New England Journal of Medicine, 2007 Vol. 356 Issue 5 Pages 486-496

102. Campbell, R.J. *Change management in health care*. The Health Care Manager, 2008 Vol. 27 Issue 1 Pages 23-39

Field Code Changed

Figure 1. PRISMA flow diagram

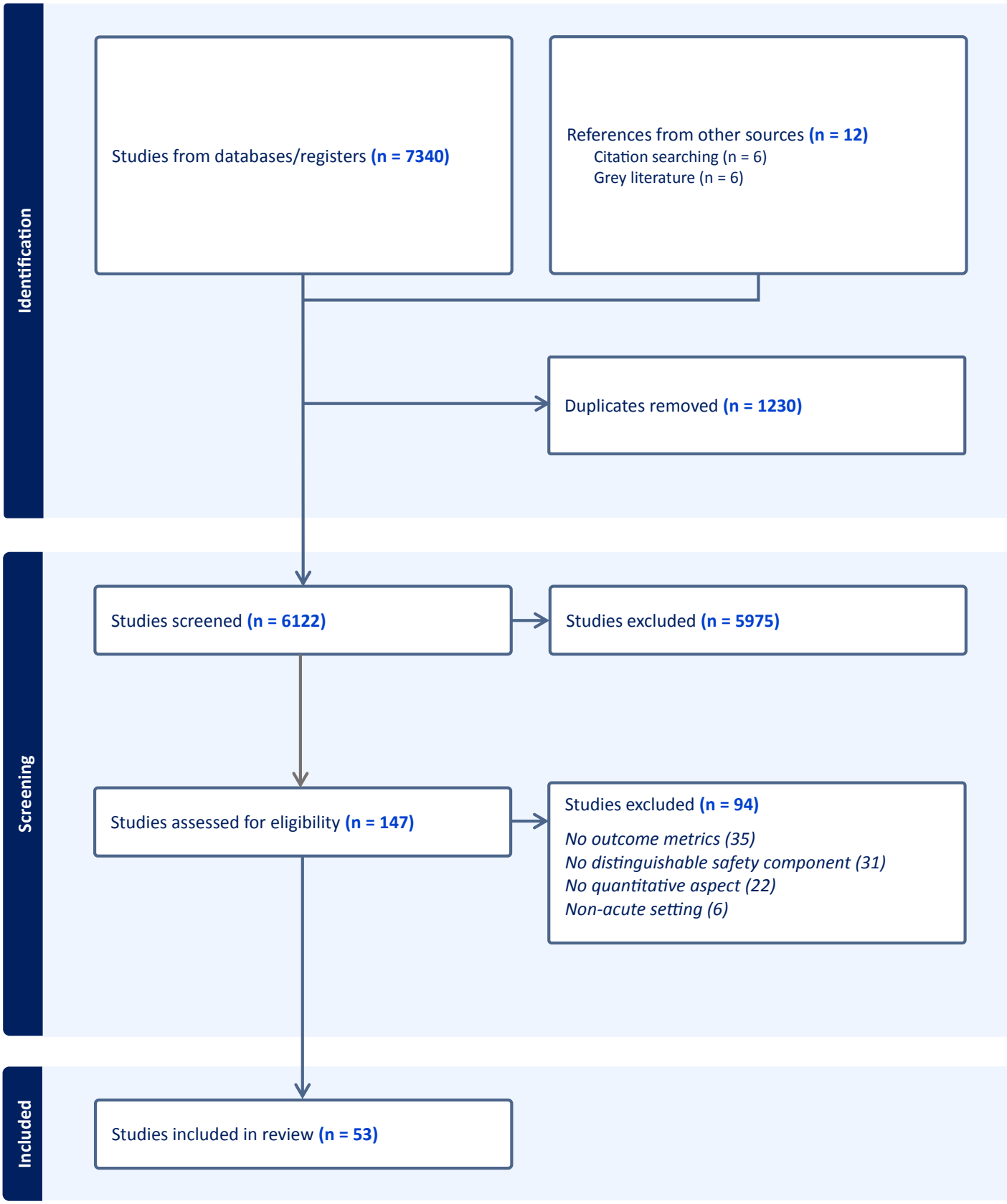

Table 1. Original investigations: summary and key findings

Reference	Location	System context	P4P Initiative	Sample & data	follow-up period (years)	Metric(s)	Effect[a]	Study quality[b]
Shih et al 2014 [40]	USA	Medicare	HQID	State Inpatient Databases (SID) 2003-09 861,411 admissions for CABG & THR/TKR	3	HACs	0*	G
						mortality	0†	
Jha et al 2012 [37]	USA	Medicare	HQID	Medicare claims data 2002-09 252 participating HQID hospitals 3363 hospitals public reporting only ~6m admissions for AMI, CHF, pneumonia, CABG	6	mortality	0†	F
Glickman 2007 [38]	USA	Medicare	HQID	CRUSADE data (voluntary AMI registry) 2003-2006 600 hospitals 105,000 admissions for AMI	3	mortality	0†	P
Ryan 2009 [39]	USA	Medicare	HQID	Medicare claims data 2000-06 3570 hospitals ~11M admissions for AMI, CHF, pneumonia, CABG,	2	mortality	0†	P

Tedesco et al 2021 [76]	USA	Medicare	HAC-POA	Healthcare Cost and Utilization Project data 91M hospitalisations 2000-13	5	HACs	Mod* (15%)	F
Kim et al 2021 [36]	USA	Medicare	HAC-POA	Healthcare Cost and Utilization Project data 20014-17 ~1.4m admissions	8	SSI	High* (72%)	F
						VTE	0*	
						mortality	0†	
Padula et al 2020 [59]	USA	Medicare	HAC-POA	Vizient Clinical Data Base 2013-16 306 hospitals ~7M admissions	3	HACs	Mod* (24%)	P
Thirukumaran et al 2018 [33]	USA (NY)	Medicare	HAC-POA	State Inpatient Database NY 2005-13. 98,729 THRs 111,361 TKRs	5	VTE	Mod* (45%)	P
Thirukumaran et al 2017 [34]	USA (NY)	Medicare	HAC-POA	State Inpatient Database NY 2005-12 867,584 admissions 159 hospitals	4	CLABSI, CAUTI, GLU, air emboli, falls & transfusion error	High* (50%)	P
Calderwood et al 2016 [53]	USA	Medicare	HAC-POA	NHSN surveillance data, American Hospital Association Annual Database	4	CLABSI (hospitals with *high* operating margin (OM))	0	F

				2007-13 358 hospitals		CLABSI (hospitals with *low* OM)	Mod (24%)	
Padula et al 2015 [58]	USA	Medicare	HAC-POA	Medicare claims data 2008-12 210 hospitals ~4M admissions	4	HAPU	High* (93%)	P
Vaz et al 2015 [55]	USA	Medicare	HAC-POA	NHSN data 2007-2014 656 hospitals	5	CLABSI	0	P
Calderwood et al 2014 [52]	USA	Medicare	HAC-POA	Medicare Claims data & NHSN surveillance data 2006-10 1234 hospitals 638,761 procedures (CABG)	2	Mediastinitis (claims data)	High* (64%)	F
						Mediastinitis (NHSN data)	0	
Gidwani & Bhattacharya 2015 [35]	USA	Medicare	HAC-POA	Medicare claims data 2007-09 136,634 admissions THR/TKR	1.25	VTE	Mod* (35%)	P
Zielinski et al 2014 [56]	USA	Medicare	HAC-POA	American College of Surgeons NSQIP data 2005-12 374 hospitals 53,879 admissions	3.25	CAUTI	Mod (29%)	P

He et al 2013 [54]	USA	Medicare	HAC-POA	Quarterly on-site prevalence surveys 2004-2011 5447 acute care nursing units 733 hospitals	3	HAPU	Mod (16%)	F
Lee et al 2012 [57]	USA	Medicare	HAC-POA	NHSN surveillance data 2006-11 398 hospitals	2.25	CLABSI, CAUTI	0	P
Waters et al 2022 [51]	USA	Medicare	HACRP	Healthcare Cost and Utilization Project Data (claims data) 2007–16 1423 hospitals	2	HACs	0*	F
						mortality	0†	
Alrawashdeh et al 2021 [47]	USA	Medicare	HACRP	NHSN surveillance data 2013-19 265 hospitals ~24M admissions	2.5	C. Diff	Mod (6%)	F
Rodriguez et al 2021 [71]	USA	Medicare	HACRP	Medicare public data 2015-18	4	C.Diff	0*	P
Arntson et al 2021 [48]	USA	Medicare	HACRP	MEDPAR claims data 2009-15 3340 hospitals ~8.9M admissions	2	HACs	Mod* (18%)	P
						mortality	0†	
Hsu et al 2020 [49]	USA	Medicare	HACRP	NHSN surveillance data 2013-18 618 hospitals	3.25	CLABSI, CAUTI, SSI	0	F

Hsu et al 2019 [50]	USA	Medicare	HACRP	NHSN surveillance data 2013-17 592 hospitals, 1185 ICUs 22 572 494 patient days	2.25	CAUTI	0	F
Rude et al 2018 [72]	USA	Medicare	HACRP	American College of Surgeons NSQIP data 2005-2012 39,000 surgical admissions	4	SSI, CAUTI, VTE	0^ [18%, *p*=0.077]	F
Figueroa et al 2016 [73]	USA	Medicare	HACRP	Medicare claims data 2008-13 4267 hospitals; ~2.4M admissions	2.25	mortality	0†	F
Keller et al 2021 [42]	USA (CA)	California Designated Public Hospitals	California Delivery System Reform Incentive Payment Program (DSRIP)	California Office of Statewide Health Planning and Development (OSHPD) patient discharge data (claims) 2009-14 17 hospitals	3	CLABSI, VTE, HAPU	0*	P
						mortality (from sepsis)	Low† (6%)	
Calicoglu, Murray & Feeney 2012 [41]	USA (MD)	Health Services Cost Review Commission	Maryland Quality-Based Reimbursement Program (Modelled on HACRP)	Maryland discharge database (claims data) 2007-10 2.8M admissions	2	HACs	Mod* (15%)	P
Berthiaume et al 2006 [74]	USA (HI)	Hawaii Medical Service Association	Hawaii Medical Service Association HQSR Program	Administrative (claims) data 2001-2004 ~150K admissions	4	HACs	0*	P

McDonald et al 2015 [45]	England	NHS	Advacing Quality	National Hospital Episode Statistics 2007-10 24 participating hospitals 130 control hospitals ~1.1M admissions	1.5	mortality (at 18 months post-intervention)	Low† (7%)	F
						mortality (at 42 months post-intervention)	0†	
Kristensen et al 2014 [44]	England	NHS	Advacing Quality	National Hospital Episode Statistics 2007-12 161 hospitals ~1.825M admissions	3.5	mortality (at 18 months)	Low† (7%)	G
						mortality (at 42 months)	0†	
Sutton et al 2012 [43]	England	NHS	Advacing Quality	National Hospital Episode Statistics 2007-10 24 participating hospitals 132 control hospitals ~1.1M admissions	1.5	mortality	Low† (6%)	F
Zogg et al 2022 [60]	England	NHS / CMS	Best Practice Tariff (BPT) hip fracture care	National Hospital Episode Statistics; ONS death certificate registrations; Medicare claims data 2000-16 ~800K admissions (England) ~3.2M admissions (USA)	6	mortality	Mod† (27%)	F

Wong et al 2022 [68]	England	NHS	BPT hip fracture care	Natioanl Hip Fracture Database (NHFD) 1 hospital 4397 admissions	8	mortality (BPT vs. non-BPT admissions)	Low† (7.8%)	P
Mubark et al 2020 [67]	England	NHS	BPT hip fracture care	Patient records 2010-14 1 hospital 378 admissions	4	mortality	Low† (13%)	P
Metcalfe et al 2019 [46]	England	NHS	BPT hip fracture care	National Hospital Episode Statistics; Office for National Statistics (ONS) death certificate registrations; Scottish Morbidity Records (SMR01) 2000-17 ~1.2M admissions	7	mortality	Low† (6%)	G
Whitaker et al (2019) [62]	England	NHS	BPT hip fracture care	National Hip Fracture Database (NHFD 2011-15 1 hospital 1,414 admissions	4.5	mortality (BPT vs. non-BPT admissions)	Low† (10.3%)	F
						mobility status	Mod (16.4%)	
						residential status	0	

Oakley et al 2016 [75]	England	NHS	BPT hip fracture care	Nottingham Hip Fracture Database 2008-14 1 hospital 2541 admissions	2	mortality (BPT vs. non-BPT admissions)	High† (71%)	P
						mortality (all hip fracture admissions, England)	0†	
Stone et al 2022 [61]	England	NHS	BPT acute exacerbation COPD	National Asthma and COPD Audit Programme (NACAP) 2017 (8 months) 181 hospitals ~28K admissions	0.7	mortality	0†	P
						readmission	0	
Tsai et al 2018 [69]	Taiwan	National Health Insurance (NHI)	NHI diabetes care P4P program	National Health Insurance (NHI) Research Database - derived from claims data 2002–2013 4200 P4P cases, 8400 propensity matched non-P4P cases	5	SSI rate	0*	F
						Surgical revisions	Mod* (47%)	

a. 0: no effect / effect not significant at 0.05; Low: 1%-14% improvement; Mod: 15%-49% improvement; High: 50%+ improvement; † mortality used as metric; * HAC rates calculated using claims/administrative data; ^ within 0.05 of alpha value and with an effect of 15% or greater.

b. G: good; F: fair; P: poor

Abbreviations: BPT: Best Practice Tariff; CAUTI: Catheter-associated urinary tract infection; C.Diff: Clostridium difficile; CE: cost-effective; CLABSI: Central line-associated bloodstream infection; COPD: chronic obstructive pulmonary disease; GLU: glucose mismanagement; HAC: Hospital-acquired condition; SSI: Surgical site infection; VTE: Venous thromboembolism

Table 2. Literature reviews: summary and key findings

Authors	Care setting	Relevant P4P initiative(s) examined	No. of papers	Method	Results specific to patient safety in acute care	Findings relevant to acute care	Quality
Van Herck et al 2010 [15]	All	All	128	Systematic review	The effectiveness of P4P implemented to date is highly variable. P4P most frequently failed to affect hospital care. Based on the evidence, the review has provided further indications on how effect findings are likely to relate to P4P design choices and context.	Inconclusive	G
Eijkenaar et al 2013 [14]	All	All	22	Systematic review of systematic reviews	Convincing evidence still lacking especially for hospital care. Many studies failed to find an effect and there are still few studies that convincingly disentangled the P4P effect from other improvement initiatives. The most rigorous evidence comes from the Hospital Quality Incentive Demonstration (HQID) where no impact was found on mortality, despite that for some conditions 30-day risk-adjusted mortality was included as a performance measure.	Inconclusive	G
Meacock, Kristensen & Sutton 2014 [16]	All	All (England)	–	Literature review	P4P can result in modest short-term improvements in the incentivised aspects of performance. There is little evidence of effort diversion, yet some to suggest positive spillovers of these schemes onto non-incentivised aspects of performance. While there is some evidence of gaming and inequitable consequences, these do not appear to be widespread. P4P programmes are likely to be most effective when introduced as a supporting part to a wider quality improvement initiative, and when results are published to encourage a reputational as well as a financial incentive for improvement	Inconclusive	G
Milstein & Schreyoegg 2016 [84]	Hospital	All (OECD member countries)	–	Literature review	The initiatives reviewed yielded modest, short-term improvements at best. The authors suggest that these may be attributed to spillover effects, such as public reporting and increased awareness of data recording. The review also found numerous methodological flaws suggesting that the full potential of P4P has not been explored yet. Overall, p4p has yet to live up to expectations.	Inconclusive	G

Bae e al 2017 [65]	Hospital	HAC-POA	14	Integrative review	Strong evidence that the CMS policy has spurred quality improvement initiatives, but the relationships between the policy and outcomes including hospital-acquired conditions are inconclusive.	Inconclusive	G
Markovitz & Ryan 2017 [12]	All	All	58	Literature review	P4P has not succeeded in satisfying its goals either efficiently or equitably. While it seems that the main effects of P4P may be masking some heterogeneity across organization types, this variation does not appear sufficient to alter the conclusion that P4P has largely failed to realize substantial quality improvements.	Inconclusive	F
Mendelson et al 2017 [11]	All	All	69	Systematic review	No clear evidence to suggest that P4P programs improve patient outcomes. In the hospital setting, P4P programs generally did not decrease mortality.	Inconclusive	G
Mathes et al 2019 [17]	Hospital	All	27	Cochrane review	Most studies showed no difference or a very small effect in favour of the P4P program. Uncertain whether P4P affects patient outcomes, quality of care, equity, or resource use as the certainty of the evidence was very low. The effects on patient outcomes of P4P in hospitals were small at best, regardless of design factors and context or setting.	Inconclusive	G

G: good; F: fair