Pay for performance at a crossroads: Lessons from taking a global perspective

Søren Rud Kristensen[1,2], Laura Anselmi[3], Garrett Wallace Brown[4], [5]Eleonora Fichera, Roxanne Kovacs[6], Rene Loewenson[7], Neha Sing[8], Nicholas Midzi[9], Fatimah Mustapha[10], Lee White[8], Josephine Borghi[8]

[1] Centre for Health Policy, Imperial College London, United Kingdom

[2] Danish Centre for Health Economics, University of Southern Denmark, Denmark

[3] Health Policy, Organisation and Economics, School of Health Sciences, The University of Manchester, United Kingdom

[4] School of Politics and International Studies, University of Leeds, United Kingdom

[5] Department of Economics, University of Bath, United Kingdom

[5] Department of Economics, University of Gothenburg, Sweden

[6] Training and Research Support Centre, Zimbabwe

[7] London School of Hygiene and Tropical Medicine, United Kingdom

[8] National Institute of Health Research, Ministry of Health and Child Care, Zimbabwe

[9] The World Bank Group, Nigeria

ABSTRACT

Purpose: The use of pay for performance as an instrument to incentivize quality improvements in health care is at a crossroads in high-income countries but has remained a commonly used tool in low- and middle-income countries. We aimed to take stock of the evidence on effectiveness and design from across income settings, to reveal insights for the future design of performance payment across income contexts

Methodology: We identified Cochrane literature reviews of the use of pay for performance in health care in any income setting, tracked the development in the quantity and quality of evidence over time, and compared the incentive design features used across high-income countries compared to low-and middle income countries.

Findings: The quantity and quality of the evidence base has grown over time but can still be improved. Scheme design varies across income settings, and although some design choices may reflect differences in context, we find that incentive designers in both income settings can learn from practices used in the other setting.

Originality: The research and literature on P4P in high-, low- and middle-income countries largely operate in silos. By taking stock of the evidence on P4P from across income settings we are able to draw out key insights between these settings which remain underexplored in the literature.

Keywords: Pay for performance; Quality of health care; Quality improvement

INTRODUCTION

Pay for performance, the idea of incentivising improvements in the quality of health care by linking provider payments to specific indicators, has intuitive appeal and has caught the interest of policymakers in both high, low and middle income countries.

In high-income countries (HICs) the term has been used mainly to refer to payment schemes aiming to improve the quality of care. Here, the use of P4P really took off in the early 2000s after the U.S. Institute of Medicine's publication of the reports "To Err is Human" and "Crossing the Quality Chasm" (2001; 2000), which raised concerns about the impact of existing activity-based provider payment models (e.g. fee for service and prospective payment system) on the quality of care. Increasingly, however, the evidence from HICs suggests that P4P has had limited success in incentivising long-lasting improvements in outcomes of care (Jha et al., 2012) and while HIC policymakers continue to use payments to improve quality and outcomes (Centers for Medicare & Medicaid Services, 2021) there have been calls for a fundamental rethinking of the role for P4P in improving the quality of care (Frakt and Jha, 2017).

Meanwhile, in low- and middle-income countries (LMICs) the use of P4P has increased substantially in the last 15 years, promoted and supported by key international funders as a means of linking

their health system aid investments to results. For example, between 2008 and 2016, the Health Results Innovation Trust Fund (HRITF) was involved in the implementation of P4P schemes in 32 LMICs, investing US $420 million linked to US $2.4 billion from the International Development Association(Kandpal, 2016). In LMICs, P4P is sometimes understood to encompass both quality and activity, and is seen as a response to fluctuating public health system funding and the lack of consistent investment effects, and involves a redesign of health system financing arrangements to maximise efficiency goals(de Walque et al., 2022; Diaconu et al., 2022). Moreover, bilateral funders and agencies such as the World Bank(Preker and Langenbrunner, 2005) and WHO (2000) have increasingly promoted P4P as a form of strategic purchasing for UHC (Mathauer et al., 2019), which further embeds P4P into development aid polices, but also has created uncertainty regarding how P4P renders purchasing 'strategic' in LMIC contexts(Paul et al., 2020).

Commented [A1]: Garrett, perhaps you could add a brief discussion about strategic purchasing vs P4P here?

It has been noted that the research and literature on P4P in high-, low- and middle-income countries (HLMIC) operate in silos (Anselmi et al., 2020). Notwithstanding the differences in context, important insights can be missed if researchers fail to share learning from across income contexts. In this paper, we aim to take stock of the evidence on P4P from HLMICs to draw out key insights between these settings which remain underexplored in the literature. We review the evidence on the

effectiveness of P4P and explore the variation in design features across income settings. We then suggest lessons that can be gained from looking at the literature from across HLMICs.

METHODS

We first review the evidence on P4P effectiveness across HLMIC and then compare the design features across HIC compared to LMIC settings.

Evidence on effectiveness

We searched the Cochrane Database of Systematic Reviews, for reviews that had examined the impact of P4P on health care provision in any income setting. We identified seven Cochrane Reviews published between 2011 and 2021. To assess P4P effects we applied Donabedian's framework of structures (e.g. availability of medicines, infrastructure), processes (e.g. clinical protocol adherence) and outcomes of care (e.g. health status)(Donabedian, 1988) and extracted evidence relating to these dimensions. We also recorded the certainty of evidence according to the GRADE framework, and key observations by review authors. We present the evidence chronologically reflecting on its evolution over time.

Comparison of design features

We identified two recent reviews that systematically developed and applied typologies of P4P schemes for classifying studies

from the HLMIC literatures. Ogundeji et al.(2018) presented a theory-driven framework for reporting design characteristics of P4P schemes, which they applied to 13 P4P schemes from nine HIC countries identified in a previous review(Eijkenaar, 2012). Kovacs et al. (2020) developed a reporting framework specifically focused on LMIC schemes. They applied the scheme to 41 P4P schemes in 29 LMICs.

We derived a unified framework based on a summary of the two reporting schemes as presented in Table one. There was direct agreement between the reporting frameworks for most dimensions, although for some dimensions the framework by Kovacs et al. was more granular. In those instances, we joined categories from the Kovacs et al. framework to match those reported by Ogundeji et al.

[Table I]

RESULTS

Evidence on effectiveness

Table Two and Figure one give an overview of how the evidence base for P4P in HLMICs has developed over time. For both settings, the quantity and quality of evidence have developed positively between 2011 and 2021, although the evidence strength could be improved in both settings.

Scott et al.(2011) examined the impact of financial incentives on the quality of primary care provision in HICs. The use of P4P was found to be positive, but with only modest effects on process quality. Moreover, the positive effects were only demonstrated for a minority of the multiple outcomes examined. For example, three of the seven studies examined the impact of incentives for smoking cessation. The studies demonstrated a positive impact on processes of care, such as recording smoking status, but no impact on outcomes, such as patients' short-term abstinence from smoking or intentions to quit.

Witter et al. (2012) conducted the first Cochrane review of P4P in LMICs. The authors found some evidence that P4P had the intended impact of improving quality for specific interventions, but the findings were mixed, with both negative and positive effects on the quality of care recorded. As a result, the review posited that the current evidence did not allow them to draw clear conclusions in favour or against the use of P4P in LMICs.

Rashidian et al. (2015) examined the effect of financial incentives on medicine prescribers in primary and secondary care in HICs. A scheme incentivising guideline adherence demonstrated only modest effects on a subset of targeted guidelines and effects were not sustained over time. The authors noted, however, that the finding could be due to the scheme design which offered the payment independently of actual performance.

Two studies found no effect on primary care prescribing or health outcomes after P4P. The authors cautioned that the effect of P4P on drug use and health outcomes was uncertain due to the quality of the evidence which, using the GRADE terminology, was rated very low.

Yuan et al. (2017) examined the impact of various payment methods for outpatient care facilities across a range of HLMIC. The authors found with moderate certainty that adding P4P to the existing payment system was associated with a slight improvement in health professionals' use of tests and treatments, yet made little or no difference in adherence to quality assurance criteria. One study compared fee-for-service (FFS) to capitation combined with P4P based on antibiotic prescriptions and patient satisfaction, finding that P4P reduced antibiotic prescriptions.

Mathes et al (2019) reviewed the evidence on P4P in hospital settings in HICs. The authors found moderate improvements in process measures of quality and smaller effects when evaluating the impact on patient outcomes. The authors found little difference in the effect size across the six programmes in the review, but the scheme that relied on penalties rather than bonuses was associated with greater improvement in outcomes. However, the authors noted that the certainty of the evidence was very low.

Diaconu et al (2021) updated the earlier review of P4P in LMICs by Witter et al. (2012) now with a significantly larger evidence base. When comparing P4P to standard care, the review was inconclusive regarding the impact on process quality while there was some improvement in intermediate quality indicators summarised by quality scores of medicines and equipment availability. The review found mixed effects of the impact on health care use and delivery including immunisation rates. For example, HIV testing seemed to increase, while children and households protected by bednets could decrease after introducing P4P. Evidence was found for modest reductions in child mortality and successful tuberculosis treatment associated with the introduction of P4P but showed inconsistent effects on neonatal mortality. When comparing P4P to comparator interventions, the review found evidence of increases in process quality and quality of family planning and antenatal care, but inconsistent, little or no effect on e.g. immunisation uptake, antenatal care and postnatal care. There was little evidence of an impact on health outcomes. The certainty of the evidence was in all cases rated low.

[Table II]

[Figure I]

Finally, Jia et al.(Jia et al., 2021) examined the impact of various payment methods for health care providers working in

outpatient settings. The authors found that P4P was associated with increased service provision including immunisation. The services could be of better quality, for example regarding the choice of medicines, but only while performance payments were in place. The evidence on the impact on patient health outcomes was mixed, however. The quality of the evidence ranged from moderate to very low.

Evidence on design features

When reflecting on the evidence, it is worth remembering that P4P is not a uniform intervention and can have multifarious design and implementation features which may be related to effectiveness.(Roland, 2012; Witter et al., 2012) Table three compares the design features of P4P schemes applied in HICs versus LMICs. In both settings, P4P schemes predominantly rely on structural or process indicators which are mostly within the control of the provider. Still, outcome-based schemes are more common in LMICs (17%) than HICs (8%).

Schemes in HIC and LMIC differ greatly in who ultimately receives the performance payment. While all included HIC schemes gave incentives at the group level (e.g. hospital or practice level) only 45% of LMIC schemes incentivised facilities, while 84% incentivised individuals.

Across both settings, schemes are predominantly based on bonuses rather than penalties. In LMICs incentives provided are large

relative to provider income, while in HICs that's only true in 50% of the schemes.

Just over 50% of HIC schemes have a long time-lag between performance and payment while in LMICs just 26% of schemes rely on annual or half-yearly payments. The remaining schemes pay out bonuses much more frequently.

With respect to the type of competition for payments, in HICs, one third of schemes measure providers' performance relative to other providers, while a tournament style performance measurement is only used in 5% of schemes in LMICs.

Finally, P4P schemes in HICs are predominantly based on government funding while this is true for only 25% of P4P schemes in LMICs. In this setting, the most common funding source is an international agency or NGO.

[Table III]

DISCUSSION

The evidence base for P4P

This review has focused on the evidence coming from research papers using methodologies that can be included in a Cochrane review, and as such is only a partial picture of the full scale of P4P research that exists.

Within this sample, the evidence base for P4P, has grown from small to rather substantial, and it is encouraging to see that both the number of studies and the quality of the evidence has increased over time. Still, there is room for improvement, and the large number of evaluations that cannot fulfil the inclusion criteria of the Cochrane Review signals necessary reflections about research quality, across all income settings, but also how we judge 'quality research' of evidence from payment reforms where experimental designs are often desirable but may not be feasible. The need to promote best practice is strengthened by recent findings that more robust study designs are less likely to find an impact of P4P(Scott et al., 2018). Moving ahead, researchers should ensure that the methods they choose are suitable for producing robust evidence and acknowledge study limitations, sources of biases as well as conflicts of interest(Paul et al., 2018).

While the collective evidence from LMICs is geographically diverse, the evidence from HICs is mostly centred on the US and

the UK. Considering that P4P is implemented in many HICs, an evidence base representing a wider range of health systems is desirable. For example, decentralisation levels and financing arrangements are important contextual factors shaping how P4P affects systems and outcomes(Singh et al., 2021).

Additional and distributed funding in P4P schemes have contributed to securing medicines, and equipment in LMICs. Schemes in these settings therefore often enable a share of the incentive to be re-invested in service delivery, and are typically accompanied by financial reforms with increased decentralisation of health financing, resulting in greater provider autonomy over resources(Singh et al., 2021). This is often expressed as a positive component of P4P, since it allows for targeting local needs while increasing program ownership. Yet, whether facilities can use this autonomy often depends on national policies and their measures for equity and efficiency of scale. For example, when authorities on human resources or drug procurement practices are centralized nationally, facility bonuses may have no impact on hiring new staff or buying more drugs. The level of autonomous control and how this translates to local health outcomes is therefore considerably determined by contextual factors.

Studies of P4P in LMICs often also measured the impact of P4P on the utilisation of services, especially for maternal and

child health services which were often incentivised together with measures of process quality. Although, as noted by Witter et al.(2012), P4P schemes are rarely intended to improve utilisation, the focus on utilisation is telling of the role of context in P4P design. Utilisation is also a function of demand-side issues, which supply-side incentives can only partially address. While some P4P schemes have included demand-side incentives, this is uncommon in LMIC contexts. Still, there are examples of facilities using payments to stimulate demand by improving structural quality (e.g. facility bonuses earmarked for general improvements or equipment) or providing free food for mothers to attract them to a facility(Kadungure et al., 2021).

The approach to assessing the impact of P4P on quality differed between settings. Evaluations in HICs typically relied on administrative data, while studies in LMICs sometimes relied on "independently assessed" measures of quality, sometimes presenting a sizeable burden in terms of data collection(Gergen et al., 2017; Kovacs et al., 2021). These assessments were often conducted by external funders or agencies, an external agent. The assessments are done e.g. by using clinical vignettes and/or rely on quality checklists to supplement quantitative reporting by facilities. Despite several studies from HICs suggesting that providers sometimes respond to P4P by attempting to "game" the system, providers' performance have mostly been taken at face

value, while there might be a reason to learn from the LMIC approach of relying more on external performance assessment when evaluating the impact of P4P. Yet, even when a third party conducts an evaluation it likely does so on behalf of the external funder or purchasing agent, thus increasing the risk of selection bias and partiality.

Across both settings, it is a limitation that few studies have examined the cost-effectiveness of P4P(Meacock et al., 2014; Paul et al., 2021). Given the lack of consistent evidence about the effectiveness of P4P for improving health outcomes, funders should consider whether other ways of spending scarce health care resources could generate greater health improvements. In Tanzania, for example, the estimated cost of administering a P4P scheme amounted to a third of the total economic cost of the programme, exceeding the cost of the financial incentives(Borghi et al., 2015). It is arguable whether the additional cost of administering incentives generates sufficient service improvements compared to what might be achieved by for example providing additional facility resources(Zeng et al., 2018).

The perspective of the studies also differs. While most studies in LMICs were looking at a relatively short-term, there are more studies in HICs that study the longer-term effects of P4P, possibly due to data availability and a longer P4P history in HICs. These studies have contributed to the realisation that P4P

was not always effective in the longer term. However, studies emerging from LMICs have also looked at longer-term effects, finding evidence of attenuation of effects over time (Borghi et al., 2021; Falisse et al., 2015), as well as improvements(Rajkotia et al., 2017). More such studies are needed to aid our understanding of the temporal dynamics of P4P.

Design characteristics

Schemes in LMICs often provide stronger incentives for performance than HIC schemes. First, LMIC schemes in predominantly provide a relatively large incentive for performance (greater than 5% of the usual salary or budget), while this is less common for schemes in HICs. Second, in HICs, often the organisational level is targeted by P4P, while in LMICs payments are typically redirected to the individual health care worker who therefore experiences a direct financial impact of better performance. Although HIC organisations often will "pass on" the incentive to individual staff this will most likely be in a weaker form, for example as greater budget control or career incentives rather than personal income. Paying individuals reduces the risk of free-riding(Kandel and Lazear, 1992) and paying closer to the level of care has been shown to be associated with greater improvements in performance(Kristensen et al., 2016). However, paying individuals can lead to unintended consequences if outcomes

constrained by inequality in facility resources or are the result of a team effort which may be undermined by giving too strong rewards for individual performance(Lazear, 1989). Third, the higher payment frequency in LMICs strengthens the link between performance and payments, which may create stronger incentives for performance(Emanuel et al., 2016).

In practice, many LMIC schemes are blended, with a proportion of payments going to facility staff, a proportion going to the facility for improvements, and possibly a proportion going to district managers or programme evaluators. Staff are incentivised to meet targets, facilities to improve services, and managers/evaluators to ensure good reporting. P4P thus tackles a greater number of issues in LMICs. This could be seen as a more holistic approach but it means that bottlenecks in one place can cause poor results in other aspects of the scheme.

Tournaments are more frequent in the HIC setting. Such schemes, where the performance of other providers sets the benchmark for performance, may create a stronger incentive due to uncertainty about the performance needed to receive payment(Lazear and Rosen, 1981). But it has been suggested that providers far from the top may feel discouraged to improve performance(Brown, 2011).

Finally, the reliance on donor-initiation and -funding may generate uncertainty about whether and how P4P in LMICs are

continued in the long run with implications for ownership and implementation(Gautier and Ridde, 2017).

Lessons from looking across HLMIC

The relationship between scheme design and effectiveness is poorly understood for many design features in any setting. Therefore, the mere existence of design differences does not alone suggest scope for the introduction of new design practices across settings. On the other hand, in some cases, a theoretical or empirical basis exists for suggesting more widespread use of certain designs. Among those is the frequency of payment, where high-frequency payments are more widely used in LMIC schemes. Future P4P designs in HICs could experiment with a higher payment frequency to enforce the link between performance and payment. These would be feasible for process indicators where the change is measurable immediately, while for outcomes, any increase in frequency is restricted by the lag between behaviour and outcome observability.

The relatively small size of payments in HIC as compared to LMICs is worthy of note. Recent research highlights that schemes paying a higher bonus increase performance more(Fardousi and al., Forthcoming) Schemes in HIC may therefore see improvements with larger bonuses. Conversely, the same research found that providers in poorer areas increased performance even in response to relatively small incentives, suggesting that also smaller

incentive sizes could lead to performance improvements in LMIC (ibid), which may increase the efficiency and financial sustainability of these schemes. That said, this is only an attractive option if measured performance is associated with true improvements in health outcomes. Moreover, small incentives in LMIC contexts may represent a significant income improvement, explaining why smaller bonuses lead to improvements. In addition, staff in LMICs may be more reliant on bonuses for basic income 'top-ups'. This difference to HICs highlights how contextual factors impact P4P design expectations and performance in different settings, and explains counter effects despite incentives in cases of high workloads, low staff numbers, inadequate training, and/or lack of equipment(Kadungure et al., 2021; Singh et al., 2021).

In some LMIC settings where resources are limited and a higher proportion is channelled through health facilities via P4P, P4P may end up partially serving the purpose of resource allocation. Notably, sometimes when the P4P implementing agent is not a local government agency, this process may effectively generate a governance/purchaser split within local health system management. In a HIC setting, the availability of resources is typically taken for granted, and the introduction of P4P is not in the same way a necessity for securing that high-quality care delivery can take place. Equally, in these settings, providers typically already benefit from substantial financial autonomy,

and the introduction of P4P does not enhance provider autonomy in the same way.

Reliance on international rather than government funding in LMICs may affect scheme credibility and sustainability, and donor conditionalities may limit the tailoring of design to the local context —potentially affecting performance. Thus, the difference in P4P design in LMIC settings may be best explained by the use of P4P often being promoted by donors. Unlike in HIC settings, for LMICs, there is a critical literature arguing that P4P is often viewed as donor-driven and not well integrated with national strategies and ownership(Barnes et al., 2015; Gautier and Ridde, 2017). As a result, donors and international bodies may examine the possibility of investing in health systems more generally, and encourage countries themselves to initiate and design P4P schemes they consider suitable for their context. Indeed, recent research suggests that direct facility financing, whereby the transfer of funds to providers are not tied to performance, but performance is tracked alongside financial flows, may provide an efficient alternative to P4P in LMIC. Decision-makers should however consider that while direct facility funding may bring similar improvements, P4P is more easily targeted to specific goals, and the relevance of either approach likely depends on the state of the health system in question and objectives of investing in performance(de Walque et al., 2022).

In both settings, systems to collect performance information may not be available prior to scheme implementation. But in LMICs there is a greater tendency to see the introduction of P4P as a tool for enhanced accountability with strengthened routine health data collection often accompanying the introduction of P4P(Borghi et al., 2015). However, the costs of performance data monitoring in LMIC have the potential to undermine the efficiency of these schemes(de Walque et al., 2022) as well as leading to parallel information systems not well integrated with existing information systems(Antony et al., 2017). In HICs, well-developed information systems often exist already, and accountability and information gathering are rarely goals of P4P implementation in their own right, because P4P is often in addition to existing information and payment systems.

CONCLUSION

For the past twenty years, health system investments in HMLICs have used P4P as a tool to improve the quality of care. Although the knowledge base on whether and how P4P works to improve performance is increasing in both quality and quantity in both income settings, it is rare that researchers attempt to take stock and draw lessons from across settings. As a result, there are missed opportunities to better understand what works in P4P, where, why, and under what circumstances. A better overview of evidence from multiple resource settings therefore helps to

clarify how best to use P4P in health care and to determine whether it is the right tool for any specific setting. In particular, better comparative knowledge helps us respond to key questions about what baseline or system readiness conditions are crucial for effective P4P implementation, the optimal amount, level and incentivising frequency for P4P bonuses, how financing conditionalities (or sources) moderate programme performance, as well as having a better understanding of what range of issues/ services P4P can be reasonably expected to address, and indeed where alternative financing and purchasing options may better achieve desired goals. These are important and lingering questions, which would benefit from greater comparative P4P research.

Competing interests

The authors have no competing interests to declare.

Funding

All authors except FM were supported by a Health Systems Research Initiative grant [MR/P014429/1] jointly funded by the U.K. Medical Research Council, Wellcome Trust, Economic and Social Research Council, Department for International Development

SRK was supported by the NIHR Imperial Patient Safety Translational Research Centre.

LA was funded by the UKRI MRC Fellowship MR/S022554/1

RK was supported by a Wellcome Research Fellowship in Humanities and Social Science 219744/Z/19/Z

Acknowledgements

We acknowledge the presenters and participants at the session "Pay for Performance: Drawing Lessons From Across High, Low and Middle Income Settings" organised by the first author for the Congress of the International Health Economics Association in Basel, Switzerland, July 15, 2019. We also acknowledge the presenters and participants in the one-day workshop at the Fifth Global Symposium on Health Systems Research: Advancing Health Systems for All in the Sustainable Development Goals Era in Liverpool, U.K. in October 2018 organised by the PEMBA team.

Commented [A2]: Needs blinding?

REFERENCES

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.