A realist review to assess for whom, under what conditions and how pay for performance programmes work in low- and middle-income countries

Neha S. Singh a,*, Roxanne J. Kovacs a, Rachel Cassidy a, Søren R. Kristensen b, c, Josephine Borghi a, Garrett W. Brown d

a Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK
b Centre for Health Policy, Institute of Global Health Innovation, Imperial College London, London, UK
c Danish Centre of Health Economics, University of Southern Denmark, Odense, Denmark
d School of Politics and International Studies (POLIS), University of Leeds, Leeds, UK

ARTICLE INFO

Keywords:
Realist synthesis
Realist review
Pay for performance
Performance-based financing
Maternal health
Child health
Low- and middle-income countries

ABSTRACT

Pay for performance (P4P) programmes are popular health system-focused interventions aiming to improve health outcomes in low-and middle-income countries (LMICs). This realist review aims to understand how, why and under what circumstance P4P works in LMICs. We systematically searched peer-reviewed and grey literature databases, and examined the mechanisms underpinning P4P effects on: utilisation of services, patient satisfaction, provider productivity and broader health system, and contextual factors moderating these. This evidence was then used to construct a causal loop diagram. We included 112 records (19 grey literature; 93 peer-reviewed articles) assessing P4P schemes in 36 countries. Although we found mixed evidence of P4P’s effects on identified outcomes, common pathways to improved outcomes include: community outreach; adherence to clinical guidelines, patient-provider interactions, patient trust, facility improvements, access to drugs and equipment, facility autonomy, and lower user fees. Contextual factors shaping the system response to P4P include: degree of facility autonomy, efficiency of banking, role of user charges in financing public services; staffing levels; staff training and motivation, quality of facility infrastructure and community social norms. Programme design features supporting or impeding health system effects of P4P included: scope of incentivised indicators, fairness and reach of incentives, timely payments and a supportive, robust verification system that does not overburden staff. Facility bonuses are a key element of P4P, but rely on provider autonomy for maximum effect. If health system inputs are vastly underperforming pre-P4P, they are unlikely to improve only due to P4P. This is the first realist review describing how and why P4P initiatives work (or fail) in different LMIC contexts by exploring the underlying mechanisms and contextual and programme design moderators. Future studies should systematically examine health system pathways to outcomes for P4P and other health system strengthening initiatives, and offer more understanding of how programme design shapes mechanisms and effects.

1. Introduction

Pay-for-performance (P4P), or the provision of financial incentives to healthcare providers based on pre-specified performance targets, is an approach with substantial variation in its design and implementation that has been adopted in numerous low- and middle-income countries (LMICs) with the aim of increasing service coverage (Remmans et al., 2016, 2017). In these settings, P4P often involves a package of reforms in addition to the financial incentives, including, for example, a shift to electronic health information systems, a new system of performance and data verification, and increased financial decentralisation. Moreover, P4P has recently been coupled as a key ‘entry point’ for strategic purchasing with the explicit aim to strengthen LMIC health systems toward the achievement of Sustainable Development Goal 3.8 for Universal Health Coverage (Mathauer et al., 2017). However, there is little evidence on how P4P affects health systems as a whole (Borghi et al., 2018). Existing systematic reviews largely examine the effectiveness of P4P initiatives in relation to performance outcomes rather than focusing on the mechanisms from which these outcomes have been delivered or how various programme and contextual factors moderate that delivery. For...
example, a systematic review of the effect of P4P programmes in LMIC reported that the evidence base was too limited to draw broad conclusions and that greater understanding of how incentive design impacts programme effectiveness was needed (Witter et al., 2012). More recent reviews of P4P effects on quality of care (Das et al., 2016) and on HIV indicators (Suthar et al., 2017) reported programme effects only in relation to these outcomes. One review focused on extracting policy recommendations from studies documenting the effects of P4P on the health system, but it did not explore the mechanisms underpinning these effects nor their contextual variation (Renmans et al., 2016). Despite some emerging evidence of the health system effects of P4P, the traditional focus within P4P systematic reviews and empirical research remains on programme outcomes, and there remains a considerable evidence gap about how P4P works, under what conditions, as well as what P4P processes best support health systems strengthening (Roland, 2012; Epstein, 2012; Ssengooba et al., 2012).

In response, our study, following Pawson and colleagues (Pawson, 2006; Pawson et al., 2005), is the first systematic realist review to assess the effects of P4P on health system inputs, focusing on financing, governance, medical commodities and human resources. This review aims to better investigate causal pathways to outcomes, thus helping to determine the ‘active ingredients’ of P4P programmes (e.g. by identifying which health system levers i.e. drugs, workforce, etc. that P4P affects), and if and how these vary by setting in LMICs. A realist approach enables the inclusion and synthesis of a much broader set of evidence than systematic reviews, including qualitative methods which are specifically designed to address ‘how and why’ questions (4). Unlike experimental and quasi-experimental studies, which are a mainstay of P4P systematic reviews, a realist approach assumes that complex interventions do not operate in isolation. Instead, realist approaches recognise that these interventions function within complex social systems, and go through numerous iterations including design, implementation and evaluation, during which time the intervention interacts with people, hierarchies, socio-cultural structures, and other factors, which are rarely linear nor result in the same outcomes in different setting.

### Table 1
Methodology for completing the systematic realist review (Molnar et al., 2015).

| Steps                              | Task(s)                                                                 |
|------------------------------------|-------------------------------------------------------------------------|
| 1 Develop initial programme theory | • Search for initial theories • Consult with experts                     |
| 2 Search strategy                  | • Search electronic peer-reviewed and grey literature databases using keywords and MeSH terms |
| 3 Select and appraise documents    | • Use inclusion and exclusion criteria to screen for relevant abstracts, articles and reports |
| 4 Extract data                     | • Use standardised Excel tool to extract relevant data                   |
| 5 Analysis and synthesis process   | • Search reference lists by hand for additional potentially relevant articles and reports |
| 6 Present and disseminate revised programme theory | • Analyse data for content and outcome pattern; and synthesise mechanisms |

Fig. 1. Initial programme theory explaining the causal pathways of the health systems effects of payment for performance programmes. Source: Borghi et al., 2018; adapted from Anselmi et al., 2017b.
The overall aim of this realist review is to help researchers and policy makers understand how and why P4P programmes implemented in LMICs result in intended or unintended outcomes and how the context within which they are implemented affects this. Specifically, the following three research questions guided the review:

1. What are the health system effects of P4P in LMICs?
2. What were the identified contextual factors and mechanisms influencing the outcomes?
3. How and why do the contextual factors affect the outcomes?

2. Methods

2.1. Study design

The study was conducted in six steps, as described in detail by Borghi et al. (2018) and summarised in Table 1 and below. According to Pawson’s proposed methodology for realist reviews, our research was not linear, but iterative (Pawson, 2006). This review adopts the Realist and Meta-narrative Evidence Syntheses Evolving Standards (RAMESES) as developed by Wong et al. (2013).

2.2. Step 1 – develop an initial programme theory

The initial programme theory was based on relevant P4P theories of change presented in the World Bank’s P4P impact evaluation toolkit (Vermeersch et al., 2012) and a P4P study in Tanzania (Anselmi et al., 2017a), as well as theories of change developed during an international workshop on the health system effects of P4P in Dar es Salaam, Tanzania in November 2015 and a Researcher Links UK-Mexico workshop in April 2015. The initial programme theory was presented to a policy and academic audience at the Fourth Global Forum for Human Resources in Medicine 2015. The initial programme theory was presented to a policy and academic audience at the Fourth Global Forum for Human Resources in Medicine 2015.

2.3. Step 2 – search strategy

The scoping review was conducted using the Medline, Embase, and EconLit databases to address the research questions. Additionally, as grey literature is a relevant source of information for both P4P and realist reviews, evaluation reports or policy documents published by LMIC governments, international organisations, non-governmental organisations and consultancy firms were also included by searching Google Scholar, Emerald Insight, and websites of key stakeholders including: World Bank, WHO, Cordaid, Norad, DfID, USAID and PEPFAR.

Search terms related to P4P, LMICs and health systems pillars were combined in the search strategy and were first developed for Medline and then adapted to the other databases (Supplementary File 1). The search period covered January 1, 1995 to March 1, 2019. The relevance of retrieved articles was assessed according to the inclusion and exclusion criteria outlined below.

In order to be included, the evidence had to meet the following criteria: (i) the exposure (intervention) was a P4P intervention targeting providers with financial incentives varying according to the achievement of quantitative health service delivery targets and/or quality of care targets. When performance was linked to quantitative outputs, it was related to “selected healthcare services” as such criterion enabled discrimination between P4P and fee-for-service (FFS) mechanisms; (ii) the study isolated (or attempted to isolate) the effects of P4P policies from broader policy reforms; (iii) incentives were allocated to public and/or non-public providers or institutions at primary, secondary or tertiary levels; and/or managers and administrators; (iv) pilot projects were included in the study; (v) the study outcome was either a quantitative or qualitative measure (or both) of the impact of the P4P initiative on one or more health system functions described in the programme theory (Fig. 1); (vi) the papers reported studies that either collected or intended to collect primary data. Where the papers referred to a different source of evidence for primary data (e.g. in the case of systematic reviews), the primary source of information was retrieved and explored; and (vii) the intervention was implemented in an LMIC, as defined by the World Bank (2015).

The review excluded documents: (i) which were evaluations of the “potential” implementation of P4P strategies not yet in place; (ii) which only measured health outcomes; and (iii) published in a language other than English, French and Spanish.

Reference lists of all publications included in the final review were explored for additional relevant literature. We also consulted P4P experts to identify additional relevant literature. The literature review search ended at the point of saturation when the research yielded no further new sources of information. References were compiled in Endnote X9 software.

2.4. Step 3 – selecting and appraising documents

The identification and selection of studies in this review was based on relevance and rigour, and their study’s ability to enrich the C-M-O system. Financial rewards for meeting targets may be invested in improving facility infrastructure and drug supply, which impacts on facility resource levels. In turn, this improves the work environment enhancing worker motivation to deliver better and more affordable services, increasing patient demand. There are a number of ways in which P4P is hypothesised to influence quality of care within the theory of change, in terms of structural quality (improved infrastructure, and drug availability), process quality (improved patient-provider interactions; adherence to clinical protocols); and outcomes (improved patient satisfaction). However, P4P can also result in unintended consequences such as mis-reporting performance (gaming), a displacement of effort away from un-incentivised services, and positive spillover effects (Binyaruka et al., 2015a; Turcotte-Tremblay et al., 2020).
configuration underlying the programme theory (Fig. 1) (Pawson, 2006). Accordingly, literature was reviewed to examine whether the programme theory (Fig. 1) was born out by evidence, and to identify knowledge gaps where evidence was weak. Where there was conflicting evidence on a given component, we explored contextual and scheme design differences that may account for differences in findings.

2.5. Step 4 – extracting data

Data were extracted and recorded in an Excel database. Initially data extraction focused on research objectives, sample size, and study subjects. Next, data were extracted to understand how, why and under what circumstances P4P effects health system pillars. Finally, all data were indexed and linked to relevant program theory as described in Fig. 1. Additionally, we appraised study quality of included articles. Quality criteria differ for qualitative and quantitative methods, because of their differing underlying assumptions, methodologies and aims. New tools have been developed for systematic reviews which include both qualitative and quantitative studies or mixed methods studies. The Mixed Methods Appraisal Tool (MMAT) was chosen for this review because it can be used quickly and reliably and includes items appraising qualitative methods (Renmans et al., 2016), quantitative methods, and (Mathauer et al., 2017) mixed methods (i.e., the approach to combining qualitative and quantitative components) (Hong et al., 2018). Six subset of Items are worded to reflect good quality and each study is rated as “yes,” “no,” or “cannot tell” for each applicable item. Because the MMAT items reflect quality of reporting as well as quality of study design, no attempt was made to obtain further details about the studies under review by contacting authors.

The first and second author familiarised themselves with the MMAT by studying the tutorial before applying the MMAT to the included literature. The mixed methods design and the design of each qualitative and quantitative component was also recorded, using the definitions supplied on the MMAT. All ratings were entered into an Excel spreadsheet and used to calculate the proportion of studies meeting each quality criteria.

2.6. Step 5 – analysing and synthesising data

Articles were screened by three study team members. Each study was read and synthesised by the first author. The second author carefully read papers and assessed if the evidence was used properly in the synthesis. The first author discussed with the study team to assess whether emerging findings supported, refuted or reinterpreted the initial programme theory (Fig. 1). By doing so, the authors critically appraised the contribution of each study to the initial theory, as well as synthesised across findings. The next refined theory, defined as a middle-range theory (Pawson et al., 2005), was finalised to highlight the links between contextual factors, mechanisms and outcomes of P4P interventions. The final synthesis was agreed upon by all authors.

2.7. Step 6 – presenting and disseminating revised programme theory

Finally, we presented and further refined our revised theoretical findings in the middle-range theory with relevant stakeholders and experts including at the Fifth Global Symposium on Health Systems Research in October 2018 and at a P4P all partners workshop in Maputo, Mozambique in March 2019. We present our final revised programme theory in a causal loop diagram (CLD) (Fig. 2), because the mechanisms for how P4P alters patient and provider behaviour, and the contextual factors that contribute to the success of the initiative identified in the realist review could not be adequately represented using a linear
Fig. 3. Detailed views of the programme theory visualised as a causal loop diagram: (A) Demand- and supply-side mechanisms through which P4P programmes affect utilisation of health services in LMIC (community outreach - R4, provider effort - R3, R4, patient trust - R3, R4); (B) Demand- and supply-side mechanisms through which P4P programmes affect utilisation of health services in LMICs (health worker adherence to clinical guidelines - R8, R9, quality of care - R5, R9, user fees - R5). Mechanisms through which P4P programmes influence patient satisfaction in LMICs (facility improvements - R12, user fees R5). Mechanisms through which P4P programmes influence healthcare providers and the broader health system (provider motivation - R2, R8); (C) Demand- and supply-side mechanisms through which P4P programmes affect utilisation of health services in LMIGs (quality of care - R6, R7, provider effort - R6, R7, R10, R2). Mechanisms through which P4P programmes influence patient satisfaction in LMIGs (increased availability of services - R7, provider responsiveness - R6, R7). Mechanisms through which P4P programmes influence healthcare providers and the broader health system (misreporting of information - R11, B1).
diagram. CLDs can be used to provide a blueprint for complex systems, e. g. health systems (Chang et al., 2017; Cassidy et al., 2019), representing important, non-linear feedback and relationships that further our understanding of how these systems operate (Sterman, 2000). CLDs use arrows with polarity to indicate causal influences between variables; health workers that were awarded incentives felt motivated to achieve targets (positive causal link), an increase in user fees resulted in reduced patient satisfaction in delivery of healthcare (negative causal link).

Supplementary file 2 provides detailed guidance on how to read a CLD. We represented delays in influence by two dashed lines across an arrow, e.g. community outreach, over time, resulted in increased trust between patients and providers (Fig. 2). Feedback loops are represented by numbered circular arrows and show either reinforcing (R) or balancing (B) behaviour. Health workers who were awarded incentives felt motivated to achieve targets and consequently adhered to clinical guidelines (captured in loop R8, Fig. 2) is an example of a reinforcing...
loop, exhibiting amplified, spiralling feedback. Loop B1 (Fig. 2) describes how an increase in provider motivation may also have resulted in providers retaining drugs to avoid stockouts (an incentivised target), thus reducing the availability of drugs for patients; the negative causal link here provides a dampening effect and stops the loop spiralling indefinitely.

The CLD shows mechanisms and mediators of programme effect that were identified in the original theory of change (grey arrows and variables), identified in the realist review but not in the original theory of change (black) and were captured in both the original theory of change and the review (green). Key mediators or contextual factors of programme effect that are not part of a causal loop are shown as notes (N) in the diagram (e.g. N1, Fig. 2).

3. Results

3.1. Study characteristics and design

We identified 38,234 records and screened 374 papers for the full-text review (Fig. 4). Finally, 117 records ((Sengooba et al., 2012), (Anselmi et al., 2017a), (Engineer et al., 2016; Berman, 2015; Celhay et al., 2015; Gertler and Giovagnoli, 2014; Rubinstein et al., 2009; Rahman et al., 2017; Rob and Alam, 2013; Bowser et al., 2013; Antony et al., 2017; Lagarde et al., 2015; Bertone and Meessen, 2013), (Bonfrer et al., 2014a, 2014b; Falisse et al., 2012, 2015; Rudasingwa et al., 2015; Ir et al., 2015; Jacobs et al., 2010; Khim et al., 2018; Matsuoka et al., 2014; Van de Poel et al., 2016), (De Walque et al., 2017; Zang et al., 2015; Ding et al., 2013; Fan, 2017; Li et al., 1982; Liu and Mills, 2005; Millar et al., 2017a; Powell-Jackson et al., 2015; Sun et al., 2016; Wang et al., 2011), (Wu, 2014; Yao et al., 2008; Yip and Eggleston, 2001; Yip et al., 2010, 2014; Zhang et al., 2017a; Cercone et al., 2005; Hindle and Kalanj, 2004; Fox et al., 2014; Huillery and Seban, 2017), (Soeters et al., 2011a; Zeng et al., 2013, 2018; Bernal et al., 2018; Mohanan et al., 2019; Singh and Masters, 2017; Chernichovsky and Bayulken, 1995; Aryankhesal et al., 2013, 2015; Sarikhani and Lankarani, 2013), (Mena et al., 2015; Saleh et al., 2013; Morgan, 2011; Lohmann et al., 2017; McMahon and De Allegri, 2016; McMahon et al., 2016, 2018; Wilhelm et al., 2016; Schuster et al., 2016, 2018), (Barnes et al., 2014; Dusburg et al., 2016; Ye et al., 2014; Regalia and Castro, 2007; Bhatnagar and George, 2016; Ogundeji et al., 2016; Witter et al., 2011; Peabody et al., 2011, 2014, 2017), (Wagner et al., 2018, Basinga, 2009; Basinga et al., 2010, 2011; de Walque et al., 2015; Gerrler and Vermeersch, 2012; Janssen et al., 2015; Kalk et al., 2015; Kantengwa et al., 2015; Lannes, 2015; Lannes et al., 2016; Meessen et al., 2006, 2007b; Ngo et al., 2017; Priedeman Skiles et al., 2012; Rusa et al., 2009; Shapira et al., 2017; Sherry et al., 2017; Skiles et al., 2015; Bertone et al., 2016), (Bertone and Witter, 2015; Binyaruka and Borghi, 2017; Binyaruka et al., 2015b, 2018a; Borghi et al., 2015; Chimbutu et al., 2014, 2015, 2016; Mayuman et al., 2017a; Olafsdottir et al., 2014), (Songstad et al., 2012; Gormus, 2015; Miller et al., 2014; Chansa et al., 2015; Friedman et al., 2016; Shen et al., 2017; Vergeer and Chansa, 2008; Feldacker et al., 2017; World Bank. Rewarding Pro, 2016; Bodson et al., 2018; Paul et al., 2018; Paul et al., 2014; Turcotte-Tremblay et al., 2018; Turcotte–Tremblay et al., 2017)) met the inclusion criteria comprising of 19 grey literature and 98 peer-reviewed articles. The included studies assessed P4P schemes in 36 countries, across Sub-Saharan Africa (n = 23 countries), Latin America and the Caribbean (n = 6 countries), East Asia and the Pacific (n = 4 countries), South Asia (n = 4 countries), the Middle East and North Africa (n = 2 countries), and Europe and Central Asia (n = 2 countries) (Fig. 5). Study designs comprised of 54 quantitative
impact evaluations (39 quasi-experimental studies, 15 randomised controlled trials (RCTs)), 33 qualitative studies, 10 observational studies, 15 mixed methods studies, 1 study protocol, and 2 reviews. Supplementary file 3 presents key characteristics of the included studies.

3.2. Study quality

The MMAT results suggest that the mixed methods studies were generally of good quality (Supplementary file 4). In the majority of studies, the mixed methods design was relevant to the research questions and the qualitative and quantitative components were integrated at some stage to address the research question. However, none of the studies acknowledged or reflected on the limitations associated with the integration of qualitative and quantitative data (or results) relevant to address the research question.

Quantitative RCT components scored high: all but one RCTs (n = 7) reported complete outcome data and all RCTs (n = 8) clearly described appropriate randomisation. However, the majority of RCTs (n = 7) did not clearly describe allocation concealment procedures. Non-randomised quantitative components were of high quality, particularly regarding the validity of measurements and the use of recruitment procedures to minimise selection bias. A majority of quantitative non-randomised components reported baseline comparisons between groups and controlled for relevant confounders, and used appropriate and valid measurements. Common weaknesses were identified in the qualitative components, including not giving appropriate consideration to the impact of the researchers or the wider context on the methods/findings. The frequency of “cannot tell” ratings for the qualitative components was also particularly high, e.g. there was often insufficient detail to evaluate data analysis procedures.

3.3. Context–mechanism–outcome configurations

This section presents findings on the potential health system effects of P4P programmes in LMICs and how context affects the mechanisms through which outcomes are achieved (i.e. C-M-O configurations). We focus on the outcomes highlighted in the initial programme theory: 1) utilisation of healthcare services; 2) patient satisfaction; and 3) P4P’s impact on healthcare providers as well as the broader health system. Findings from included studies were used to refine the initial programme theory, with the final revised programme theory presented in Fig. 2. Detailed views of the revised programme theory are presented in Fig. 3, in which loops have been predominantly grouped to describe behaviour around a common outcome.

3.4. P4P and the utilisation of health services

We found 32 studies reporting data on the effect of P4P programmes on the utilisation of healthcare services, of which 19 studies showed significant positive effects on at least one indicator (Anselmi et al., 2017a; Berman, 2015; Celhay et al., 2015; Gertler and Giovagnoli, 2014; Rahman et al., 2017; Rob and Alam, 2013; Jacobs et al., 2010; Matsuoka et al., 2014; Liu and Mills, 2005; Zhang et al., 2017a; Soeters et al., 2011a; Zeng et al., 2013; Regalia and Castro, 2007; Basinga, 2009; Basinga et al., 2011; Basinga et al., 2010; de Walque et al., 2015; Sherry et al., 2017; Binyaruka et al., 2015b; Soeters et al., 2011b; Singh et al., 2015), six studies reported non-significant effects on the utilisation of healthcare services (Engineer et al., 2016; Falisse et al., 2015; De Walque et al., 2017; Powell-Jackson et al., 2015; Wang et al., 2011; Skiles et al., 2015), and a further six studies whose design did not allow for the interpretation of positive or negative effects or correlations (Meessen et al., 2007a; Morgan, 2011; Schuster et al., 2016; Witter et al., 2011; Janssen et al., 2015; Meessen et al., 2006). Studies examined a range of outcomes such as care seeking for pre-natal care (number of ANC visits, timing of first ANC visit) (Berman, 2015), preventive care seeking (testing for HIV) (de Walque et al., 2015), facility delivery (Rob and Alam, 2013; Sherry et al., 2017) and immunisation coverage (Matsuoka et al., 2014). Positive programme effects on facility-based delivery were the most consistently reported in the literature, with less consistent effects on other indicators. The identified studies provided some level of evidence on eight potential demand- and supply-side pathways (i.e. mechanisms) through which P4P programmes affect the utilisation of health services in LMICs: 1) demand-side strategies including community outreach; 2) health worker adherence to clinical guidelines; 3) quality of care including patient-provider interaction, and access to essential drugs and equipment; 4) provider effort; 5) patient trust; 6)
facility autonomy, and 7) user fees. It should be noted that some of these effects are intermediate, so they affect domains which affect outcomes, i.e. they do not affect use directly. For example, availability of drugs affects access to services and trust, which in turn affects demand. Financial autonomy in itself will not affect demand, but it will affect health workers ability to deliver services and improve drug supply which will ultimately affect demand.

Several studies suggest that P4P schemes might improve care seeking by encouraging providers to invest more time in demand-side strategies to achieve incentivised P4P targets (Fig. 3A, Loop R4). For example, a qualitative study on the Plan Nacer programme in Argentina found that providers taking part in the programme conducted more home-visits to encourage pregnant women to seek ante-natal care and that, potentially as a result, eligible women sought care 1.5 weeks earlier (Berman, 2015). An important contextual factor that likely moderates this effect is cultural norms about care seeking (Fig. 3A, Note 1). A study in Rwanda found that because of health care seeking behaviour of women where they generally seek care only in the later stages of pregnancy, the local P4P scheme, which did not have a community outreach component to address this cultural norm, was unable to stimulate demand for ANC in the early stages of pregnancy (Basinga et al., 2011).

A number of studies suggest that P4P programmes could improve care seeking by increasing staff adherence to clinical guidelines (Fig. 3B, R8). In a quasi-experimental study in Rwanda, providers participating in a P4P scheme incentivising both coverage and quality of care delivered higher clinical quality care to pregnant women and children as measured by a standardised total quality score (Basinga et al., 2011). Facilities receiving P4P incentives also had improved access to skilled personnel, medical equipment and drugs (Fig. 3B, R9). However, overall levels of provider knowledge (Fig. 3B, N2) and staff capacity (Fig. 3B, N3) were identified as a key contextual moderator. For example, evidence from Malawi suggests that provider skills and educational level as well as institutional capacity in terms of staffing and other resources were at too low a level when the P4P scheme was introduced (McMahon et al., 2016), thus limiting programme effects, a finding that is also echoed in studies from Uganda (Ssengooba et al., 2012) and Tanzania (Chimbutu et al., 2014).

A small number of studies suggest that P4P schemes can improve quality of care by altering patient-provider interaction (Fig. 3C, R6, R7). The main mechanism identified is the amount of time providers spend with patients (Fig. 3C, R7). A mixed-methods study in Cambodia found that providers extended their working hours – thereby giving them more time to see patients (Matsuoka et al., 2014). Similarly, in a qualitative study in Rwanda a P4P scheme incentivised providers to hire more staff – thereby improving the ratio of patients to staff (Kalk et al., 2010). One of the contextual factors identified in the literature that likely moderates this relationship, is the availability of human resources (Fig. 3C, N3). Studies in Malawi (McMahon et al., 2016), Cambodia (Matsuoka et al., 2014) or Afghanistan (Engineer et al., 2016) found that when the ratio of patients to providers is heavily skewed, P4P is unlikely to allow providers to spend more time with patients. The attitude of providers towards patients was also identified as a mediator of programme effect on demand, with greater provider kindness during deliveries being an important pathway towards increased facility-based deliveries in Tanzania (Anselmi et al., 2017a).

A change in the availability of drugs and equipment in healthcare facilities was one of the mechanisms through which P4P affects quality of care (Fig. 3C, R9). For example, an RCT in Cameroon found that the availability of drugs, medical supplies and equipment was higher in facilities taking part in the P4P scheme and that vaccination coverage and contraceptive use also improved (Fig. 3B, R9) (De Walque et al., 2017) as a result of this. However, the literature suggests improvements in structural quality are not always associated with increased demand for care. One study in Tanzania reported that facilities in rural areas were more likely to observe improvements in the availability of drugs and medical supplies. Other potentially relevant contextual factors not explicitly reported on within the literature include the financial autonomy of facilities (i.e. to decide how to use funds) and broader consultation fees (Fig. 3B, R5). Indeed, it is a priori plausible that even in settings where patients’ willingness to pay for care improves because of improved structural quality, their ability to pay – and therefore the level of care seeking, remain unchanged. Supportive of this, a number of studies reported that the effects of the P4P programme were greater among wealthier groups with greater ability to pay (Van de Poel et al., 2016; Lannes et al., 2016; Bonfrer et al., 2014c). However, it is also plausible that if drugs are available at the facility, this can reduce the cost of care seeking as it mitigates the purchase of drugs from private pharmacies, in such cases the scheme may preferentially benefit poorer groups, as was reported in Tanzania (Binyaruka et al., 2018b). Therefore, programme effects on care seeking, and their distribution across socio-economic groups, will depend on financing arrangements, the affordability of drugs, and existence and extent of other patient fees.

Included studies also suggest that P4P schemes could improve care seeking by focusing provider effort on specific activities (Fig. 3A, R3, R4) and (Fig. 3C R6, R7, R10). For example, a quasi-experimental study in Rwanda indicates that because healthcare workers received large financial bonuses for jointly testing couples for HIV (US$4.59 per couple/partner jointly tested, compared to US$0.92 per new individual tested), they focused on this activity, and the number of jointly tested couples increased (de Walque et al., 2015). Nonetheless, P4P schemes are not always able to focus healthcare worker effort on incentivised activities, even when large financial rewards are provided. The type of activity that is incentivised together with available staffing levels (Fig. 3C, N3) will moderate the programme effect. For example, providers in Rwanda did not improve performance for time consuming activities or those requiring substantial effort, such as the provision of modern contraception (Basinga, 2009; Basinga et al., 2011; Basinga et al., 2010). The amount of time providers spend recording and verifying data (Fig. 3C, B2) will also impact provider ability to focus on incentivised clinical tasks, with a higher administrative burden reducing provider response to incentives (Antony et al., 2017; Millar et al., 2017b). Furthermore, it seems that for delivery care, the level of baseline performance at the facility can moderate programme effects on providers, with greater effects reported in lower performing facilities in Tanzania (Binyaruka et al., 2018b).

Some studies indicate that a potential channel through which P4P programmes influence care seeking is by altering the level of patient trust (Fig. 3A, R3, R4). A study in Bangladesh found that, because of a referral system for complicated deliveries introduced by a local P4P scheme, the level of confidence and trust in healthcare providers improved among women – thereby increasing demand for facility deliveries (Fig. 3A, R3) (Rahman et al., 2017). Similarly, a study conducted in Tanzania (Anselmi et al., 2017a) indicates that improved access to essential drugs in healthcare facilities (specifically oxytocin), associated with the local P4P scheme, might have increased patient trust, though this variable was not explicitly measured.

Some studies indicate that P4P could affect care seeking by increasing the financial autonomy of facilities in terms of access to and use of funds (Fig. 3B, N4) (Van de Poel et al., 2016; Wilhelm et al., 2016). For instance, a mixed-methods study in Cambodia found that facilities taking part in the P4P scheme were able to independently procure medicines, vaccines and consumables (Fig. 3B, R9), which improved facilities’ ability to plan ahead and increased immunisation coverage (Matsuoka et al., 2014). However, such an increase in facility autonomy will likely only influence care seeking in contexts where facilities have little autonomy at the outset – which is not usually the case in higher income settings such as China. In some P4P schemes provider autonomy was not sufficiently embedded within the programme design which limited programme effects on care seeking. For example, in Malawi it was reported that adaptations to the P4P scheme to bolster autonomy in relation to procurement and other local priorities could not be made. A District Management Officer noted: “… let me tell you, I
wanted them to bring us a skeleton [of P4P]. A skeleton and then together we would put on some flesh. Build something together. But they came from Lilongwe and they brought a prince. He could not be touched, nothing could be changed or altered” (Wilhelm et al., 2016). P4P schemes can enhance provider autonomy by building in design features to enable this; however, programme effects and autonomy will be constrained in contexts where procurement is centrally organised and local level procurement decisions (to various degrees) are not permitted despite the facility having a surplus of funds via P4P.

Finally, several studies suggest that P4P schemes could influence the level of care seeking by lowering user fees (Fig. 3B, R5). A study in Cameroon found that out-of-pocket spending, particularly on consultation fees, was lower in facilities taking part in the P4P scheme, and that service utilisation increased (De Walque et al., 2017). Similarly, a study in Tanzania reported reductions in the probability of paying for delivery care, which mediated the programme effect on the rate of institutional deliveries (Anselmi et al., 2017a). Furthermore, a study from DRC provides evidence for two contextual factors moderating programme effect on user fees (Remnans et al., 2017): the level of user fees, as if they are very low, then P4P is less likely to impact them; and (Remnans et al., 2016) the level of incentive (i.e. programme design) and extent to which it would offset lost revenue from reduced fees, with providers being less likely to reduce user fees to achieve targets where incentive payments were insufficient to offset the lost user fee revenue from reduced fees (Maini et al., 2017). A number of studies introduced user fee removal policies alongside P4P, and in such cases programmes would have no effect on user fees. Though the introduction of demand side incentives such as user fee subsidies or maternity care vouchers can enhance programme effects on service coverage (Van de Poel et al., 2016). Although there was no evidence of this in our review, the wider literature suggests that even where user fees are reduced, effects on demand will be lower in remote areas where distance/geographical barriers limit access (Sengooba et al., 2012).

P4P and patient satisfaction (perceived quality of care).

Overall, only seven studies offer some level of evidence on the potential effect of P4P programmes on patient satisfaction, i.e. a measure of perceived quality of care, of which five studies showed significant positive effects (Rob and Alam, 2013; De Walque et al., 2017; Cercone et al., 2005; Soeters et al., 2011a; Binyaruka et al., 2018b) and two studies reported significant negative effects (Wang et al., 2011; Binyaruka et al., 2015b). Patient satisfaction is measured with very little consistency across studies – which makes it challenging to compare findings between settings. Studies included in this review provide some level of evidence on four potential pathways (i.e. mechanisms) through which P4P programmes influence patient satisfaction in LMICs: 1) facility improvements; 2) increased availability of services; 3) user fees, and; 4) provider responsiveness.

Several studies suggest that one of the channels through which P4P schemes might affect patient satisfaction is via improvements in facility infrastructure, i.e. a measure of structural quality of care. For instance, a study using data from an RCT in Cameroon found that the local P4P scheme was associated with infrastructural improvements at the facility level (Fig. 3B, R12), improved access to drugs and equipment (Fig. 3B, R9) as well as an increase in measures of patient satisfaction (De Walque et al., 2017). A contextual factor that is frequently highlighted in the literature as moderating this mechanism is the quality of infrastructure at the start of the P4P intervention. For example, evidence from DRC (Soeters et al., 2013b; Binyaruka et al., 2014) indicates that in settings where the quality of facility infrastructure is very poor at the outset, potential improvements achieved by P4P schemes are not sufficient to translate into better outcomes.

The evidence also suggests that P4P schemes could improve patient satisfaction by increasing the availability of services. For example, a study in Costa Rica found that healthcare facilities taking part in the scheme extended their opening hours (Fig. 3C, R7) and offered home delivery of medication (Fig. 3C, R10), and that levels of patient satisfaction improved in these facilities (Cercone et al., 2005). Nevertheless, none of the studies reviewed examined what exogenous factors may or may not have influenced incentivised providers to increase their community engagement.

A number of studies indicated that P4P could improve patient satisfaction by reducing out-of-pocket payments (Fig. 3B, R5). For example, a study in Cameroon found that P4P facilities had lower user fee levels and higher patient satisfaction (De Walque et al., 2017). The literature highlights that patients’ perception of the quality of care is an important contextual factor moderating the effect of user fees on patient satisfaction. For instance, a study in Cambodia found that shortages of essential consumables and medicines were prevalent in facilities (Matsuoka et al., 2014), and that this reduced programme effect on user fees, as patients had to purchase supplies at private pharmacies. They argue that such shortages are “important from the service users’ viewpoint since they tend to perceive the availability of medical supplies and products to be a token of better health services” (p.463) (Matsuoka et al., 2014). As is also suggested by the broader literature on patient satisfaction (Christopher et al., 2009) patients’ perception of the quality of care provided in facilities is important for satisfaction.

Finally, some studies suggested that provider responsiveness is a potential channel through which P4P schemes might influence patient satisfaction (Fig. 3C, R6, R7). In a before-and-after study in Bangladesh, authors claim that in P4P facilities, “service providers were found more responsive to the clients and behaved well due to the incentive” (p.9) (Rob and Alam, 2013). As with community engagement, none of the studies we reviewed reported the existence of contextual moderators nor whether other policies external to the programme affected provider responsiveness or its perceived effects on patients.

3.5. P4P and healthcare providers as well as the broader health system

Of the studies included in the review, 24 papers offer some level of evidence on the potential effect of P4P programmes on healthcare provider motivation and behaviour and the broader health system (Anselmi et al., 2017a; Gertler and Giovagnoli, 2014; Bowser et al., 2013; Antony et al., 2017; Bertone and Meessen, 2013; Falisse et al., 2012; De Walque et al., 2017; Ding et al., 2013; Soeters et al., 2011a; Singh and Masters, 2017; Aryankhesal et al., 2015; Peabody et al., 2014; Kalk et al., 2010; Ngo et al., 2017; Bertone et al., 2016; Binyaruka and Borghi, 2017; Chimbhuta et al., 2016; Mayumana et al., 2017a; Gormus, 2015; Shen et al., 2017; Vergeer and Chansa, 2008; Feldacker et al., 2017; Abduvalieva et al., 2016; Ageypong et al., 2014). Studies discussed in this section examine a range of outcomes including, provider motivation and efficiency and governance, as well as spill-overs on non-incentivised activities, and misreporting of information. The mechanisms relevant to each of these outcomes are discussed below.

A small number of studies examined the potential effect of P4P schemes on provider productivity, all of which reported significant positive results (Meessen et al., 2007a; Bowser et al., 2013; Bertone and Meessen, 2013; Cercone et al., 2005; McMahon et al., 2018). Studies highlight that one of the potential pathways through which P4P schemes might affect provider productivity is by improving provider motivation (Fig. 3B, R2, R8). For instance, a qualitative study in Burundi found that in P4P facilities, provider motivation improved and providers worked “more, more quickly [and] more focused on objectives and priorities” (pp.853) (Bertone and Meessen, 2013). Similarly, a study in Belize reported that providers working in P4P facilities had significantly higher levels of productivity (Fig. 3B, R2), in terms of patients seen per hour (Bowser et al., 2013). Identified studies hint towards several programme schemes on provider productivity, all of which reported significant effects (Fig. 3B, N5). A prolonged lag time between reporting and rewarding outcomes likely delinks the relationship between rewards and productivity. The extent of payment delays will depend on programme factors, such as procedural
A number of studies provide evidence on the potential negative effects, or spillovers, of P4P schemes on non-incentivised activities (Cel-hay et al., 2015; Gertler and Giovagnoli, 2014; De Walque et al., 2017; Ngo et al., 2017; Binyuruka et al., 2017b; Feldacker et al., 2017b). For example, a qualitative study on a P4P scheme focusing on male circumcision (MC) found that the perceived quality of other services suffered (Fig. 3C, R1) (Feldacker et al., 2017b). Specifically, they found evidence suggesting that providers prioritised circumcision work over their other duties, even if they were more urgent. One respondent noted that “For instance, if there is a patient who needs to have a caesarean done and at the same time the doctor has to go out for MC, if the remains doing the C-section he doesn’t get any incentive for that C-section so he would rather go and do MC” (p.10 (Feldacker et al., 2017b)). These negative spillover effects are also reported in other settings, such as in Argentina, where prenatal care utilisation of non-beneficiary populations in clinics covered by P4P scheme decreased (Gertler and Giovagnoli, 2014) and in Rwanda where resources were shifted away from non-incentivised areas (Fig. 3C, R1) (Ngo et al., 2017b). A number of programme factors likely moderate this mechanism. For example, one might expect that neglect of non-incentivised services will be more likely in schemes that reward a limited set of activities, as it is easy for providers to identify non-incentivised tasks in their daily work. In addition, one might argue that providers are more likely to focus on activities that are heavily rewarded – although it is unclear at what point a payment becomes “too large” in this sense. One contextual factor affecting negative spillover effects in Tanzania, was the level of care; with spillover effects being more likely at lower level primary care facilities, as they had limited staffing to meet demand and thus prioritised incentivised over other services (Chimbhutu et al., 2014; Gormus, 2015).

Another common finding in the literature is that P4P schemes encourage misreporting of information and gaming (Fig. 3D, R11) (Engineer et al., 2016; Aryankhesal et al., 2015; de Walque et al., 2015; Kalk et al., 2010). A qualitative study in Iran found that stakeholders often misreported information and data to auditors (Aryankhesal et al., 2015). Similarly, a qualitative study in Rwanda found that “information was regularly distorted. Such distortion included the arbitrary and retrospective filling of forms” (p.186) (Kalk et al., 2010). They also found evidence for other forms of gaming such as not distributing the last box of a given drug to avoid stock-outs (which were disincentivised in the scheme) (Fig. 3D, Loop B1). The theoretical rationale for gaming is the desire to maximise P4P bonuses. In addition, the aforementioned study in Rwanda suggests that providers perceiving some indicators as inappropriate or lacking time to “do the job properly” (p.186) likely plays a role (Kalk et al., 2010). Evidence from Rwanda also suggests that misreporting could be inherent to P4P as authors note that “some people defended the view that such behaviour was incompatible with medical ethics, though it was fostered by the P4P approach.” (p.186) (Feldacker et al., 2017b). Several programme design factors were identified in the literature as being associated with misreporting and gaming within P4P schemes (Engineer et al., 2016). Although this was not supported by empirical studies within this review, from a theoretical point of view, gaming should be more likely in schemes that provide penalties rather than rewards (Fig. 3D, N9). This is because of loss-aversion (Kahneman, 1991), or people’s tendency to feel that the pain of losing is more powerful than the pleasure of an equivalent gain. A study in Benin (Antony et al., 2017) found that misreporting was more likely to occur where there was inadequate verification (Fig. 3D, N10) and in the absence of sufficient sanctions for misreporting. Furthermore, misreporting was not always intentional. This sometimes arose due to flaws in data reporting systems (Mayumana et al., 2017b), or a lack of human resources impacting on the quality and reliability of timely reporting (Fig. 3D, N12).

4. Discussion

Ours is the first systematic realist review to assess how and why P4P schemes work, how they help to strengthen health system capacity, and how they are perceived by facility managers. Our study was motivated by a growing recognition of the need to improve the implementation and design of P4P schemes in order to ensure that they achieve their intended outcomes. We conducted a broad and systematic review of the literature on P4P schemes, including both quantitative and qualitative studies, and found evidence of positive and negative effects on provider motivation, productivity, and service delivery. Our findings suggest that P4P schemes have the potential to improve provider motivation and productivity, but also the potential to cause negative spillover effects and misreporting.

One of the main findings of our study is the potential negative spillover effects of P4P schemes on non-incentivised activities. These effects can be seen in a number of ways, such as when providers spend more time on incentivised activities and less time on non-incentivised activities, or when providers shift their focus to incentivised activities at the expense of non-incentivised activities. This can lead to a decrease in the quality of non-incentivised activities, and to a decrease in the overall efficiency of the health system. Our findings also suggest that P4P schemes can encourage misreporting and gaming. This can happen when providers misreport information in order to claim higher incentives, or when they engage in gaming by misreporting or not reporting information.

To address these negative effects, it is important to consider the design and implementation of P4P schemes. Our findings suggest that P4P schemes should be designed in a way that minimises negative spillover effects and misreporting. This can be done by ensuring that the incentives are designed to be compatible with medical ethics, and by providing adequate verification and sanctions for misreporting. It is also important to consider the context in which P4P schemes are implemented, as this can affect the likelihood of negative spillover effects and misreporting.

In conclusion, our study provides evidence on the potential negative effects of P4P schemes, and suggests ways to design and implement these schemes to minimise these effects. Our findings highlight the importance of considering the context in which P4P schemes are implemented, and of designing these schemes in a way that minimises negative spillover effects and misreporting. This is crucial if P4P schemes are to be implemented effectively and efficiently, and if they are to achieve their intended outcomes.
programmes implemented in LMICs result in intended or unintended outcomes by exploring the underlying mechanisms and contextual and programme design moderators. Although we specifically targeted studies that in some way captured the associative mechanisms between P4P and relevant outcomes, it is important to note that the general evidence on the potential outcome effects of P4P was mixed and indeterminate, which is consistent with other systematic reviews on P4P (Witter et al., 2009; Turcotte-Tremblay et al., 2016; Oxman and Fretheim, 2009). For example, many of the studies included in this review did not find evidence that P4P schemes are consistently associated with increased patient demand for healthcare (Engineer et al., 2016; Powell-Jackson et al., 2015; Basinga et al., 2011) or with improving healthcare provider motivation (Schuster et al., 2016).

It is also clear that existing P4P studies, as a body of knowledge, remain insufficient for coming to a clear determination on a full set of pathways and mechanisms, and that variation in study design, programme design, implementation and contextual influences makes it challenging to make generalisations on P4P in LMICs. That said, the review did pinpoint a number of common pathways and contextual factors that demonstrated how P4P works in LMIC settings, namely by increasing the utilisation of healthcare services, patient satisfaction, healthcare provider productivity, and improving governance arrangements. The review also identified pathways to unintended consequences.

In terms of the utilisation of health services in LMICs, common pathways that were suggested to affect outcomes included supply-side changes including a resource effect: the improved availability of drugs and equipment and greater provider effort; and an effect on procedures within the organisation: greater adherence to clinical care guidelines, improved interactions with patients, and a reduction in user fees charged. Facility autonomy around financial management and decision making which was reported in some schemes supported these supply side changes. Provider initiatives to stimulate demand within the community, such as outreach activities, were also important. Changes in supply and demand side factors resulted in greater patient trust which was a further mechanism underpinning programme demand. We found that P4P effects on patient satisfaction in LMICs were driven by improvements in the quality of facility infrastructure and the availability of drugs and supplies; an increased availability of services and reductions in user fees charged increasing the affordability of care, and provider responsiveness. The availability of drugs and supplies, and reductions in user fees emerged as key mechanisms stimulating both service utilisation and patient satisfaction. Lastly, we found that P4P schemes effect on health worker productivity was enabled by increased provider motivation and that P4P can strengthen internal and external accountability. Unintended consequences such as spillover effects on non-incentivised activities and misreporting of information were commonly reported. Although these spillovers were not present or reported across all cases, reference to these issues as representing a potential problem common to P4P was generally noted.

From our review, it was also possible to identify a number of contextual factors that likely moderated the effects of P4P schemes in LMICs, either enhancing or undermining the ability of these schemes to strengthen the health system. These can be classified as distal factors characterising the wider health system; and proximal factors characterising the facility environment within which the P4P scheme was introduced and the characteristics of the population served by the facility. In terms of distal factors, the level of decentralisation of the health system was of key importance, as this shapes the degree of financial and management autonomy at the facility level and affects the ease of equipment and medicine procurement and the extent to which facility staff can determine and control the allocation of resources towards local priorities and needs. Greater autonomy was associated with stronger programme effects on facility resourcing (availability of drugs, supplies and equipment) and staff motivation and productivity. A second distal factor is the efficiency of the banking system and ease of bonus transfers, affecting timeliness of bonus payments; where bonus payments were timely this increased motivation and productivity. A third distal factor is the financing of public services, and the extent to which this is dependent on user charges: where user charges were higher, P4P was more likely to affect demand through a reduction in user charges; however, P4P effects on demand were generally higher where user fees were absent or where concurrent demand side programmes such as insurance or voucher schemes were in place as services were more affordable.

Proximal factors at the facility level include staffing levels, staff training, knowledge and motivation levels and the quality of facility infrastructure before the programme started. These factors were important in shaping the provider response to the programme, as where staff levels were higher they could better absorb the additional reporting tasks and demand associated with P4P; where skills and motivation were higher, staff were more likely to increase adherence to clinical care guidelines, resulting in greater patient satisfaction. Conversely where resources were limited at the outset, P4P was less likely to overcome these constraints and improve outcomes. The risks of unintended consequences such as negative spillovers and gaming were also higher in facilities with more limited staff capacity.

Proximal factors at the level of the wider community included pre-existing social norms about care seeking, which generally constrained the ability of P4P programmes to improve demand through outreach activities; and geographical access, with user fee effects being attenuated among more remote communities, but rural facilities generally performing better under P4P schemes due to lower baseline performance levels and more scope for improvement.

A number of programme design features were also identified within the review as supporting health system effects of P4P. In terms of what is incentivised, schemes with a wide range of incentivised indicators appeared less prone to negative spillover effects; and tasks requiring a lower degree of effort were often prioritised by providers irrespective of the associated reward. In terms of who receives the bonus payment, ideally everyone who has a role to play in service delivery needs to be incentivised to avoid system bottlenecks and bad feeling among those who are left out of the incentive system. Incentivising district managers and governing committees can enhance the governance function of the health system. Fairness in the distribution of bonus payments among facility staff, was also important in ensuring the programme motivated providers. The use of bonus payments for facility improvements, which was a common feature of programmes in Sub-Saharan Africa, was critical to improving structural quality resulting in better quality care and greater patient satisfaction.

The level of the health worker incentive, and its size in relation to health worker income is clearly important in determining programme effects, although the review did not allow us to determine minimum thresholds. Where the facility incentive was equal or greater than user fee revenue, P4P schemes appear more likely to motivate a reduction in user fees.

Auditing of performance data through verification is an important feature of programme design. Where verification systems involved communities and district managers this increased external and internal accountability respectively. The verification system could trigger payment delays or result in long lag times between reporting and payment, which would reduce motivation effects. Strong verification systems reduce the risk of misreporting, but where these represent a heavy time burden for facility staff, this can result in negative spillovers. There is therefore a need to balance rigour with efficiency in the design of verification systems.

While many of the mechanisms hypothesised in the original theory of change were supported by the review, a number were not. This does not mean that P4P schemes do not affect these areas, just that these areas have been understudied and there is currently a lack of evidence that these were mechanisms underpinning programme outcomes. A number of new mechanisms were identified, especially around the unintended effects of P4P, with penalty systems encouraging providers to retain
drugs at facilities and making misreporting of results more likely; and the mechanisms underpinning provider motivation, notably the timeliness of payments, fairness in the distribution of incentives and clarity of roles. Overall, the review enabled us to build a more nuanced understanding of the relationships between supply and demand side elements of the health system, contextual and programme moderators and outcomes using a CLD.

We found that some of the mechanisms which emerged as key, were not obviously linked to the financial incentive component of P4P. For example, the increased availability of drugs and supplies, and improved facility infrastructure, is a reflection of the ‘resource effect’ of P4P, the fact that providers have more revenue. Such results may have been achieved by simply increasing facility budgets. The greater autonomy due to financial decentralisation and ability to plan and prioritise how funds are spent were also important; as was the monitoring and auditing of performance. Future research should compare financial incentives to these alternative reforms to determine their relative effectiveness and cost-effectiveness.

Within the literature reviewed there was often a tendency to conflate heterogeneous programme-specific moderators (such as design and/or implementation shortcomings) with non-scheme-related contextual factors. As a result, it was often difficult to determine what factors moderated P4P outcomes. For example, a reported lack of essential consumables and medicines could have been a result of multiple factors, such as centralised procurement inefficiencies (contextual), a restrictive list of medicines available to be purchased locally (P4P design feature or contextual political restrictions), insufficient bonus amounts (design), and/or cultural beliefs regarding traditional medicines (contextual). However, often the reasons for a shortage of consumables and medicines was not identified clearly, instead focusing on the fact that P4P performed well or poorly relative to a specific set of facility, quality, or medical supply targets. As another example, many of the studies reviewed highlighted the moderating effect of low staff numbers suggesting that human resource deficiencies significantly undermined P4P performance, as well as facilities being able to ensure sufficient staffing to absorb additional demand and tasks associated with data reporting. However, it was often not clear whether these staff shortages were due to underestimations within the P4P programme design, and/or whether staff shortages were a result of broader contextual health system conditions that affected the results of an otherwise well designed P4P scheme. Overall, there is a need for better conceptual and methodological tools to more reliably understand, classify and measure P4P pathways and how multifarious heterogeneous and exogenous factors influence the performance of P4P schemes.

Several studies in our review also reported concerns about the overall cost-effectiveness of P4P and whether the gains in certain indicators are offset by higher P4P procedural costs (e.g. verification) (Ssengooba et al., 2012; Antony et al., 2017; McMahon and De Allegri, 2016; McMahon et al., 2016; McMahon et al., 2018; Borghi et al., 2015). Studies suggested a need to find a balance between rigorous verification processes and their practical feasibility and costs in terms of financial resources and time, as these funds would otherwise be available for other activities including providing funds to providers. For example, a study in Tanzania found that managing the P4P programme was the most costly component of ongoing implementation and exceeded the costs of financial incentives by between 1.7 times (in financial costs) and 1.9 times (in economic costs) (Borghì et al., 2015).

Finally, a number of studies also reported concerns of sustainability of P4P programmes where investment is less driven by government (Ssengooba, et al., 2012; Bertone and Meessen, 2013; Matsuoka et al., 2014; Van de Poel et al., 2016; Borghi et al., 2015; Chimbutu et al., 2015). Authors of a study in Burundi advocate for using external agencies in the short-term to build capacity and coach, especially in weak health systems, with a focus on later transitioning to local health system actors (Bertone and Meessen, 2013). They argue that “an external agency is able to tap into highly qualified national and international expertise and, because of its role in the verification, it has at its disposal information on facilities’ performance that could help effective coaching” (Bertone and Meessen, 2013). However, authors of a study in Uganda note that coaching is a powerful enforcement mechanism, and entrusting it to an external agency may create a dependency, which could undermine the long-term sustainability of P4P interventions and behavioural changes (Ssengooba et al., 2012). It was also noted that as local health system actors’ capacities are strengthened, reliance on external agencies would become counterproductive, as it would create duplication in fundamental responsibilities (Bertone and Meessen, 2013).

There were numerous limitations and weaknesses in the reviewed evidence, so the findings should be interpreted with caution and viewed in light of these. Evidence in the review is weak in terms of mechanisms, though clearly a number of studies do assess health system effects. Admittedly, there are also issues with generalisability and design differences across schemes. Most of the evidence also examined programme effects relative to singular health system pillars or levers and only one study examined the overall effects of programmes on the health system as a whole, using methodology grounded in complexity science (Borghì and Chalabi, 2017). This lack of research evidence thus highlights a significant gap in our understanding of how P4P effects change, through which sorts of mechanisms, and under what conditions.

In addition, very few of the studies included in the review were specifically designed to study pathways though which P4P outcomes were achieved. While most studies reviewed were able to show that P4P is associated with an outcome (e.g. care seeking) as well as with a plausible mechanism (e.g. structural quality), very few provided further evidence that a particular mechanism was indeed relevant for a specific outcome (for example, evidence to suggest that structural quality is a channel through which P4P improved care seeking). As much of the included evidence in the review is observational and/or qualitative, the effects described are not based on experimental or quasi-experimental data in most part, hence the review sets out to describe relationships rather than attribute causal effects to P4P in the way of a systematic review.

A further limitation of this review relates to our level of confidence about the generalisability of the common P4P pathways and moderators. Generalisability issues arise because most P4P studies focus on outcomes and not specifically on the combination of programme mechanisms and contextual influences delivering those outcomes. Where these factors are discussed they are often presented as side notes or as bi-product findings or explanations to support initial research questions. Therefore, a level of interpretation by the review team was necessary. Second, most studies of P4P in LMICs generally focused on a single country programme with specific attention given only to its particular set of programme indicators, with little to no discussion of health system effects. Since each programme has its own unique P4P design and implementation mechanisms, it remained difficult to make clear comparative determinations on any common set of P4P features. Third, each study utilised different theoretical approaches, methodologies, and programme areas of focus. As a result, direct comparisons across a set of similar programming criteria or programme design and implementation features was often not available, thus limiting generalisability.

Although we used the MMAT tool to assess the quality of included studies, we acknowledge the limitation of using quality assessment tools designed to assess the quality of studies that were not primarily designed to unpack how and why P4P impacts the health system. Accordingly, our own assessment may have inflated the quality of certain studies that were very well-designed to assess the impact of P4P on a specific health outcome (e.g. facility delivery), but not to analyse how and why P4P affected the outcome in the specific LMIC context.

Despite its limitations, this is the first study of its kind to conduct a realist review to describe and explain how and why P4P initiatives work (or fail to work) in different LMIC contexts by exploring the underlying programme theories and the interactions between contextual factors, programme design, mechanisms of change and outcomes. Furthermore,
our review moves away from a static, linear representation of a programme theory by drawing on complexity science methodology and a CLD to represent the final revised programme theory. Doing so allowed us to fully represent the dynamic and complex nature of this intervention and its impact on the health system as a whole, including by being able to illustrate non-linear feedback loops and relationships in the CLD to further our understanding of how P4P programmes operate in LMIC contexts.

Our synthesis of the current state of the evidence of how and why P4P affects health systems to produce varied outcomes in different LMIC contexts can inform donors, policymakers and implementers to design more effective P4P programmes to strengthen health systems and achieve sustainable service delivery and health impact and minimise unintended effects. Building on our study findings, Fig. 6 presents key features underpinning effective P4P programmes in LMICs and poses 20 key questions to donors, policymakers and evaluators in charge of designing or studying P4P programmes in LMICs, with an aim of helping this audience to think about important contextual and programme design factors that may have an impact on their P4P programme’s intended outcomes.

**Authors’ contributions**

Neha S. Singh and Josephine Borghi conceived the study and all authors contributed to the final structure and content of the paper. Neha S. Singh, Roxanne J. Kovacs and Josephine Borghi screened articles, and Neha S. Singh led the overall data analysis with inputs from Roxanne J. Kovacs and Garrett W. Brown. Neha S. Singh wrote the first draft of the paper with inputs from Roxanne J. Kovacs, Rachel Cassidy, Josephine Borghi and Garrett W. Brown, and all authors contributed to critical interpretation of the results and development of the paper. All authors saw and approved the final version.

**Declaration of competing interest**

The authors have not declared competing interests. The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views of the organisations listed.

**Acknowledgements**

We would like to acknowledge Dr. Laura Anselmi, who helped with
screening articles and provided feedback on the review’s emerging findings. We would also like to thank relevant stakeholders and experts who provided inputs on the review’s middle-range theory at the Fifth Global Symposium on Health Systems Research in Liverpool, UK in October 2018 and at a P4P all partners workshop in Maputo, Mozambique in March 2019. This research was supported by a Health Systems Research Initiative grant number MR/P014429/1 jointly funded by the Medical Research Council, Wellcome Trust, Economic and Social Research Council, Department for International Development. SRK was also funded by the NIHR Imperial Patient Safety Translational Research Centre.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2020.113624.

References

Abduvalieva, S., Akerov, A., Rakhmanov, C., 2016. Improving the quality of medical care for mothers and newborns in Kyrgyzstan by funding health care based on the results. J. Matern. Fetal Neonatol Med. 29, 210–211.

Agustina, L.A., Aryono, P., Nusrinov, J., Anom-Bendi, F., Dzikuho, H., Antwi, E., et al., 2014. Advancing the application of systems thinking in health: provider payment and service supply behaviour and incentives in the Ghana National Health Insurance Scheme - a systems approach. Health Res. Pol. Syst. 12 (1).

Arembipe, L., Binyaruka, P., 2017a. Understanding causal pathways within health systems policy evaluation through mediation analysis: an application to payment for performance (P4P) in Tanzania. Implement. Sci. 12 (1), 10.

Arembipe, L., Binyaruka, P., Borghi, J., 2017. Dec. Understanding causal pathways within health systems policy evaluation through mediation analysis: an application to payment for performance (P4P) in Tanzania. Implement. Sci. 12 (1), 1–8.

Antony, M., Bertone, M.P., Barths, O., 2017. Exploring implementation practices in results-based financing: the case of the verification in Benin. BMC Health Serv. Res. 17 (1), 204.

Aryankhesal, A., Sheldon, T.A., Mannion, R., 2013. Role of pay-for-performance in a hospital performance measurement system: a multiple case study in Iran. Health Pol. Plann. 28 (2), 206–214.

Aryankhesal, A., Sheldon, T.A., Mannion, R., Mahdjoub, S., 2015. The dysfunctional consequences of a performance measurement system: the case of the Iranian national hospital grading programme. J. Health Serv. Res. Pol. 20 (3), 138–145.

Bank, W., 2015. The World Bank: Data Worldbank Org, 2016.

Barnes, A., Brown, G.W., Harman, S., Papamichail, A., Banda, P., Hayes, R., et al., 2014. African Participation and Partnership in Performance-Based Financing: A Case Study in Global Health Policy. EQUINET Discussion Paper.

Basinga, P., 2009. Impact of Performance-Based Financing on the Quantity and Quality of Maternal Health Services in Rwanda. Tulane University.

Basinga, P., Gertler, P.J., Binagwaho, A., Soucat, A.L.B., Sturdy, J.R., Vermeersch, C.M.J., 2013. Hospital grading programme. J. Health Serv. Res. Pol. 20 (3), 138–145.

Bernal, D., 2015. Argentina: can short term incentives change long term behaviour?. Bernal, P., Martinez, S., Celhay, P., 2018. Is results-based aid more effective than conventional aid?. In: Evidence from the Health Sector in El Salvador. Inter-American Development Bank.

Bertone, M.P., Meessen, B., 2013. Studying the link between institutions and health system performance: a framework and an illustration with the analysis of two performance-based financing schemes in Burundi. Health Pol. Plann. 28 (8), 847–857.

Bertone, M.P., Witter, S., 2015. An exploration of the political economy dynamics of Maternal Health Services in Rwanda. Trop. Med. Int. Health 22 (1), 92–103.

Bhatnagar, A., George, A.S., 2016. Motivating health workers up to a limit: partial effects of payment for performance (P4P) in Tanzania. Implement. Sci. : IS 12 (1), 1074.

Bonfrer, I., Soeters, R., Van de Poel, E., Basenya, O., Longis, G., van de Leeji, F., et al., 2014a. Introduction of performance-based financing in Burundi was associated with improvements in care and quality. Health Aff. 33 (12), 2179–2187.

Bortone, M.P., Lagarde, M., Witter, S., 2016. Performance-based financing in the context of the complex remuneration of health workers: findings from a mixed-method study in rural Sierra Leone. BMC Health Serv. Res. 16, 286.

Chernichovsky, D., Bayulken, C., 1995. A pay-for-performance system for civil service employment in the Indonesian health sector. Soc. Sci. Med. 41 (2), 155–161.

Chansa, C., Das, A., Qamruddin, J., Friedman, J., Mkandawire, A., Vledder, M., 2015. Improving the quality of medical commodities in Tanzania. Trop. Med. Int. Health 22 (1), 92–103.

Chang, A.Y., Ogbuoji, O., Atun, R., Verguet, S., 2017. Dynamic modeling approaches to shaping health worker incentives in three districts in Sierra Leone. Soc. Sci. Med. 173, 217–226.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.

Chollet, P., Gervais, M., 1998. Introducing payment for performance in the health sector of Tanzania- the policy package. BMJ Glob Health 2 (3), e000406.
Falisse, J.B., Meessen, B., Ndyashimiye, J., Bossuyt, M., 2012. Community participation and voice mechanisms under performance-based financing schemes in Burundi. Int. J. Health Plann. Manag. 27 (2), 574–582.

Falisse, J.-B., Ndyashimiye, J., Kamenyosi, V., Bossuyt, M., 2015. Performance-based financing in the context of selective free health-care: an evaluation of its effects on the use of primary health-care services in Burundi using routine data. Health Pol. Plann. 30 (10), 1307–1317.

Fan, E.L., 2017. Counting results: performance-based financing and HIV testing among pregnant women in Malawi. Soc. Sci. Med. 179, 61–73.

Mayumana, I., Borghi, J., Arslanli, S., Mamdani, M., Lange, S., 2017a. Effects of payment for performance on accountability mechanisms: evidence from pwani, Tanzania. Soc. Sci. Med. 179, 73–82.

Mayumana, I., Borghi, J., Arslanli, S., Mamdani, M., Lange, S., 2017b. Effects of payment for performance on accountability mechanisms: evidence from pwani, Tanzania. Soc. Sci. Med. 179, 61–73.

McMahon, S., De Allegri, M., 2016. Final Results of the SSDR-PBF Process and Impact Evaluation [Available from: https://www.ditp.act/malawi/documents/SSDR_final_report.pdf].

McMahon, S.A., Brenner, S., Lohmann, J., Makwero, C., Torbica, A., Mathanga, D.P., et al., 2016. Evaluating complex health financing interventions: using mixed methods to inform further implementation of a novel PBF intervention in rural Malawi. BMC Health Serv. Res. 16 (1), 414.

McKee, M., Peabody, J., Shinkhada, R., Quimbo, S., Solon, O., Javier, X., McCulloch, C., 2014. The money can be a motivator, to me a little, but mostly PBF just helps me to do what I need to do. PBF program in Rwanda. Soc. Sci. Med. 138, 1–7.

McManus, I., Borghi, J., Arslanli, S., Mamdani, M., Lange, S., 2017a. Effects of payment for performance on accountability mechanisms: evidence from pwani, Tanzania. Soc. Sci. Med. 179, 73–82.

Peabody, J.W., Shinkhada, R., Quimbo, S., Solon, O., Javier, X., McCulloch, C., 2014. The impact of performance incentives on child health outcomes: results from a cluster randomized controlled trial in the Philippines. Health Pol. Plann. 29 (5), 615–621.
S. N. Singh et al.  

Social Science & Medicine 270 (2021) 113624

17

Peabody, J.W., Quimbo, S., Florentino, J., Shimkhada, R., Javier, X., Pascualo, D., et al., 2017. Comparative effectiveness of two disparate policies on child health: experimental evidence from the Philippines. Health Pol. Plann. 32 (4), 563-571.

van De Poel, E., Flores, G., Ir., P., O’Donnell, O., 2016. Impact of performance-based financing in a low-resource setting: a decade of experience in Cambodia. Health Econ. 25 (6), 688-700.

Powell-Jackson, T., V., W.C.M., Han, W., 2015. Realigning demand and supply side incentives to improve primary health care seeking in rural China. Health Econ. 24 (6), 755-772.

Priedeman Skles, M., Curtis, S.L., Basinga, P., Angelle, G., 2012. An equity analysis of performance-based financing in Rwanda: are services reaching the poorest women? Health Pol. Plann. 28 (8), 825-837.

Rahman, S., Choudhury, A.A., Khanam, R., Moin, S.M.I., Ahmed, S., Begum, N., et al., 2017. Effect of a package of integrated demand- and supply-side interventions on facility delivery rates in rural Bangladesh: implications for large-scale programs. PLoS One 12 (10), e0186182.

Regalla, F., Castro, L., 2007. Performance-based incentives for health: demand and supply-side incentives in the Nicaraguan Red de Protección Social.

Rennoms, D., Holvoet, N., Orach, C.G., Criel, B., 2016. Opening the ‘black box’ of performance-based financing in low- and lower-middle-income countries: a review of the literature. Health Pol. Plann. 31 (9), 1297-1309.

Rennoms, D., Holvoet, N., Criel, B., Meessen, B., 2017. Performance-based financing: the same is different. Health Pol. Plann. 32 (6), 860-868.

Rob, U., Alam, M.M., 2013. Performance-based incentive for improving quality of maternal health services in Bangladesh. Int. Q Community Health Educ. 34 (4), 303-312.

Roland, M., 2012. Pay-for-performance: not a magic bullet. Ann. Intern. Med. 157 (12), 830.

Rupa, L., Ngirabega, J., Janssen, W., Van Bastelaere, S., Porignon, D., 2017. Incentives to change: effects of performance-based financing on health-care quality and utilization in urban areas of Cameroon. Rural China. Health Aff. 33 (3), 502-511.

Sun, X., Liu, X., Sun, Q., Yip, W., Wagstaff, A., Meng, Q., 2016. The impact of a pay-for-performance scheme on prescription quality in rural China. Health Econ. 25 (6), 706-722.

Suthar, A.B., Nagata, J.M., Nsamzinama, S., Barnighausen, T., Negussie, E.K., Doherty, M. C., 2017. Performance-based financing for improving HIV/AIDS service delivery: a systematic review. BMC Health Serv. Res. 17 (1), 6.

Turcotte-Tremblay, A.M., Sengu, P., Shisana, O., McKee, M., Riddle, V., 2016. Does performance-based financing increase value for money in low- and middle-income countries? A systematic review. Health Economics Review 6 (1).

Vermeersch, C., Van Donkkenhoven, E., Sturdy, J., 2012. Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health. World Bank, Washington, DC, p. 321 http://www.worldbank.org/health/impactevaluationtoolkitConclusion.

Vermeersch, C., Rothenbuhler, E., 2008. Payment for Performance (P4P) Evaluation: 2008 Zambia Country Report for Cordaid. Royal Tropical Institute (RIT), Amsterdam.

Yip, W., Powell-Jackson, T., Chen, W., Fei, E., Hu, M., Jian, W., et al., 2014. Capitation payments to improve quality of care in rural Shandong, China. J. Health Econ. 33 (5), 1120-1130.

Ye, M., Aninanya, G.A., Sie, A., Kakoko, D.C.V., Chatio, S., Kagone, M., et al., 2014. Establishing sustainable performance-based incentive schemes: views of rural health workers from qualitative research in three sub-Saharan African countries. Rural Aem. Health 14 (3), 2681.

Yip, W., Eggleston, K., 2001. Provider payment reform in China: the case of hospital reimbursement in Hainan province. Health Econ. 10 (4), 325-339.

Zhang, W.L.H., Ma, Y., Guo, Y., Fang, Q., Yang, Z., Zhang, X., Jia, M., Chen, X. S., 2017. The unintended consequences of community verifications for performance-based financing in Burkina Faso. Soc. Sci. Med. 191, 226-236.

Zhang, W.L.H., Pibiri, P., McPake, B., Palmer, N., 2012. Why performance-based contracting failed in Uganda: an ‘open box’ evaluation of a complex health system intervention. Soc. Sci. Med. 75 (2), 377-383.

Zerman, J.D., 2000. Business Dynamics: Systems Thinking and Modelling of the Complex World. Irwin McGraw Hill, Boston.