CURATIO INTERNATIONAL FOUNDATION

25 Years for Better Health Systems

Effects of Pay for Performance on utilization and quality of care among Primary Health Care providers (in private settings) in Middle and High-Income countries

Evidence synthesis

Prepared by:
Ivdity Chikovani, Lela Sulaberidze, Alisa Tsuladze

August 2020

Abbreviations

ACO	Accountable Care Organization
AHPSR	Alliance for Health Policy and Systems Research
AMSTAR	A MeaSurement Tool to Assess systematic Reviews
ANC	Antenatal Care
CHD	coronary heart disease
COPD	Chronic obstructive pulmonary disease
DRG	Diagnosis-related group
ED	Emergency department
ERA	Embedded RApid Reviews in Health Systems Decision Making
GP	General practitioner
HbA1c	Hemoglobin A1C
IOM	US Institute of Medicine
LDL	Low-density lipoprotein
LMIC	low- and middle-income countries
MCH	Maternal and child health
NCD	Non-communicable disease
NCDC	National center for disease control and public health of Georgia
NHS	United Kingdom National Health Service
OECD	Organization for Economic Co-operation and Development
P4P	**Pay for Performance**
PBF	Performance-Based Financing
PHC	Primary health care
QOF	Quality Outcome Framework
RBF	Results-Based Financing
TB	Tuberculosis
UHC	Universal Health Care
UK	United Kingdom
US	United States
VBP	Value-Based Purchasing
WHO	World Health Organization

Acknowledgment

The document has been prepared by the Curatio International Foundation for the Health and Social Issues Committee of the Parliament of Georgia and Ministry of Internally Displaced Persons from the Occupied Territories, Labor, Health and Social Affairs of Georgia (MoILHSA) in the frame of the Embedded RApid Reviews in Health Systems Decision Making- ERA Platform.

This work was funded by the Alliance for Health Policy and Systems Research/WHO. The Alliance is able to conduct its work thanks to the commitment and support from a variety of funders. These include our long-term core contributors from national governments and international institutions, as well as designated funding for specific projects within our current priorities. For the full list of Alliance donors, please visit: https://www.who.int/alliance-hpsr/partners/en/

CIF would like to thank the Alliance for financing the ERA platform aimed at responding policy needs in the country with evidence products.

Table of Content

Executive Summary

The evidence review summarizes existing literature on Pay for Performance (P4P) effectiveness on utilization and quality of primary health care in private settings in middle-income and high-income countries. The evidence review was developed in response to the request of the Parliament Committee on Health and Social Issues in Georgia in the frame of Embedded Rapid Reviews in Health Systems Decision Making (ERA) platform in Georgia.

We conducted a narrative review of systematic reviews and review papers, evaluation reports and individual studies. We included studies that evaluated P4P programs in primary care targeting individual, group, or institutional practices and included quality dimension. Date and language restrictions were applied.

We included 45 publication in the review. Although the papers mention the private sector in some form, it was impossible to make a public-private distinction or any kind of data synthesis, therefore, in coordination with the main client, we slightly changed our research question which has been formulated in the following way: What is the effectiveness of P4P on the utilization and quality of primary care services in middle and high-income countries. The review also looked at: What unintended consequences or spillover effects are associated with P4P? and What factors should be considered during P4P program design and implementation?

The quality outcomes have been synthesized and structured in four groups: 1st group includes process of care, intermediate and hospitalization level outcomes, 2nd group includes patient health outcomes, 3rd group covers equity, coordination and continuity of care and the 4th group includes unintended consequences and spillover effects.

The review found that there was significant heterogeneity in terms of the contexts in which the P4P schemes were implemented, services and populations targeted, types of outcome measures and incentives used. Most of the P4P interventions targeted preventive care, management of chronic and maternal and child health (MCH) conditions. The review papers are dominated by studies from the UK (QOF scheme) and the US.

Although some systematic reviews showed contradictory outcomes on *PHC service utilization,* P4P was found to be an effective intervention scheme to increase the utilization of preventive care services for MCH. Evidence on the utilization of screening services for chronic diseases and cancer showed inconclusive results.

Earlier systematic reviews and reviews of the systematic reviews reported insufficient evidence on the effectiveness of P4P interventions in improving *quality of care*. The major deficiency was related to a lack of studies with strong designs. The later studies with robust designs tend to show less positive results compared with the studies with weak designs.

The review showed that there is more consistent evidence that P4P schemes improve *process of care outcomes*. there is low-strength contradictory evidence that the P4P programs improve process-of care over the short-term, while evidence is limited on long term outcomes. Mixed evidence was found with regards to *intermediate and proxy outcomes* such as emergency department and hospital admissions due to aggravation of chronic conditions, institutional deliveries.

The evidence on the P4P effect on *health outcomes* such as disease prevalence, disease-specific or overall mortality is limited.

There is inconclusive evidence whether P4P influences positively or negatively *equity*, *continuity of care, coordination* between the health workers.

Few studies report about *unintended negative consequences* such as gaming practices in the P4P schemes where physicians try to manipulate with the data in order to prove achievement of certain indicators and be eligible for incentive payments. Adverse selection of patients and distortion, as well as concentration on incentivized activities have also been examined by certain reviews. P4P have *positive spillover effects,* such as improved performance on unincentivized measures or medical conditions, improved intermediary or health outcomes in non-target populations.

The P4P interventions vary widely by its design features, contexts where they operate, cultural factors, implementation specifics, etc. Many P4P programs have evolved over time by adjusting design, introducing a mechanism to mitigate negative spillover effects, adding quality improvement and cost parameters to achieve desired goals. These process changes are not sufficiently captured by empirical studies and thus largely remain unknown to researchers.

The review discusses set of conditions that could help in P4P program design and implementation. These conditions are engagement of providers in scheme design and alignment of measures with their professional values, fairness of the incentives distribution and their flexibility, acknowledgment of baseline performance level and practice size as well as target population characteristics, existence of effective reporting and robust monitoring systems, etc.

In sum, P4P programs have likely been effective in increasing the utilization of care and the process of care outcomes while evidence on P4P long-term effect is limited. The heterogeneity of evidence does not allow to conclude that provider-targeted financial incentives have failed to improve the quality of care. To fully realize its potential in quality improvement P4P programs need to be carefully planned, implemented and rigorously evaluated. Consideration of important preconditions suggested by theoretical concepts and empirical evidence helps P4P programs to achieve desired goals.

Introduction

Pay for Performance (P4P) is a relatively new strategy aiming at improving the performance of the health care providers through incentivizing and motivating behavior change for the desired output. P4P is one of the forms of a wider umbrella Results-Based Financing (RBF) concept and comes with a variety of labels and metrics (Musgrove P., 2011). P4P can be used to pay individuals, groups of people or organizations and includes a wide range of interventions that vary with respect to the level of care at which incentives are targeted, how performance is measured, features related to incentive structure, size and frequency (Witter et al., 2012). Many country health systems adopted P4P as a complement to other reimbursement practices. Its wide application has become increasingly common in primary care.

A number of systematic reviews have examined different angles of P4P across different economies, health system configuration and schemes. While various reviews focus on P4P from different angles, there is no summarized evidence on how P4P works in a private environment.

The public-private mix varies by countries. The private sector is increasingly recognized as playing an important role in health systems across the world. Many high-income countries have a long history of engagement with private providers, while in low- and middle-income countries (LMIC) private sector has emerged more recently and is growing and capturing an increasing share of health market (WHO, 2010; Wadge et al., 2017).

There are theoretical claims about the positive effect of private ownership in primary health care (PHC) (Alonso et al., 2015). However, studies show mixed results. One systematic review comparing the performance of private and public health-care systems in LMIC reported about poor quality in both types of providers with the private sector performing better in drug availability and possibly being more client oriented (Berendes et al., 2011). Another systematic review showed that evidence does not support the claim that private sector is usually more efficient, accountable, or medically effective than the public sector, although the public sector appears to lack timeliness and hospitality towards patients (Basu et al., 2012). Studies in Malta, Hong-Kong and South Korea showed better performance in private primary care services in comparison with public services (Pullicino et al., 2015; Sung et al., 2010; Wong et al., 2010). In the contrary, no significant difference was found in Canada (Mayo-Bruinsma, Liesha, 2013).

Moreover, there are concerns about the role of the private sector in LMIC in the context of universal health coverage (UHC) (WHO, 2010; Morgan et al., 2016; Wadge et al., 2017). Private sector complexity and diversity requires specific policy approaches to engage and manage it. These capacities are limited in LMIC, leading to a failure of systems to deliver adequate outputs (McPake & Hanson, 2016; WHO, 2018c). This context is especially relevant to Georgia. Georgia is an upper-middle income country with a highly privatized healthcare system that has been implementing the Universal Health Care (UHC) program since 2013. Despite several changes in the design, the program faces challenges: PHC services are significantly underutilized. Since the UHC introduction outpatient per capita visits per annum increased by 61% and reached 3.7 visits in 2018 in Georgia (NCDC, 2019), however, it is twice lower compared to the WHO European region estimate - 7.53 visits (WHO Regional Office for Europe, 2020).

The current payment model for primary care in Georgia is mostly input-based, such as fixed payments for rural family doctors and nurses and per-capita payment for urban primary health care providers. There is no link between reimbursement mechanisms and quality parameters at this stage. No incentive schemes are in place with the exception of a pilot project on TB

outpatient care (Curatio International Foundation, 2018). The PHC fails to fulfill a gatekeeping role. There are high referral rates from family doctors to specialists leading to out-of-pocket expenditures. Accountability for performance is largely absent, and the PHC providers are rarely held accountable for their performance. Mechanisms for quality improvement including supervision and feedback is absent, compliance to standards of care is not routinely audited unless there is a complaint requiring further investigation (Chikovani & Sulaberidze, 2017; WHO Regional Office for Europe, 2018, 2019).

All these emphasize that PHC in Georgia has a poor gatekeeping role, it fails in effective management of chronic diseases, in preventing a patient from using costly specialized and inpatient services and averting the health system from increased health spending/expenditure.

Reimbursement mechanism and P4P specifically has been discussed as one of the policy options to address some of the challenges at the PHC level. There is a need for empirical evidence to inform policy discussion. In 2018 Alliance for Health Policy and Systems Research (AHPSR) supported the establishment of Embedded Rapid Reviews in Health Systems Decision Making (ERA) platform in Georgia to facilitate evidence-informed policy making. The ERA is a collaboration between researchers and policy-makers (the Parliament Committee on Health and Social Issues, the Ministry of Internally Displaced persons from the Occupied Territories, Labour, Health and Social Affairs of Georgia), where researchers would respond to the identified policy issues with evidence products. The research question discussed below was commissioned by the Parliament Committee on Health and Social Issues in Georgia.

The objective of our evidence review is to summarize existing literature on P4P effectiveness on utilization and quality of primary health care in private settings in middle-income and high-income countries. As Georgia is the upper-middle income country by the World Bank Group definition, it would have been expected to look at experience of low and middle-income settings, as these income level group countries are commonly studied together. However, we took a slightly different approach. While there are significant differences between the countries of varying socio-economic development, there are similarities as well. Experience from the developed world could be valuable for other economies. Moreover, a substantial literature on P4P programs comes from high-income countries where their application started in the 1990s in the United States (US) and early 2000s in the United Kingdom (UK) (Cromwell et al., 2011; Gemmill, 2007). In addition, the private sector is well developed in primary care in high-income countries (WHO, 2018b).

To respond fully to our main research question - whether P4P affect utilization and quality of primary care services in private settings in middle and high-income countries - we also looked at the following questions: What unintended consequences or spillover effects are associated with P4P? and What factors should be considered during P4P program design and implementation?

Methods

Defining the terms

Quality of care, despite its universal acknowledgment, is not commonly referenced in the literature. Definitions of quality include those of Donabedian, which specified that quality in relation to processes and linked it to patient-welfare (Donabedian A, 1980). About a decade later, the United States Institute of Medicine (IOM) defined the quality in the following way: "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge." In addition to the basic definition, the IOM identifies six dimensions or aims the quality care should fulfill: safe, effective, patient-centered, timely, efficient, and equitable (Institue of Medicine, 2001). Later this concept has been adopted and expanded by different organizations like the WHO, the European Commission, the OECD.

The latest definition of the quality by the WHO is as follows: "*Effective*: providing evidence-based health care services to those who need them. *Safe*: avoiding harm to people for whom the care is intended. *People-centred*: providing care that responds to individual preferences, needs and values. In order to realize the benefits of quality health care, health services must be *Timely*: reducing waiting times and sometimes harmful delays for both those who receive and those who give care; *Equitable*: providing care that does not vary in quality on account of age, sex, gender, race, ethnicity, geographical location, religion, socioeconomic status, linguistic or political affiliation; *Integrated*: providing care that is coordinated across levels and providers and makes available the full range of health services throughout the life course; and *Efficient*: maximizing the benefit of available resources and avoiding waste (WHO, 2018a).

In this review, we mainly focus on effectiveness, safety and equity dimensions of quality. We have not included the efficiency dimension in our review. Included studies should demonstrate a link between P4P and quality measures. There are different types of indicators used by health systems and programs to measure quality at different levels of care. It could be the availability of resources also defined as the structural quality, adherence to standards of care also defined as process quality or process-of-care outcomes, or technical quality, emergency department and hospital admissions for ambulatory care sensitive conditions as a proxy measure of primary care performance, patient-level outcomes including expenditures and experience with service, and lastly disease prevalence, mortality rates as subnational, national level outcomes.

Primary care could be defined in multiple ways in the literature. It includes terms such as general medicine, family medicine, family practice, outpatient care.

P4P conceptually means incentivizing for a better payoff from health care (Musgrove P., 2011). Different country programs use various labels such as Performance-Based-Payment, Results-Based-Payment, Performance-Based-Financing (PBF), and an umbrella term RBF. A relatively recent term includes Value-Based-Purchasing (VBP), more typical to the US context. VBP models assume a variety of forms but are operationally defined as financial incentives that aim to improve clinical quality and outcomes for patients, while simultaneously containing or reducing health care costs (Conrad et al., 2016). For our review purposes, we are looking at financial incentive schemes that seek to motivate and change providers' behavior by aligning payment with any output, outcome, achievement, value.

Search strategy, data collection

We conducted a literature review. We searched systematic as well as non-systematic review papers.

PubMed search strategy included the following syntax (primary care AND performance-based financing) and respective vocabulary terms in titles/abstracts of the papers.

Initial search also included “private practice,” however this strategy did not result in any hints; therefore, for a wider search we omitted this term.

Terms

primary care/ or primary healthcare/ or primary health care/ or ambulatory care

results based financing/ or performance based financing/ or performance based payment/ or performance based contracting/ or pay for performance

We searched review papers published in PubMed and Health Systems Evidence database as well as evaluation reports at the RBF Health web site (https://www.rbfhealth.org). Review papers and reports published from January 2009 up to 19 December 2019 and written in English were eligible for inclusion.

Two authors reviewed all titles and abstracts generated by the search. We included studies that evaluated primary care P4P programs targeting individual, group, or institutional practices and included quality dimension. We excluded studies reviewing P4P programs of low-income countries only. The review papers covering both low-income and middle-income countries were included, however, information on middle-income countries was extracted. Similarly, the papers examining both outpatient and hospital services were selected, but information related to only outpatient services was extracted. The review papers examining P4P effects on specific conditions such as mental illness/ serious mental illness, Parkinson's disease and asthma were excluded. However, we included review papers looking at P4P effect on diabetes care as one of the main focus of primary care practice.

Considering the search date of the last systematic review of interest, we expanded our search by primary studies in middle-income countries published between March 2016 and January 2020. We have also complemented our search by scanning references of included review papers.

A calibration exercise was conducted to ensure the reliability of data extraction. About 15% of studies were extracted jointly by the two researchers. When researchers achieved good agreement on more than 90% of t studies to be included, the two researchers continued data extraction independently. Data was extracted according to the predetermined data extraction form. Among the other information, we extracted a summary of results and, in some instances, primary study level findings for a better understanding of the results.

We did not determine the methodological quality of the studies but report the AMSTAR score whenever available.

We also identified two reviews of systematic reviews published during the last decade that are relevant to our topic. Cochrane review of systematic reviews by Wiysonge et al. included three prior reviews of the effects of P4P (Akbari et al., 2008; Scott et al., 2011; Witter et al., 2012) and one review of the effects of incentives to practice in underserved areas. The paper concluded that

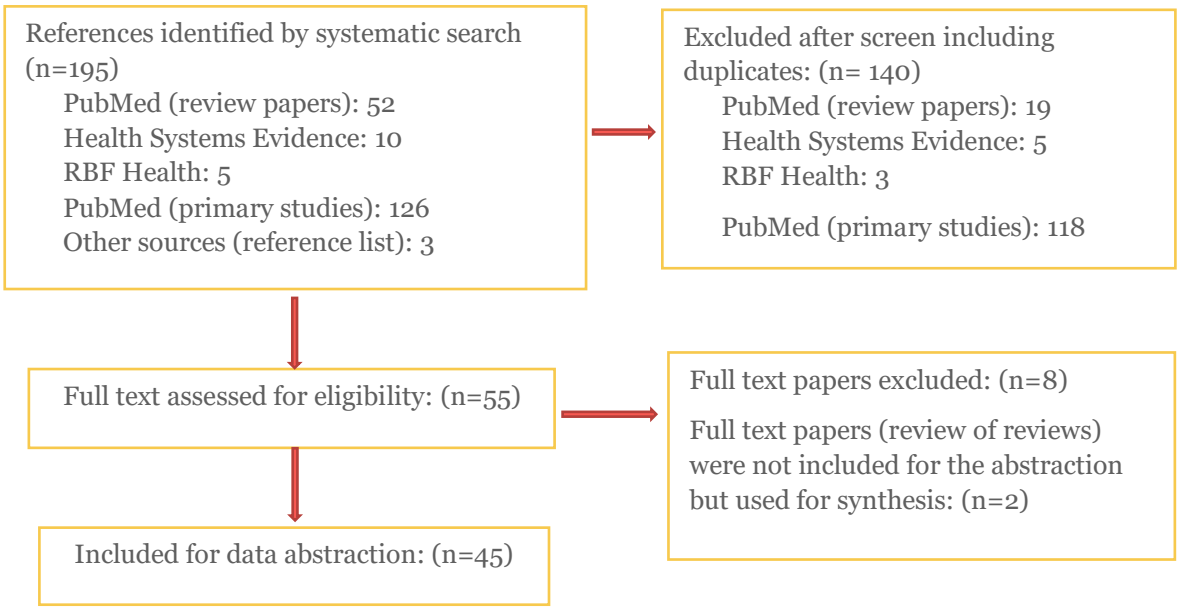
the effects of provider incentive are uncertain, including provider incentives on quality of care or outpatient referrals to secondary care and P4P for provider performance, utilization of services and patients outcomes (Wiysonge et al., 2017). The other overview is authored by Eijkenaar et al. and included 22 systematic reviews. The authors concluded that there is insufficient evidence to support or not support the use of P4P (Eijkenaar et al., 2013).

Results

Search results

A total of 195 titles and abstracts were assessed for eligibility. Following the screening and full-text assessment, 45 publications were identified and included in the final review.

Figure 1. Literature flow diagram



Change in the research question

The primary studies included in the review papers were originated from the UK (22 papers), the US (18), Taiwan (17) and Canada (12). Several papers included studies from Italy, Germany, Sweden, Argentina, Brazil, Philippines, Cambodia, Nigeria. Important to mention that the primary studies are overlapped in the review papers; therefore, duplication of findings is not ruled out.

Annex 1 presents the main characteristics of the included studies. The studies differ widely with respect to contexts in which P4P programs are implemented, primary care organization, an organizational culture within a setting, P4P program design, patient population. For example, some reviews focus only on one disease (e.g. diabetes), type of service (cancer screening), a subset of primary health care area (NCDs, Maternal and Child Health), or broadly on general primary care services.

Most importantly, although the papers mention the private sector in some form, it was impossible to make a public-private distinction or any kind of data synthesis in this regard to answer our research question. From 45 papers included in our review, only 13 mentioned public or private ownership in any form, and the rest of the papers did not provide any information with this regard. None of 13 papers analyzed the effectiveness of P4P in primary care through public/private lens. We investigated each of 13 papers deeply to find out whether included

primary studies, as shown in the reviews, provided more granular information. However, none of the primary studies pertaining to primary care where presented in this form. We did not investigate further each of the primary studies individually or characteristics of the P4P schemes or national systems to find out what type of ownership the PHC providers had during the scheme implementation.

We discussed this information gap with our main client and the decision was made to proceed further with the synthesis without the private setting dimension. Therefore, our research question has changed and has been formulated in the following way: What is the effectiveness of P4P on the utilization and quality of primary care services in middle and high-income countries.

Effectiveness of the P4P

We structure the results section in the following way: P4P effects on service utilization are followed by P4P effects on quality. Considering that the quality is a broad concept, we present the outcomes in four groups.

The following subsections describe findings on implementation considerations of P4P programs.

In the results section, we provide a more detailed description of the P4P schemes in the US, UK, Taiwan and Argentina derived from the review papers to give a better understanding of how the schemes evolved and what was captured by the review studies.

Annex 2 presents key findings from the studies included in the review.

Outcomes of P4P effect on quality of care

1st group:

Process of care: Adherence to guidelines such as recommended tests e.g. blood glucose monitoring, eye exam, etc.

Intermediate outcomes: e.g. changes in laboratory value; controlled blood pressure;

Hospitalizations: Ambulatory care-sensitive ED or hospital admissions; All-cause ED or hospital admissions;

2nd group:

Patient health outcomes: Disease prevalence; disease specific mortality; all-cause mortality;

3rd group:

Other quality outcomes: Equity; coordination of care; continuity of care;

4th group:

Unintended consequences and spillover effects: Gaming, patient adverse selection, distortion, positive spillover effects.

P4P effects on primary care services utilization

Sixteen papers in total (including 11 systematic reviews) examined the effect of P4P on PHC service utilization.

P4P effects on utilization of preventive and screening services:

- Immunization
- Antenatal Care (ANC)
- Diabetes, coronary heart disease (CHD), hypertension

We found positive effects on vaccination service utilization, where most of the studies described the increased immunization coverage as one of the main achievements of P4P scheme (Van Herck et al., 2010; Houle et al., 2012; Soranz & Pisco, 2017; Yuan et al., 2017; Patel, 2018). Significance and magnitude of change varied across the studies, e.g. in Cambodia P4P resulted in only 2.3% increase (Patel, 2018), while in Nigeria P4P scheme in three states resulted in the increase of the average coverage for completely vaccinated children from 1.4% to 49.2% during two years period (Odutolu et al., 2016). A randomized controlled trial in the US (New York city) showed statistically significant 5.9% increase of childhood immunization after bonus introduction, although these increases were largely due to improvements in documentation rather than actual immunization practices, maintaining an accurate vaccination history is a critical component of the immunization process (Fairbrother et al., 2001; Houle et al., 2012). Another observational study in the US in Accountable Care Organizations (ACO), that used Difference in Difference method found a significant increase in coverage for 8 out of 10 vaccines with less than 80% coverage rate among the incentivized physicians. No change was observed for two vaccines which starting coverage rate was close to 80%. Comparison with non-incentivized physicians showed that incentivized physicians had greater improvements in performance on 4 out of 10 vaccines coverage. In the former group large performance improvements could be explained by lower starting points and other quality improvement efforts including electronic records (Gleeson et al., 2016).

P4P schemes also increased ANC service utilization (Gertler et al., 2014; Kandpal, 2016; Khim et al., 2018; Patel, 2018; Soranz & Pisco, 2017; Wekesah et al., 2016). Argentina's P4P scheme under Plan Nacer program showed a significant positive effect on prenatal visits (6.8% point increase) and provision of tetanus toxoid vaccine for mothers (5.6% point increase). The second evaluation in one of the provinces, where increased incentives were given for a temporary period, found that early initiation of prenatal visits was 34% higher in the treatment group compared with the control. The effect in this province sustained following a year after the incentives ended (Gertler et al., 2014; Kandpal, 2016; Patel, 2018). Although there was no P4P scheme evaluation in Armenia, one study claimed that the RBF scheme played a role in improving maternal and child health and non-communicable disease services in PHC facilities and meeting annual targets (Petrosyan et al., 2017).

P4P also positively affected the utilization of screening services for hypertension and coronary heart disease (CHD) (Cattel & Eijkenaar, 2019; Lin et al., 2016a; Scott et al., 2011). However, the other study contrasts by showing no effect concerning CHD care (Mendelson et al., 2017).

Contradictory results were found for P4P schemes' effect on cancer (cervical, breast and colorectal cancer) screening service utilization across the selected systematic reviews. The studies have demonstrated that financial incentives had heterogeneous effects (positive, little, or

no effect) on improving cancer screening service utilization (Van Herck et al., 2010; Mauro et al., 2019; Houle et al., 2012).

P4P effects on Quality

Twenty-eight studies describe P4P effectiveness on 1st and 2nd group of outcomes. The studies mostly focus on chronic conditions and maternal and child health in primary care, some studies do not specify the focus of their analysis. Diabetes care is most frequently referenced condition by the studies.

1st group of quality outcomes

Nineteen studies mention a positive effect on at least some type of outcomes from the 1st group.

Almost all review papers include primary studies examining the UK's Quality Outcome Framework (QOF) (see box for more details). Relatively earlier systematic reviews base their analysis on more dated studies and tend to report about the association between improvements in the management of diabetes care and QOF (Alshamsan et al., 2010; S. J. Gillam et al., 2012). The authors mention strengthened processes that lead to an improved process of care and intermediate outcomes, particularly during the first year of the QOF introduction (S. Gillam, 2015; S. J. Gillam et al., 2012). Van Herck et al. also reports that diabetes showed the highest rates of quality improvement due to P4P implementation, with positive results also reported for asthma and smoking cessation (Van Herck et al., 2010). Lin et al. analyzed 36 studies worldwide and found that all studies on coronary heart disease and on diabetes management reported significant improvement mostly on the process of care outcomes rather than clinical (intermediate) outcomes. According to another systematic review, P4P implementation was influenced by baseline quality level: the practices with lower baseline performance levels showed greater improvement compared with practices with a better quality of care (Lin et al., 2016).

P4P effects on quality outcomes:

- Process of care
- Intermediate outcome
- ED or Hospital admissions

One of the latest reviews that exclusively focus on the QOF role in long-term (chronic disease) care found a modest reduction in emergency admission rates (mainly driven by coronary heart disease), a modest increase in consultation rates in severe mental illness, and modest improvements in certain limited aspects of diabetes care (Forbes et al., 2017).

A systematic review by Yuan et al. included twelve studies in the effect analysis and found moderate-certainty evidence that adding of P4P to an existing payment method (capitation or different kinds of input-based payment) slightly improved the care provided by health professionals compared with the existing method. Comparison of P4P plus capitation versus fee for service included one randomized trial in China showing that capitation combined with P4P targeting control of antibiotic prescriptions led to a reduction of antibiotic prescriptions in village and township health facilities (Yuan et al., 2017).

The other systematic review authored by Patel et al. included overall 13 studies and 7 impact evaluations to investigate the P4P effect on MCH outcomes and quality of care in LMIC (Patel,

2018). The countries of our interest were Argentina, the Philippines and Cambodia. In Philippines bonus payments to physicians resulted in clinical Mean Vignette score increase for child health (by 9.7% points) and positive outcomes among children expressed by averting age-adjusted wasting as well as improvement in general self-reported health (7-9% improvement) over time of the P4P duration (Patel, 2018). This review builds on a previous review conducted by Das and colleagues with a similar objective. The latter paper, among other countries, looked at the Philippines as well. The authors used the same primary studies in the Philippines and report that the P4P scheme increased physicians' knowledge to manage under-five diarrhea and pneumonia and a small improvement in patient-reported health measure for under-five (Das et al., 2016).

Inconsistency in the results was found by a number of studies (Gillam, 2015; Langdown & Peckham, 2014; Mendelson et al., 2017). Mendelson et al. examined P4P effects on health, health care use and process of care by reviewing studies from the UK, the US, Taiwan and other western countries. The authors found low-strength contradictory evidence that the P4P programs may improve process-of-care outcomes over the short-term, while evidence on the longer-term effects is limited. The biggest improvement is seen among practices with poor baseline performance and improvements of process-of-care outcomes were found in early stages which slowed down over time. No clear evidence was found on intermediate health outcomes (Mendelson et al., 2017).

Gupta and Ayles looked at P4P schemes' effect on diabetes outcomes such as patient-level experience and population-level outcomes (hospitalizations, or premature deaths) in single-payer systems. The authors analyzed eight P4P interventions in seven countries and identified two types of incentives: (1) high-powered incentives when additional payments to the achievement of verifiable targets are aligned with policymakers' expectations; and (2) low-powered incentives, where bonuses are not necessarily linked to specific patient-oriented objectives. Schemes in the UK, Taiwan and Sweden belong to the first group of incentives, and Australia, Canada (2 schemes), Denmark and Italy belong to the second group. The review concludes that the P4P schemes tied to physician performance metrics can have important effects that could be attributed to enhanced clinical practices and counselling for patient self-management. This effect was examined in Taiwan and Sweden P4P schemes, where the introduction of the incentives resulted in process and intermediated outcomes improvement. In the contrary, low-powered incentives showed little evidence of improved processes of care, and mixed associations with the risk of diabetes-related hospitalization. In Denmark, the scheme failed to show any effect largely due to small incentives to promote behavior change. The authors suggest that in some settings, P4P implementation occurred in parallel with improved information systems leading to the completeness of data and thus, overestimation of the P4P effect. The authors claim that more research with rigorous evaluations is needed (Gupta & Ayles, 2019).

Scott et al. reviewed 44 schemes (from 80 empirical studies) and summarized their impact on cost and quality in the context of VBP and key design features of these schemes. The review targeted both primary and hospital care levels. From all 44 schemes majority were in the United States (25 of which 15 were implemented by the private sector), followed by the UK, Taiwan and Canada. Actual effects were not summarized due to heterogeneity in the types and number of

outcomes; instead, vote counting was used and the findings were not disaggregated by levels of care or public/private ownership.

Of all outcome measures reported by the 44 schemes, 46% were positive (including expenditures and quality of care). The paper reported slightly higher, although not statistically significant, positive outcomes with schemes targeting primary care. The authors found *that weaker study designs were more likely to show positive effects*, suggesting that as study designs improve, the likelihood of finding stronger effects will be lower. The review found that in terms of the proportion of positive outcomes, the *VBP, a key innovation in the United States, that combines P4P rewards with rewards for reducing costs, did not show better results than P4P alone*. The authors conclude that many shared savings models are in their early stages; and more evidence is required to examine if this persists over time. *Another key finding of the review is that schemes that reward improvements in performance over time have a lower probability of being effective than those that do not*. The latter include single threshold schemes but also other scheme types, such as value-based pricing of DRGs (Scott et al., 2018). Considering that this comparison mentions DRG, we presume that the finding pertains to hospital settings as well (that was the case in 30% of all schemes included in the review). The authors further argue that specific factors that affect behavior were difficult to capture due to heterogeneity and small sample sizes. The review reported Taiwan experience based on nine studies, eight of which showed positive results. The authors also pointed out patient selection bias in the incentive schemes that could lead to biased results (Scott et al., 2018). More details are provided in the country scheme description (see box).

Cattel and Eijkenaar in a systematic review to inform the US VBP reform processes examined the effectiveness of payment initiatives in improving value and described design features of these initiatives. The authors included and compared 18 initiatives (targeting primary and hospital levels) from the US (15), Germany, Netherlands and Spain. Nine initiatives were implemented by private insurance companies, two by public-private partnerships, and seven by public payers. Contracted entities were ACOs, private contractors, or other networks and private groups. The paper describes three methods of linking payment to quality: 1) quality incentives as add-on payment in combination with a provider share of realized savings/losses depending on quality (most common); 2) savings/ losses also depending on quality but no direct add-on payment; 3) only add-on payments. The paper further describes quality measurements and the quality incentive structure of the schemes. Evaluation of five VBP initiatives with difference-in-difference design demonstrated similar or reduced spending growth and equal or improved quality (Cattel & Eijkenaar, 2019).

The primary studies from middle-income countries included in our review examined P4P interventions in Nigeria, Cambodia and Brazil. Mabuchi et al. report about large variations in performance among participating primary care centers under the World Bank funded scheme in Nigeria. Although coverage of institutional delivery was around 10% of catchment population before the intervention in all target centers, high-performers achieved 80–90% coverage while low-performers struggled with 20–30% coverage (Mabuchi et al., 2018). Another study from Nigeria explains that the P4P scheme did not result in health facility performance improvements as the scheme suffered from serious implementation challenges (Ogundeji et al., 2016). The other author reports that interventions that involved supply and demand side incentives in

Nigeria and targeted maternal health services resulted in a 60% increase in maternity services use (Wekesah et al., 2016). Renmans et al. report about a rise in institutional deliveries as an impact of the P4P program, however, no effect on neonatal mortality was observed (Renmans et al., 2016). In Brazil, P4P intervention resulted in improved quality of care for specific conditions and thus reduced hospitalizations for sensitive conditions (Soranz & Pisco, 2017).
The review papers indicate that higher-quality studies, those with control groups and controlling for secular trends fail to confirm the positive impact of P4P (Houle et al., 2012; Scott et al., 2018).

2nd group of quality outcomes

The six studies mention the positive effect of P4P on patient health outcomes (2nd group). All these studies refer to the programs in Argentina, Taiwan, the US, Germany and the UK's QOF scheme (Gertler et al., 2014; Gillam et al., 2012; Gupta & Ayles, 2019; Patel, 2018; Scott et al., 2018).

Scott et al. in an earlier Cochrane review looking at the P4P schemes in the US, the UK (the NHS scheme prior to the QOF) and Germany found modest effect of financial incentive to improve quality of care defined as improved health outcomes and patients self-perceived well-being, noting about the substantial risk of bias for the majority of the studies (particularly self-selection into schemes by physicians) (Scott et al., 2011).

P4P effects on health outcomes:

- Disease prevalence
- Disease specific mortality
- All-cause mortality

There are no disagreements about the positive effect of the Argentinian Plan Nacer program on low-birth weight births and neonatal mortality reduction (19% and 74% reduction respectively in the beneficiary group) (Gertler et al., 2014; Patel, 2018).Taiwan P4P initiative that showed a lower risk of mortality among diabetes patients and all-cause mortality and diabetes-related mortality among patients having survived cancer was also criticized for patients selection bias in the earlier primary studies captured by the review papers (Mendelson et al., 2017; Patel, 2018; Scott et al., 2018), however, later research proves this positive effect (Gupta & Ayles, 2019). As for the QOF, at least one review paper reported a modest reduction in mortality shown by earlier studies (S. J. Gillam et al., 2012), other authors found contradictory results (Peckham & Wallace, 2010) and more recent review papers argue that there is no clear evidence that P4P improves patient health outcomes (Gillam, 2015; Mendelson et al., 2017; Forbes et al., 2017; Gupta & Ayles, 2019) More information is given in the country case descriptions.

3rd group of quality outcomes

P4P influence on *equity* has been addressed by nine papers. The authors, most likely referring to the same primary studies, identified that women and certain ethnic minority groups had not benefited equally from QOF implementation (Alshamsan et al., 2010; Boeckxstaens et al., 2011; So & Wright, 2012). Importantly, after correcting the practice characteristics, the inequalities were reduced, indicating that existing differences between socio-

P4P effects on quality outcomes:

- Equity
- Coordination of care
- Continuity of care

economic groups were mainly due to differences at the practice level (Boeckxstaens et al., 2011). Gillam and colleagues noted that P4P could reduce inequality by reducing the gap between socioeconomic groups. E.g. the gap in median achievement narrowed from 4.0% to 0.8% between 2004 and 2007 while comparing practices from the most deprived and least deprived quintiles in the UK (S. Gillam, 2015). The review on impact of reimbursement systems on equity in access and quality of primary care looking mostly at the UK and the US schemes did not find association between P4P and socioeconomic and racial inequity (Tao et al., 2016). The review focusing on P4P modifying factors found patients' poor socio-economic status and representation of minority groups to be associated with poor P4P performance (Markovitz & Ryan, 2017). The same authors indicate some suggestive evidence that higher density and rurality may harm response to the incentives. As for other patients' factors like age, gender, or patients' health, the reviews conclude that evidence on P4P performance is not consistent (Markovitz & Ryan, 2017; Mendelson et al., 2017).

Continuity of care, coordination between team members, satisfaction. Earlier studies on QOF indicate that P4P might have changed the nature of the practitioner-patient consultation through declining personal/relational continuity of care between doctors and patients (Boeckxstaens et al., 2011; Latham & Marshall, 2015). In the systematic review, Mendelson and colleagues mention studies from the UK (from 2012-214) and from the US (2016) documenting concerns related to the considerable burden on health workers related to reporting on measures (the UK, the US) and threatening clinical autonomy (the UK) (Mendelson et al., 2017) .One of the latest systematic reviews examining the QOF indicates that there is no evidence that the scheme influences positively or negatively, other aspects of care, such as integration or coordination of care, holistic or personalized care, or self-care, nor any evidence of its effects on patients' quality of life, experience, or satisfaction (Forbes et al., 2017). Another study on incentive scheme in Brazil resulted in improved practices of coordination of care at PHC level, however, P4P was part of a larger reform including investments in an information system, equipment, capacity building (Soranz & Pisco, 2017).

The studies found that recognition of the contribution of all team members encourages smooth collaboration and communication (Korda & Eldridge, 2011). The QOF experience suggests that even though achieving the targets required a team effort, the doctors who owned the practice received the bonus payments themselves, which led to resentment by the other team members and may have altered nurse-patient interaction (Latham & Marshall, 2015).

4th group of quality outcomes

Wiysonge and colleagues provide the classification of unintended effects of P4P schemes that may include: gaming (i.e. inaccurate or false reporting); adverse selection (i.e. excluding high-risk people from care to obtain better performance) and distortion (i.e. ignoring essential tasks that are not rewarded with incentives) (Wiysonge et al., 2017).

Unintended effects of P4P schemes may include:

- gaming
- adverse selection
- distortion

We define *"gaming"* as exception reporting, that is, exclusion of patients from denominators to improve percentage target achievement, falsifying of data, and measurement fixation. Three systematic reviews and one review paper reported about this event

out of all selected papers: Gaming by over exception reporting and over classifying patients was not widespread in the UK QOF scheme (median, 6%) (S. J. Gillam et al., 2012; Houle et al., 2012; Van Herck et al., 2010). Misreporting was also found in Cambodia P4P scheme (Renmans et al., 2016).

The adverse selection occurs when doctors prefer to treat patients with a milder disease condition or better socioeconomic status, which not only intensifies the inequity but also is likely to exaggerate the improvement of clinical performance. Eight systematic reviews examined the adverse selection of P4P. Taiwan confirmed the results that primary practices with a lower baseline level of medical quality tended to exclude patients with a severe condition, so as to show great promotion in clinical performance apparently (Lin et al., 2016).

Based on the studies conducted in 2009 - 2011, concerns have also emerged that patients from disadvantaged and vulnerable populations may be disproportionately excepted from the QOF because their diabetes may be more challenging to manage. Patients with longstanding diabetes or multiple comorbidities were also more likely to be excluded from the A1C indicator. The same study of 2011 and other earlier studies found that QOF does not address ethnic disparities in diabetes care adequately (Peckham & Wallace, 2010; Latham & Marshall, 2015)(Peckham & Wallace, 2010; Latham & Marshall, 2015a). However, one systematic review claimed that patients' adverse selection in the UK QOF scheme was unknown (Houle et al., 2012). P4P participants find ways to maximize measurable results by skimming of healthier patients for treatment by physicians (Korda & Eldridge, 2011).

Distortion, which is ignoring unincentivized activities to perform, was discussed in seven systematic reviews and one single study. A potential problem here is that P4P could lead to the neglect of those non-incentivized areas of care, which continues to rely on the professionalism or moral motivation of medical professionals participating in P4P. There is some evidence of concern amongst general practitioners that non-incentivized areas like acute care, preventive care, care for specific groups such as children or older people and care for patients with multiple comorbidities would suffer as these professionals chased targets (Mendelson et al., 2017; Yuan et al., 2017; Langdown & Peckham, 2014; Peckham & Wallace, 2010; Boeckxstaens et al., 2011; Alshamsan et al., 2010; Korda & Eldridge, 2011; Gertler et al., 2014).

Four systematic reviews mentioned ***positive spillover effects* of P4P**. Three out of these four reviews discussed this effect in conjunction with the quality of care: The study in Argentina found an overall 22% reduction in neonatal mortality (beneficiaries and non-beneficiaries) using the same clinics (Patel, 2018). Other studies found positive effects on P4P targets concerning coronary heart disease, COPD, hypertension and stroke when applied to non-incentivized medical conditions (10.9% effect size) (Van Herck et al., 2010) or improved immunization coverage for non-incentivized vaccines (Gleeson et al., 2016). In addition to the intermediate and patient outcomes, rates of recording were also found to have increased for all the various groups of patients used (Allen et al., 2014). In the US, there was evidence of a reduction in expenditure growth for Medicare patients who were not covered, but who were enrolled with the same provider organizations participating in the Alternative Quality Contract (Scott et al., 2018).

Interestingly, there is evidence that when the measures are no longer incentivized, improvement is sustained or continued. The studies in the US found that after incentives removal, all

improvements were sustained for up to three years. Similarly, a QOF study indicated that the level of performance achieved prior to the incentive withdrawal was generally maintained (Kondo et al., 2016).

Implementation considerations

The P4P interventions vary widely by its design features, contexts where they operate, cultural factors, implementation specifics, etc.

Lack of understanding, perception and acceptance of P4P interventions by healthcare personnel can undermine the potential impact of P4P schemes by limiting the behavioral response of health workers (Patel, 2018). It is therefore important that providers are actively involved in designing the program, especially in developing and maintaining the aspects of performance to be measured. This increases the likelihood of provider support and alignment with their professional norms and value (Saddi & Peckham, 2018). The authors found that the schemes showed better results when providers were involved and collaborated on the scheme development (Allen et al., 2014; Kondo et al., 2016).

A systematic review examining implementation processes through primary studies review and experts interview found that incentives linked to measures of clinical quality (a process of care and clinical outcomes) may inspire more positive change than programs using measures targeted efficiency or productivity (Kondo et al., 2016).

Small practices demonstrate better results compared to big practices as per QOF experience. Although the overall quality of care was below the required level, process indicators of P4P, such as physicians' prescriptions of examination or drugs, management of chronic obstructive pulmonary disease, diabetes, hypertension and coronary heart disease, improved more in smaller practices with a higher proportion of female and younger physicians. However, another study showed that a sufficient number of staff (physicians, nurses, and administrative personnel) in big cities create better conditions for physicians to manage chronic diseases. Nurses are dealing with urgent diseases and physicians have more time managing chronic diseases (Lin et al., 2016). The studies examining the USA schemes found that larger practices outperformed smaller practices (Markovitz & Ryan, 2017).

P4P implementation and outcomes are affected by the baseline level of facility performance. Studies demonstrated that the practice with a better quality of service before improved less than the practices with worse baseline (Lin et al., 2016). Interestingly, low performing practices do not give up when the targets are unrealistically far to reach (Markovitz & Ryan, 2017).

Indicators selection plays a critical role. Implementation of too many indicators can lead to increased bureaucracy and administrative work instead of spending time with patients. The UK QOF scheme with 134 indicators raised concerns about health workers' administrative workload (Kolozsvári et al., 2014).

The incentive structure is one of the critical factors to be considered during the program design (Gupta & Ayles, 2019, 2019; Houle et al., 2012; Kolozsvári et al., 2014; Kondo et al., 2016; Paul & Renmans, 2018; Peckham & Wallace, 2010). Incentive structure needs to carefully consider several factors, including incentive size, frequency, and target (Kondo et al., 2016; Yuan et al.,

2017). Research in the USA showed that although the size and structure of incentives do seem to be important in promoting effective physician activity, financial incentive may be less important than giving public recognition for improved quality (Gillam, 2015).

There is evidence that compared to larger incentives, small incentives were associated with greater improvement in provider communication and interaction with the patient (Kondo et al., 2016; Scott et al., 2018). However, too low incentives are not likely to be effective. According to Scott and colleagues, contrary to expectations, the size of the incentives as a percentage of revenue was not associated with the probability of an effect (Scott et al., 2018). According to Kolozsvari et al., there are different opinions. No exact, universal percentage can be established for incentive calculation in different countries, but an increase of at least 5-10% could be appropriate (Kolozsvári et al., 2014). In Europe, the bonus amount ranges between 1-25% of the total income of the practice (Kolozsvári et al., 2014). In the UK, family doctors' practices receive 10-15% of their income from the QOF scheme (Forbes et al., 2017). Peckham and colleagues argue that incentives have to be large enough to influence behavior and designed in such a way that they cannot be played off so as to reward both process and improved outcomes (Kondo et al., 2016; Peckham & Wallace, 2010). Too high incentives can cause unintended consequences (e.g. data manipulation, "gaming"/cheating) (Kolozsvári et al., 2014). In Cambodia, financial incentives accounted for 42% of the average total income of a health worker and were associated with higher job motivation (Paul & Renmans, 2018). Although high incentives were paid to health workers in Cambodia, unintended consequences of the scheme were averted thanks to regular monitoring, random verification and web-based reporting (Renmans et al., 2016).

A ceiling effect was shown by earlier studies in Argentina and in the UK, indicating that after achieving the upper limit, quality would reach a plateau and no further improvement is observed. Even more, other indicators, unrelated to payment, saw a drop to a certain degree (Lin et al., 2016).

Monitoring and verification are essential to ensure P4P meets predetermined objectives that are the predetermined quantity and quality targets (Kandpal, 2016; Patel, 2018). Best practice requires that monitoring be implemented by an independent agent for improved autonomy in issuing penalties for poor performance (Khim et al., 2018). At the same time, the studies identify the administrative burden associated with this function (Renmans et al., 2016).

Performance feedback to providers and managers facilitates performance improvement (Kandpal, 2016; Patel, 2018; Saddi & Peckham, 2018). It is suggested that the 'easier' structural quality indicators are addressed first and then programs can move onto introducing process measures of clinical care. This allows health providers to address less complex quality of care issues first, develop a better understanding of bonus scheme and quality of care, and then shift gradually toward more demanding measures of care under the P4P programs (Kandpal, 2016; Patel, 2018).

The study examining pathways to high and low performance of P4P intervention in Nigeria identified contextual factors such as staffing, access and competition with other providers, management including system of accountability, various measures to improve staff motivation and team-work drove performance improvement among the primary care centers. Interestingly, drivers leveraged positive contextual and health system factors and mitigated negative factors (Mabuchi et al., 2018).

Selectively rewarding primary care physicians and leaving nurses may discourage teamwork and coordinated care with other members of the team (Latham & Marshall, 2015).

Our review was not specifically looking at P4P effects on provider behavior, namely, satisfaction and motivation, although we bring evidence where such patterns were described. According to Ogundeji et al., poor motivation of health workers was caused by a combination of factors such as poor salaries, poor working conditions, inadequate infrastructure and limited opportunity for career development or training, lack of government ownership of this health financing mechanism, lack of understanding of the P4P scheme, delayed incentive payments (Ogundeji et al., 2016).

The UK, QOF

The UK's Quality and Outcomes Framework (QOF) is the world's largest P4P scheme in primary care, which was introduced in 2004. QOF was designed to provide a mechanism to motivate GPs and to increase funding for their practices. The QOF is voluntary, by 2019, nearly 95% of practices in England participated in the scheme.

In the early years of introduction, QOF addressed four domains of care: clinical care in 10 areas, organizational aspects of care, patient experience and additional services such as cervical screening and reported on 147 indicators developed by the National Institute of Health and Care Institute. QOF has undergone a series of reforms and developments. A decade after the introduction, QOF focused on other challenges. Total number of indicators reduced to 77 indicators focusing primarily upon clinical aspects of care and public health. Incentives consisted of 15-20% of total practice income upon implementation and reduced to approximately 8% as of 2018 (NHS, 2019).

Much of the academic work on QOF was conducted during the first five years of its introduction. The studies examined incentives scheme effects as well as implementation features. Diabetes care was extensively studied under QOF. Almost all studies reported that performance improved in incentivized areas. Use of evidence-based methods for better management of chronic patients such as improved use of computers and patient records, decision support, clinicians prompts and patient reminders lead to improved intermediate outcomes immediately after the QOF introduction (Roland & Guthrie, 2016). (S. J. Gillam et al., 2012). Data from the first year of the QOF showed that practices in less deprived areas delivered a higher quality of care, in addition, the practices with lower performance have a greater chance to improve rather than practices with already high performance. By the third year of the QOF, the performance gap between the most and least deprived had disappeared (Allen et al., 2014; Latham & Marshall, 2015). Important to note that quality improvement in diabetes care was observed prior to the QOF introduction. The trend observed at the beginning of the initiative slowed down in 2007 and no performance improvement was observed in later years (Latham & Marshall, 2015).

Nationwide expansion of the QOF scheme led to the absence of the control sites for adequate experimental studies. Majority of the study use Interrupted Time Series design. Methodological challenges of the studies and confounders not fully examined also hampered interpretation of findings (Latham & Marshall, 2015).

The majority of unintended consequences of QOF have emerged following the introduction of indicators that were not aligned with core professional values and increasingly been described as becoming a 'tick box' exercise (NHS, 2019).

The systematic review examining the QOF effect on long term care (NCDs) published in 2017 found no convincing evidence that the QOF promotes better care and outcomes for people with long-term conditions (Forbes et al., 2017). The authors argue that even the indicators were based on high-quality evidence of the effectiveness of interventions, other wider determinants of health may play a role, or non-incentivized activities are more important in determining mortality in the patient population. The review identified a modest reduction in emergency admission rates in long-term conditions (both for QOF and non-QOF conditions), and modest improvements in certain limited aspects of the care of diabetes, however, whether these changes were attributable to the QOF was difficult to determine as an introduction of new standards and technologies may have driven such improvements as well (Forbes et al., 2017). The same systematic review reports that there is no evidence to suggest that the QOF influences positively or negatively integration or coordination of care, holistic or personalized care, or self-care, nor any evidence of its effects on patients' quality of life, experience, or satisfaction (Forbes et al., 2017).

In 2017 the NHS England undertook a review of the QOF and number of improvements were incorporated in line with its recommendations.

The US, Value Based Purchasing

In the US, P4P initiatives were introduced into the traditional reimbursement system from the 1990s and rapidly diffused within private and public health insurance plans (Gemmill, 2007). Currently, P4P is part of the Value-based purchasing (VBP) strategy. VPB includes incentives to reward quality increase combined with rewards for slowing expenditure growth (shared savings) (Scott et al., 2018). The quality aspects may include structure (e.g. labour, facilities, and materials), process, outcome, or a combination of all three types of measures (Gemmill, 2007). VBP is applied by both public (e.g. Medicaid, Medicare) and private payers (e.g. Integrated Health Care Association Program, Bridges to Excellence program, etc.). VBP is currently tested under Accountable Care Organizations (ACOs). ACOs were introduced by the Affordable Care Act into Medicare since 2010, and a range of private ACOs was established. The key design feature of ACO model is that the providers are rewarded for an improvement in performance between two time points, like directly measuring a change in performance, as well as schemes that have more than one threshold so that providers can move to a higher threshold over time. Risk-sharing could be one-sided (where providers share in any savings) and two-sided, when providers share risk for deficits (Scott et al., 2018).

Despite P4P widespread use, they typically make up a small proportion of provider reimbursement. Most payers only put 5% or less of provider compensation at risk of profit or loss from the P4P system (Gemmill, 2007).

The systematic review of Scott et al. that analyzed 25 schemes from the US (from all 44 schemes included in the review) presents an example of a two-sided, private scheme that showed an impact on both reducing spending and improvements in quality after 4 years of its implementation. There was no positive change about ED admissions and pharmaceutical expenditures. Medicare ACO (public scheme) showed some evidence of reductions in the growth of spending and the patient experience was no worse than before, however, none of the studies examined effects on other measures of quality of care (Scott et al., 2018). The review found that one- and two-sided risk-sharing models which combine rewards for P4P with rewards for reducing costs, yet seems no better than P4P alone in terms of the proportion of positive outcomes (10% lower for only P4P but not statistically significant difference). However, these shared savings models are in their early stages; therefore more evidence is required to examine if this persists over time (Scott et al., 2018).

There are a number of factors that influence the probability of the scheme having an effect that were not captured by the studies. The list includes unobserved factors related to how the scheme was developed, the extent to which scheme participants were involved, and the extent of already existing quality improvement initiatives and public reporting (Scott et al., 2018).

Cattel and Eijkennar, in their systematic review, analyzed the effectiveness of VPB initiatives in public and private programs and demonstrated that these initiatives generally show promising results in terms of lower spending growth with equal or improved quality (Cattel & Eijkenaar, 2019).

Argentina, Plan Nacer

Plan Nacer was launched in 2004 following the deterioration in maternal and child health indicators resulting from the 2001 economic crisis.

Plan Nacer was designed to improve the health status of uninsured pregnant women and children by channeling more resources to the public health care system and creating incentives to use those resources more efficiently. The program covers women during pregnancy and up to 45 days after birth (or the loss of the fetus) and children up to age six and concentrates on services during the first year. All other care outside of the Plan Nacer package of benefits is covered by regular provincial health services (Gertler et al., 2014)

Plan Nacer represents additional funding beyond the historical administrative budgets and supplements the existing public financing system with an innovative P4P model that incentivizes the provision of quality priority maternal and infant health services. Through Plan Nacer, the national government reimburses provinces on a per capita basis at a maximum cost of $8 per person per month. The provinces receive $5 (60 percent of the maximum per capita payment) for every eligible individual enrolled in the program and up to an additional $3 (40 percent of the maximum payment) if health targets for the eligible population are achieved. Thus, the program provides explicit incentives to enroll the target population of uninsured mothers and children and to provide services that improve the health outcomes of the eligible population (Gertler et al., 2014).

General guidelines for the use of resources by providers are set at a national level, and provinces are allowed to impose additional restrictions for service providers in their jurisdictions. Resources may be used at the discretion of the provider to improve the quality of health services (Gertler et al., 2014).

Impact Evaluation (IE) found that scheme had a positive impact on the patient health outcomes; specifically, beneficiaries have a 74% lower chance of in-hospital neonatal mortality in larger facilities, it improved toxoid vaccine uptake as well as an increase in the number of prenatal care visits in general (Gertler et al., 2014; Kandpal, 2016).

Another evaluation focused on the Misiones province and suggested that the rate of early initiation of prenatal care was higher in the intervention group compared to the control one. The study found that large short-term incentives appeared to be more cost-effective concerning the motivation of the providers, rather than permanent incentives with fixed costs in terms of changing clinical practice (Kandpal, 2016).

Another P4P initiative in Argentina was implemented in 2005 in Buenos Aires as part of multimodal intervention. It aimed at family physician's performance quality improvement with incentives accompanied by continuous education, audit and feedback. The study examining this initiative two years after its introduction found that clinical effectiveness improved across all indicators (e.g. cancer screening, blood pressure measurements, cholesterol levels, etc.), but performance on comprehensive practices showed contradictory results. There was a significant improvement in the detection and management of depression, but well-child visits targets decreased. Insignificant improvement was found in the documentation of relevant data and coordination of care of family physicians (Rubinstein et al., 2009).

Taiwan, P4P diabetes program

A diabetes P4P program was introduced in 2001 by Taiwan's National Health Insurance Administration (NHIA) to improve the quality of health care for diabetes patients. The scheme has the following features: First, only physicians who specialize in metabolic disorders or endocrinology or who attended a training program for diabetes care are eligible to participate in and voluntarily enroll patients into this special P4P for diabetes care. Second, medical care teams are expected to work as coordinated physician-led multidisciplinary teams adhering to the American Diabetes Association's clinical guidelines. Third, in addition to regular and usual care, P4P patients received extra comprehensive care, including medical history assessment, physical examination, laboratory evaluation, management plan evaluation, and self-management and health education. Fourth, participating P4P physicians receive extra incentive payments in addition to regular physician fees depending on incentive targets for improving processes (e.g. documented HbA1c or LDL tests) and intermediate outcomes (e.g. higher percentages of patients with controlled HbA1c or LDL) (Hsieh, Chiu, et al., 2017).

Taiwan P4P experience is based on several cross-sectional studies that were captured by different review papers (Gupta & Ayles, 2019; Latham & Marshall, 2015; Lin et al., 2016; Scott et al., 2018). The majority of the studies showed positive effect indicating that patients enrolled in the program were more likely to receive guideline-recommended tests and examinations related to diabetes care, cancer screening and quality care, tuberculosis treatment adherence and lengths. Some of the studies showing a positive spillover effect. According to a longitudinal study, published in 2010, the effect of the Taiwanese program on hospitalization rates found that patients enrolled in the incentive program were less likely to be hospitalized after 3 years of care compared with non-enrolled patients (Cheng et al., 2012). In the contrary, one later study found an increase in emergency admissions for diabetic patients (Scott et al., 2018).

The authors of the systematic reviews (Mendelson et al., 2017; Scott et al., 2018) note that Taiwan scheme findings should be treated with caution as there may have been substantial selection bias of patients enrollment in the schemes so that positive effects could have been due to selection rather than the impact of the program. However, the l*ater studies, after controlling selection bias, found that P4P increased physician continuity of care among patients with diabetes that in turn was associated with a lower risk of mortality* (Pan, Chien-Chou et al., 2017), other authors (Hsieh, Chiu, et al., 2017; Hsieh, He, et al., 2017) reported significantly lower risks of cancer-specific mortality in newly diagnosed cancer and reduction of the 5-year risk of all-cause mortality and diabetes-related mortality among patients having survived cancer (Gupta & Ayles, 2019).

Discussion

Our aim, initially, was to synthesize evidence on P4P effectiveness on utilization and quality of primary care in private settings in middle and high-income countries. Our review could not identify review studies that compare P4P effectiveness through public/private ownership lens. Although P4P interventions are implemented in diverse contexts by both public and private payers and providers, there is a lack of primary studies, and consequently reviews, examining comparative effectiveness of public and private players. Therefore, we slightly modified the research question by dropping the "private settings" dimension.

The review found that there was significant heterogeneity in terms of the contexts in which the P4P schemes were implemented, services and populations targeted, types of outcome measures and incentives used. Most of the P4P interventions targeted preventive care, management of chronic and MCH conditions. The review papers are dominated by studies from the UK (QOF scheme) and the US.

Although some systematic reviews showed contradictory outcomes on ***PHC service utilization***, P4P was found to be an effective intervention scheme to increase the utilization of preventive care services for MCH. The studies examining P4P effect on MCH care in middle-income countries documented positive results such as increased utilization of antenatal care services and provision of antenatal tetanus toxoid in Argentina, positive clinical outcome for children under-5 in the Philippines, and small but positive effect on antenatal care and vaccination in Cambodia (Gertler et al., 2014; Patel, 2018). In Nigeria, where the P4P package covers a wide range of services the studies show mixed results with significant changes in high-performer primary care centers and minor or no improvements in poor performers (Mabuchi et al., 2018; Ogundeji et al., 2016). With regards to childhood immunization, the P4P scheme implemented in three states of Nigeria resulted in the increase of average coverage for completely vaccinated children from 1.4% to 49.2% during two years period (Odutolu et al., 2016). The primary studies in high-income countries and specifically in the US showed from modest to large improvements in childhood immunization coverage with a greater effect when starting point was low (Fairbrother et al., 2001; Gleeson et al., 2016). Evidence on the utilization of screening services for Chronic Diseases and cancer showed inconclusive results.

To summarize evidence on the ***effectiveness on quality of care.*** We tried to stratify the quality indicators by four categories, nevertheless, heterogeneity of the outcomes and evaluation approaches complicate the synthesis of the results.

Earlier systematic reviews and reviews of the systematic reviews reported insufficient evidence on the effectiveness of P4P interventions in improving quality of care (Scott et al., 2011; Houle et al., 2012; Witter et al., 2012; Eijkenaar et al., 2013). The major deficiency was related to a lack of studies with strong designs to control for observable and unobservable factors and time trends. Literature has grown considerably since that time and studies with more robust designs have been proliferated.

In general, as noted by the review papers, results from the studies with rigorous methodology tend to show less positive results compared with the studies with weak designs (Eijkenaar et al., 2013; Houle et al., 2012; Mendelson et al., 2017; Scott et al., 2018).

Our review showed that there is more consistent evidence that P4P schemes improve *process of care outcomes*. Lin and colleagues were more positive and report on significant improvement on

CHD and diabetes process of care measures, while Mendelson concludes that there is low-strength contradictory evidence that the P4P programs improve process-of care over the short-term, while evidence is limited on long term outcomes. Both authors conclude that the largest improvements are seen in areas where baseline performance was poor (Lin et al., 2016; Mendelson et al., 2017).

Mixed evidence was found with regards to *intermediate and proxy outcomes* such as emergency department and hospital admissions due to aggravation of chronic conditions, institutional deliveries. Gupta and Ayles showed that the schemes tied to physician performance metrics could have an important effect in limiting disease progression over the long term (Gupta & Ayles, 2019). The latest review focusing on the QOF found a modest reduction of emergency admissions after coronary heart disease (Forbes et al., 2017). The study in Nigeria demonstrated large variations in institutional delivery increase among three participating states with similar starting point (Mabuchi et al., 2018).

Less favorable results concerning patient-level outcomes compared to process of care outcomes could be explained by the fact that the processes precede improvements in outcomes and could not be captured by the studies, or that when more attention is diverted, it is easier to influence measures (Vlaanderen et al., 2019).

The evidence on the incentives effect on *health outcomes* such as disease prevalence, disease-specific or overall mortality is limited. The robust impact evaluation of Argentina Plan Nacer program demonstrate a significant reduction of low-birth weight births and neonatal mortality (Kandpal, 2016). The scheme in Taiwan was successful in the reduction of diabetes-related mortality (Mendelson et al., 2017). While earlier QOF studies demonstrate modest reduction in mortality, the recent paper argues that there is no clear evidence that P4P improves patient health outcomes under this scheme (Mendelson et al., 2017).

Eijkenaar et al., in the review of systematic reviews, concluded that although the P4P long-term effect on *inequalities* remained largely unknown P4P seems to have narrowed socio-economic inequalities (Eijkenaar et al., 2013). The later reviews indicate either inconsistent results or no association between P4P and socioeconomic and racial inequity (Tao et al., 2016).

There is inconclusive evidence on how P4P influences positively or negatively *continuity of care, coordination* between the health workers. A scheme design defines interprofessional collaboration. Selective rewarding may discourage teamwork and coordinated care, while recognition of all team members' contributions stimulates smooth collaboration (Korda & Eldridge, 2011; Latham & Marshall, 2015). Earlier QOF studies report about reduced clinical autonomy following the scheme introduction. At the same time, the later review states that there is no evidence that the scheme influences positively or negatively on coordination of care, holistic or personalized care, patients satisfaction (Forbes et al., 2017).

As P4P proliferates, questions have arisen about its *unintended negative consequences*. Few studies report about gaming practices in the P4P schemes where physicians try to manipulate with the data in order to prove achievement of certain indicators and be eligible for incentive payments. Gaming was not widespread in the QOF scheme, although some cases were reported (S. J. Gillam et al., 2012). Cambodia P4P scheme decreased gaming by implementing regular monitoring and random verification and the availability of web-based reporting (Renmans et al.,

2016). Adverse selection of patients and distortion have also been examined in certain reviews. Skimming of healthier patients for treatment by physicians and concentration on incentivized activities only where mentioned as an unintended consequences of P4P schemes (Alshamsan et al., 2010; Korda & Eldridge, 2011; Peckham & Wallace, 2010; Van Herck et al., 2010).

There is evidence on *positive spillover effects,* with some studies finding improved performance on unincentivized measures or medical conditions, improved intermediary or health outcomes in non-target populations (Allen et al., 2014; Patel, 2018; Scott et al., 2018). Moreover, the sustained effect has been demonstrated by some of the studies when the improvements maintained at the level achieved prior to the discontinuation of incentives or continued thereafter (Kondo et al., 2016).

P4P is a complex intervention and its effect is influenced greatly by different factors like scheme design, internal and external factors. Many P4P programs have evolved over time by adjusting design, introducing a mechanism to mitigate spillover effects, adding quality improvement and cost parameters to achieve desired goals. These process changes are not captured by experimental or quasi-experimental studies and thus largely remain unknown to researchers. Contextual and health system factors have been identified as drivers of performance improvement in Nigeria. Importantly, it was found that the drivers influence each other and leverage positive factors (Mabuchi et al., 2018).

It is well known that pure economic theory could not explain all the nuances of the financial incentives effect (Himmelstein et al., 2014). Behavioral economics provide a better understanding of P4P. However, evidence is not always consistent with the behavioral economics concepts (Emanuel et al., 2016). As an example, contrary to the "goal gradients" effect, low performing practices do not necessarily give up when there are unrealistically far from the threshold. The other example from the US studies when bonus size increased and became more achievable the practices did not appear to respond more robustly than other practices with lower benefits and harder to reach threshold, which was opposite to the "threshold effect" when providers would try harder to reach the target when they are close to it (Markovitz & Ryan, 2017; Emanuel et al., 2016).

A key challenge to researchers is to successfully separate the P4P effect from parallel interventions that may have overestimated the effect (Gupta & Ayles, 2019; Markovitz & Ryan, 2017). Moreover, the introduction of a new financial model is accompanied by supporting quality improvement interventions (trainings, guidelines, job-aids, new information systems), or the model could be part of a larger reform processes; therefore causality is difficult to establish (Soranz & Pisco, 2017). The same challenge was pointed by Eijkenaar and colleagues in the review of systematic reviews suggesting that rigorous evaluations are needed to disentangle the effects of different components to draw firm conclusions (Eijkenaar et al., 2013). The other challenge is that empirical studies do not sufficiently describe contextual information where the schemes are implemented and lack comparison of program design characteristics (like institutional settings, resources, incentive size, frequency, structure, etc.) (Eijkenaar et al., 2013; Markovitz & Ryan, 2017). Due to this deficiencies, our review was not able to examine P4P effectiveness among private health care providers.

Most of the studies included in the review papers were conducted in high-income countries. We have broadened our review by including high-income settings as breadths of evidence is coming

from these countries that might be informative for Georgia and other middle-income countries. No firm conclusion could be done about the direct applicability of these findings. However, there is less uncertainty about the transferability of findings from high-income to middle-income levels than vice-versa (Wiysonge et al., 2017).

The majority of the studies focus on the UK and US examples. The UK has considerable experience in P4P programs through its QOF initiative. Higher-quality studies that account for time trends failed to replicate positive effects shown by studies with less strong designs (Mendelson et al., 2017), and there is no sufficient data to conclude long-term effects (Mendelson et al., 2017).

The United States has a long history of P4P implementation and a strong private sector. The US P4P program has been testing the value-based approach (when rewards for quality and costs are linked) which is a unique model for this country. The studies show that process of care outcomes improved in a medium-term period (4-years), although there is limited evidence on long term effects. The studies in the US indicate that greater improvements include culture change interventions along with the P4P and clinical support tools (Kondo et al., 2016). Compared to the UK QOF, incentives are lower in the US P4P schemes (Mendelson et al., 2017) and some experts argue that this could have contributed to better success of the UK schemes (Kondo et al., 2016). Distinctions between the UK and the US P4P programs (such as purchasing systems, ownership of payers and providers, incentive size) preclude strong comparative judgment.

Our review has several *limitations*. Our search for review papers was comprehensive, however less systematic for individual studies as we included primary studies from middle-income countries only (after the last systematic review) and not for high-income countries. We may have missed papers that have not explicitly mentioned primary care in the title/abstract, as per our inclusion criteria, but have discussed it in the body text. The overlap of primary studies in the reviews is significant. We did not check for potential inaccurate representation of the studies by review paper authors. Acknowledging this limitation, we tried to mention earlier and later reviews in the narrative, however, did not make strict distinctions by their primary studies publication periods. We did not assess the quality of the studies and did not do a sub-analysis. Our results are limited to a narrative synthesis. And lastly, our review did not look at motivational factors and the cost-effectiveness of P4P.

What should be considered during P4P design and Implementation

Due to heterogeneity of the findings, it is difficult to draw firm conclusions and recommendations, however there are several patterns that could help health system planners in P4P program design and implementation.

Providers' buy-in is crucial. Engagement of providers in scheme design and alignment of measures with their professional norms and values improves scheme acceptance and could demonstrate bigger change (Eijkenaar et al., 2013; Patel, 2018). Evidence shows that measures that are clinically important lead to better results, rather than measures targeted productivity and efficiency (Kondo et al., 2016).

Incentive structure should consider size, frequency and target. At this point, optimal payment structure (size) still remains uncertain (Gupta & Ayles, 2019). It should be large enough to motivate behavior, but not too high to cause gaming and be not cost-effective (Kolozsvári et al., 2014). Non-financial incentives like giving public recognition for improved quality, was also found important (Gillam, 2015).

Incentives should acknowledge the contribution of clinical and other team members. The evidence shows that communication between the team members is influenced by the fair distribution of rewards between them (Latham & Marshall, 2015). Delayed incentives lead to poor motivation of health workers (Ogundeji et al., 2016). As no consistency exists on the optimal frequency of incentives, behavioral economy principles should be taken into account, suggesting that due to the "immediacy factor," people respond more strongly to immediate benefits, while tend to discount delayed benefits (Emanuel et al., 2016).

The incentive should be *flexible* enough, evaluated continuously and measures adjusted (Kondo et al., 2016). Careful consideration of measures is critical to balance between *sufficient number* to measure quality and avoid overburden of reporting. At the same time, the sufficient number of measures is still unknown. Complexity of incentive structure and poor understanding of the P4P scheme leads to poor motivation and ultimately could fail to promote quality improvement (Markovitz & Ryan, 2017; Ogundeji et al., 2016).

When planning the P4P schemes, one should bear in mind that practices with *low baseline performance* appear to show better results, and far-reaching thresholds do not preclude the practices from performance improvement. Therefore, P4P should target areas of poor performance and de-emphasize areas that achieved high performance (Mendelson et al., 2017). The studies indicate that achieved high performance could be sustained following de-incentivization (Kondo et al., 2016).

There is no consistent evidence about *practice size* and quality improvement under incentive programs. While there is still room for quality improvement in small practices, confounding factors such as smaller staff (particularly nurses) could play a role (Lin et al., 2016). In the contrary, large facilities may have more potential for improvement. For instance, one of the latest systematic reviews looking broadly at outcome-based payment models found that large private providers with initially poor quality scores tend to show better improvements than other providers (Vlaanderen et al., 2019).

High certainty evidence proved that patient characteristics, such as low income and attribution with minority groups (race/ethnicity) is associated with worse P4P performance (Markovitz & Ryan, 2017). This finding suggests that in order to achieve desired patients' outcomes for vulnerable population groups with different socio-economic needs, social interventions should also be in place along with P4P.

The programs should have well developed *electronic records and claim system*, or specially developed data system to evaluate P4P effects (Yuan et al., 2017). *Verification* of reporting is essential to avoid manipulation with the results. *Accountability* and *performance feedback* to managers and providers is crucial to facilitate performance improvement.

Conclusion

P4P programs have likely been effective in increasing the utilization of care and the process of care outcomes. Although there are successful examples of improving patient-level outcomes, in sum, evidence on P4P long-term effect is limited. Heterogeneity of evidence does not allow to conclude that provider-targeted financial incentives have failed to improve the quality of care. To fully realize its potential in quality improvement P4P programs need to be carefully planned, implemented and rigorously evaluated. Consideration of important preconditions suggested by theoretical concepts and empirical evidence helps P4P programs to achieve desired goals.

References

Akbari, A., Mayhew, A., Al-Alawi, M. A., Grimshaw, J., Winkens, R., Glidewell, E., Pritchard, C., Thomas, R., & Fraser, C. (2008). Interventions to improve outpatient referrals from primary care to secondary care. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD005471.pub2

Allen, T., Mason, T., & Whittaker, W. (2014). Impacts of pay for performance on the quality of primary care. *Risk Management and Healthcare Policy*, *7*, 113–120. https://doi.org/10.2147/RMHP.S46423

Alonso, J. M., Clifton, J., & Díaz-Fuentes, D. (2015). Did New Public Management Matter? An empirical analysis of the outsourcing and decentralization effects on public sector size. *Public Management Review*, *17*(5), 643–660. https://doi.org/10.1080/14719037.2013.822532

Alshamsan, R., Millett, C., Majeed, A., & Khunti, K. (2010). Has pay for performance improved the management of diabetes in the United Kingdom? *Primary Care Diabetes*, *4*(2), 73–78. https://doi.org/10.1016/j.pcd.2010.02.003

Basu, S., Andrews, J., Kishore, S., Panjabi, R., & Stuckler, D. (2012). Comparative Performance of Private and Public Healthcare Systems in Low- and Middle-Income Countries: A Systematic Review. *PLoS Medicine*, *9*(6), e1001244. https://doi.org/10.1371/journal.pmed.1001244

Berendes, S., Heywood, P., Oliver, S., & Garner, P. (2011). Quality of Private and Public Ambulatory Health Care in Low and Middle Income Countries: Systematic Review of Comparative Studies. *PLoS Medicine*, *8*(4), e1000433. https://doi.org/10.1371/journal.pmed.1000433

Boeckxstaens, P., Smedt, D. D., Maeseneer, J. D., Annemans, L., & Willems, S. (2011). The equity dimension in evaluations of the quality and outcomes framework: A systematic review. *BMC Health Services Research*, *11*, 209. https://doi.org/10.1186/1472-6963-11-209

Carter, R., Riverin, B., Levesque, J.-F., Gariepy, G., & Quesnel-Vallée, A. (2016). The impact of primary care reform on health system performance in Canada: A systematic review. *BMC Health Services Research*, *16*(1), 324. https://doi.org/10.1186/s12913-016-1571-7

Cattel, D., & Eijkenaar, F. (2019). Value-Based Provider Payment Initiatives Combining Global Payments With Explicit Quality Incentives: A Systematic Review. *Medical Care Research and Review: MCRR*, 1077558719856775. https://doi.org/10.1177/1077558719856775

Cheng, S.-H., Lee, T.-T., & Chen, C.-C. (2012). A longitudinal examination of a pay-for-performance program for diabetes care: Evidence from a natural experiment. *Medical Care*, *50*(2), 109–116. https://doi.org/10.1097/MLR.0b013e31822d5d36

Chikovani, I., & Sulaberidze, L. (2017). *Primary health care systems (PRIMASYS): Case study from Georgia*. World Health Organization; Geneva.

Conrad, D. A., Vaughn, M., Grembowski, D., & Marcus-Smith, M. (2016). Implementing Value-Based Payment Reform: A Conceptual Framework and Case Examples. *Medical Care Research and Review: MCRR*, *73*(4), 437–457. https://doi.org/10.1177/1077558715615774

Cromwell, J., Trisolini, M., Pope, G., Mitchell, J., & Greenwald, L. (2011). *Pay for Performance in Health Care: Methods and Approaches* (1st ed.). RTI Press. https://doi.org/10.3768/rtipress.2011.bk.0002.1103

Curatio International Foundation. (2018). *Results4TB project*. http://results4tb.curatiofoundation.org

Das, A., Gopalan, S. S., & Chandramohan, D. (2016). Effect of pay for performance to improve quality of maternal and child care in low- and middle-income countries: A systematic review. *BMC Public Health*, *16*, 321. https://doi.org/10.1186/s12889-016-2982-4

Donabedian A. (1980). *The Definition of Quality and Approaches to Its Assessment*. (Vol. 1). Health Administration Press.

Eijkenaar, F., Emmert, M., Scheppach, M., & Schöffski, O. (2013). Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy*, *110*(2–3), 115–130. https://doi.org/10.1016/j.healthpol.2013.01.008

Emanuel, E. J., Ubel, P. A., Kessler, J. B., Meyer, G., Muller, R. W., Navathe, A. S., Patel, P., Pearl, R., Rosenthal, M. B., Sacks, L., Sen, A. P., Sherman, P., & Volpp, K. G. (2016). Using Behavioral Economics to Design Physician Incentives That Deliver High-Value Care. *Annals of Internal Medicine*, *164*(2), 114. https://doi.org/10.7326/M15-1330

Fairbrother, G., Siegel, M., & Friedman, S. (2001). Impact of financial incentives on documented immunization rates in the inner city: Results of a randomized controlled trial. *Ambul Pediatr*, *1*(4), 206–212. https://doi.org/10.1367/1539-4409(2001)001<0206:IOFIOD>2.0.CO;2

Forbes, L. J., Marchand, C., Doran, T., & Peckham, S. (2017). The role of the Quality and Outcomes Framework in the care of long-term conditions: A systematic review. *British Journal of General Practice*, *67*(664), e775–e784. https://doi.org/10.3399/bjgp17X693077

Gemmill, M. (2007). Pay-for-Performance in the US: What lessons for Europe? *Eurohealth*, *13*(4), 21–3.

Gertler, P., Giovagnoli, P., & Martinez, S. (2014). *Rewarding Provider Performance to Enable a Healthy Start to Life. Evidence from Argentina's Plan Nacer*. World Bank.

Gillam, S. (2015). Financial incentive schemes in primary care. *Journal of Healthcare Leadership*, *7*, 75–80. https://doi.org/10.2147/JHL.S64365

Gillam, S. J., Siriwardena, A. N., & Steel, N. (2012). Pay-for-performance in the United Kingdom: Impact of the quality and outcomes framework: a systematic review. *Annals of Family Medicine*, *10*(5), 461–468. https://doi.org/10.1370/afm.1377

Gleeson, S., Kelleher, K., & Gardner, W. (2016). Evaluating a Pay-for-Performance Program for Medicaid Children in an Accountable Care Organization. *JAMA Pediatrics*, *170*(3), 259. https://doi.org/10.1001/jamapediatrics.2015.3809

Gupta, N., & Ayles, H. M. (2019). Effects of pay-for-performance for primary care physicians on diabetes outcomes in single-payer health systems: A systematic review. *The European Journal of Health Economics*, *20*(9), 1303–1315. https://doi.org/10.1007/s10198-019-01097-4

Himmelstein, D. U., Ariely, D., & Woolhandler, S. (2014). Pay-for-Performance: Toxic to Quality? Insights from Behavioral Economics. *International Journal of Health Services*, *44*(2), 203–214. https://doi.org/10.2190/HS.44.2.a

Houle, S. K. D., McAlister, F. A., Jackevicius, C. A., Chuck, A. W., & Tsuyuki, R. T. (2012). Does performance-based remuneration for individual health care practitioners affect patient care?: A systematic review. *Annals of Internal Medicine*, *157*(12), 889–899. https://doi.org/10.7326/0003-4819-157-12-201212180-00009

Hsieh, H.-M., Chiu, H.-C., Lin, Y.-T., & Shin, S.-J. (2017). A diabetes pay-for-performance program and the competing causes of death among cancer survivors with type 2 diabetes in Taiwan. *International Journal for Quality in Health Care*, *29*(4), 512–520. https://doi.org/10.1093/intqhc/mzx057

Hsieh, H.-M., He, J.-S., Shin, S.-J., Chiu, H.-C., & Lee, C. T.-C. (2017). A Diabetes Pay-for-Performance Program and Risks of Cancer Incidence and Death in Patients With Type 2 Diabetes in Taiwan. *Preventing Chronic Disease*, *14*, 170012. https://doi.org/10.5888/pcd14.170012

Institue of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. *Washington DC: National Academy Press*. Washington DC: National Academy Press.

Kandpal, E. (2016). *Completed Impact Evaluations and Emerging Lessons from Health Results Inovation Trust Fund Learning Portfolio*. World Bank.

Khim, K., Jayasuriya, R., & Annear, P. L. (2018). Administrative reform and pay-for-performance methods of primary health service delivery: A comparison of 3 health districts in Cambodia, 2006-2012. *The International Journal of Health Planning and Management*, *33*(2), e569–e585. https://doi.org/10.1002/hpm.2503

Kolozsvári, L. R., Orozco-Beltran, D., & Rurik, I. (2014). Do family physicians need more payment for working better? Financial incentives in primary care. *Atencion Primaria*, *46*(5), 261–266. https://doi.org/10.1016/j.aprim.2013.12.014

Kondo, K. K., Damberg, C. L., Mendelson, A., Motu'apuaka, M., Freeman, M., O'Neil, M., Relevo, R., Low, A., & Kansagara, D. (2016). Implementation Processes and Pay for Performance in Healthcare: A Systematic Review. *Journal of General Internal Medicine*, *31 Suppl 1*, 61–69. https://doi.org/10.1007/s11606-015-3567-0

Korda, H., & Eldridge, G. N. (2011). Payment incentives and integrated care delivery: Levers for health system reform and cost containment. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, *48*(4), 277–287. https://doi.org/10.5034/inquiryjrnl_48.04.01

Langdown, C., & Peckham, S. (2014). The use of financial incentives to help improve health outcomes: Is the quality and outcomes framework fit for purpose? A systematic review. *Journal of Public Health (Oxford, England)*, *36*(2), 251–258. https://doi.org/10.1093/pubmed/fdt077

Latham, L. P., & Marshall, E. G. (2015). Performance-based financial incentives for diabetes care: An effective strategy? *Canadian Journal of Diabetes*, *39*(1), 83–87. https://doi.org/10.1016/j.jcjd.2014.06.002

Lin, Y., Yin, S., Huang, J., & Du, L. (2016). Impact of pay for performance on behavior of primary care physicians and patient outcomes. *Journal of Evidence-Based Medicine*, *9*(1), 8–23. https://doi.org/10.1111/jebm.12185

Mabuchi, S., Sesan, T., & Bennett, S. C. (2018). Pathways to high and low performance: Factors differentiating primary care facilities under performance-based financing in Nigeria. *Health Policy and Planning*, *33*(1), 41–58. https://doi.org/10.1093/heapol/czx146

Markovitz, A. A., & Ryan, A. M. (2017). Pay-for-Performance: Disappointing Results or Masked Heterogeneity? *Medical Care Research and Review*, *74*(1), 3–78. https://doi.org/10.1177/1077558715619282

Mauro, M., Rotundo, G., & Giancotti, M. (2019). Effect of financial incentives on breast, cervical and colorectal cancer screening delivery rates: Results from a systematic literature review. *Health Policy (Amsterdam, Netherlands)*, *123*(12), 1210–1220. https://doi.org/10.1016/j.healthpol.2019.09.012

Mayo-Bruinsma, Liesha. (2013). Family-centred care delivery Comparing models of primary care service delivery in Ontario. *Can Fam Physician*, *59*(11), 1202-1210.

McPake, B., & Hanson, K. (2016). Managing the public–private mix to achieve universal health coverage. *The Lancet*, *388*(10044), 622–630. https://doi.org/10.1016/S0140-6736(16)00344-5

Mendelson, A., Kondo, K., Damberg, C., Low, A., Motúapuaka, M., Freeman, M., O'Neil, M., Relevo, R., & Kansagara, D. (2017). The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care. *Annals of Internal Medicine*, *166*(5), 341–353. https://doi.org/10.7326/M16-1881

Morgan, R., Ensor, T., & Waters, H. (2016). Performance of private sector health care: Implications for universal health coverage. *The Lancet*, *388*(10044), 606–612. https://doi.org/10.1016/S0140-6736(16)00343-3

Musgrove P. (2011). *Financial and Other Rewards for Good Performance or Results: A Guided Tour of Concepts and Terms and a Short Glossary*. Washington, D. C.: World Bank. https://www.rbfhealth.org/sites/rbf/files/RBFglossarylongrevised_0.pdf

NCDC. (2019). *Health Care. Statistical Yearbook. 2018 Georgia*. National Center for Disease Control and Public Health. https://www.ncdc.ge/Pages/User/News.aspx?ID=bec659c0-56a2-4190-9c0c-e47a63bcca4f

NHS. (2019). *Report of the Review of the Quality and Outcomes Framework in England*. https://www.england.nhs.uk/wp-content/uploads/2018/07/05-a-i-pb-04-07-2018-qof-report.pdf

Odutolu, O., Ihebuzor, N., Tilley-Gyado, R., Martufi, V., Ajuluchukwu, M., Olubajo, O., Banigbe, B., Fadeyibi, O., Abdullhai, R., & Muhammad, A. J. G. (2016). Putting Institutions at the Center of Primary Health Care Reforms: Experience from Implementation in Three States in Nigeria. *Health Systems and Reform*, *2*(4), 290–301. https://doi.org/10.1080/23288604.2016.1234863

Ogundeji, Y. K., Jackson, C., Sheldon, T., Olubajo, O., & Ihebuzor, N. (2016). Pay for performance in Nigeria: The influence of context and implementation on results. *Health Policy and Planning*, *31*(8), 955–963. https://doi.org/10.1093/heapol/czw016

Pan, Chien-Chou, Kung, Pei-Tseng, & Chiu, Li-Ting. (2017). Patients With Diabetes in Pay-For-Performance Programs Have Better Physician Continuity of Care and Survival. *Am J Manag Care*, *23*(2), e57-e66.

Patel, S. (2018). Structural, institutional and organizational factors associated with successful pay for performance programmes in improving quality of maternal and child health care in low and middle income countries: A systematic literature review. *Journal of Global Health*, *8*(2), 021001. https://doi.org/10.7189/jogh.08.021001

Paul, E., & Renmans, D. (2018). Performance-based financing in the heath sector in low- and middle-income countries: Is there anything whereof it may be said, see, this is new? *The*

International Journal of Health Planning and Management, *33*(1), 51–66. https://doi.org/10.1002/hpm.2409

Peckham, S., & Wallace, A. (2010). Pay for performance schemes in primary care: What have we learnt? *Quality in Primary Care*, *18*(2), 111–116.

Petrosyan, V., Melkom Melkomian, D., Zoidze, A., & Shroff, Z. C. (2017). National Scale-Up of Results-Based Financing in Primary Health Care: The Case of Armenia. *Health Systems and Reform*, *3*(2), 117–128. https://doi.org/10.1080/23288604.2017.1291394

Pullicino, G., Sciortino, P., Calleja, N., Schafer, W., Boerma, W., & Groenewegen, P. (2015). Comparison of patients' experiences in public and private primary care clinics in Malta. *The European Journal of Public Health*, *25*(3), 399–401. https://doi.org/10.1093/eurpub/cku188

Renmans, 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 Policy and Planning*, *31*(9), 1297–1309. https://doi.org/10.1093/heapol/czw045

Roland, M., & Guthrie, B. (2016). Quality and Outcomes Framework: What have we learnt?: *BMJ*, i4060. https://doi.org/10.1136/bmj.i4060

Rubinstein, A., Rubinstein, F., Botargues, M., Barani, M., & Kopitowski, K. (2009). A multimodal strategy based on pay-per-performance to improve quality of care of family practitioners in Argentina. *The Journal of Ambulatory Care Management*, *32*(2), 103–114. https://doi.org/10.1097/JAC.0b013e31819940f7

Saddi, F. C., & Peckham, S. (2018). Brazilian Payment for Performance (PMAQ) Seen From a Global Health and Public Policy Perspective: What Does It Mean for Research and Policy? *The Journal of Ambulatory Care Management*, *41*(1), 25–33. https://doi.org/10.1097/JAC.0000000000000220

Scott, A., Liu, M., & Yong, J. (2016). Financial Incentives to Encourage Value-Based Health Care: *Medical Care Research and Review*. https://doi.org/10.1177/1077558716676594

Scott, A., Sivey, P., Ait Ouakrim, D., Willenberg, L., Naccarella, L., Furler, J., & Young, D. (2011). The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD008451.pub2

So, J. P. P., & Wright, J. G. (2012). The use of three strategies to improve quality of care at a national level. *Clinical Orthopaedics and Related Research*, *470*(4), 1006–1016. https://doi.org/10.1007/s11999-011-2083-8

Soranz, D., & Pisco, L. A. C. (2017). Primary Health Care Reform in the cities of Lisbon and Rio de Janeiro: Context, strategies, results, learning and challenges. *Ciencia & Saude Coletiva*, *22*(3), 679–686. https://doi.org/10.1590/1413-81232017223.33722016

Sung, N. J., Suh, S.-Y., Lee, D. W., Ahn, H.-Y., Choi, Y.-J., Lee, J. H., & for the Korean Primary Care Research Group. (2010). Patient's assessment of primary care of medical institutions in South Korea by structural type. *International Journal for Quality in Health Care*, *22*(6), 493–499. https://doi.org/10.1093/intqhc/mzq053

Tao, W., Agerholm, J., & Burström, B. (2016). The impact of reimbursement systems on equity in access and quality of primary care: A systematic literature review. *BMC Health Services Research*, *16*(1), 542. https://doi.org/10.1186/s12913-016-1805-8

Van Herck, P., De Smedt, D., Annemans, L., Remmen, R., Rosenthal, M. B., & Sermeus, W. (2010). Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research*, *10*, 247. https://doi.org/10.1186/1472-6963-10-247

Vlaanderen, F. P., Tanke, M. A., Bloem, B. R., Faber, M. J., Eijkenaar, F., Schut, F. T., & Jeurissen, P. P. T. (2019). Design and effects of outcome-based payment models in healthcare: A systematic review. *The European Journal of Health Economics*, *20*(2), 217–232. https://doi.org/10.1007/s10198-018-0989-8

Wadge, H., Roy, R., Sripathy, A., Fontana, G., Marti, J., & Darzi, A. (2017). How to harness the private sector for universal health coverage. *The Lancet*, *390*(10090), e19–e20. https://doi.org/10.1016/S0140-6736(17)31718-X

Wekesah, F. M., Mbada, C. E., Muula, A. S., Kabiru, C. W., Muthuri, S. K., & Izugbara, C. O. (2016). Effective non-drug interventions for improving outcomes and quality of maternal health care in sub-Saharan Africa: A systematic review. *Systematic Reviews*, *5*(1), 137. https://doi.org/10.1186/s13643-016-0305-6

WHO. (2010). *Strengthening the capacity of governments to constructively engage the private sector in providing essential health-care services* (Provisional Agenda Item A63/25). World Health Organization. https://apps.who.int/gb/ebwha/pdf_files/WHA63/A63_25-en.pdf

WHO. (2015). *Building primary care in a changing Europe*. European Observatory on Health Systems and Policies. https://www.euro.who.int/__data/assets/pdf_file/0011/277940/Building-primary-care-changing-Europe-case-studies.pdf

WHO. (2018a). *Handbook for national quality policy and strategy – A practical approach for developing policy and strategy to improve quality of care*. Geneva: World Health Organization.

WHO. (2018b). *Private sector engagement for advancing universal health coverage. Regional Committee for the Eastern Mediterranean Sixty-fifth session*. WHO. https://applications.emro.who.int/docs/RC_Technical_Papers_2018_8_20546_EN.pdf

WHO. (2018c). *The private sector, universal health coverage and primary health care*. World Health Organization. file:///Users/macbook/Downloads/WHO-HIS-SDS-2018.53-eng.pdf

WHO Regional Office for Europe. (2018). *Quality of primary health care in Georgia*. Division of Health Systems and Public Health. https://www.euro.who.int/__data/assets/pdf_file/0003/373737/geo-qocphc-eng.pdf

WHO Regional Office for Europe. (2019). *Can people afford to pay for health care?*

WHO Regional Office for Europe. (2020). *Health for All Database*. WHO. https://gateway.euro.who.int/en/datasets/european-health-for-all-database/

Witter, S., Fretheim, A., Kessy, F. L., & Lindahl, A. K. (2012). Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD007899.pub2

Wiysonge, C. S., Paulsen, E., Lewin, S., Ciapponi, A., Herrera, C. A., Opiyo, N., Pantoja, T., Rada, G., & Oxman, A. D. (2017). Financial arrangements for health systems in low-income

countries: An overview of systematic reviews. *The Cochrane Database of Systematic Reviews*, *9*, CD011084. https://doi.org/10.1002/14651858.CD011084.pub2

Wong, S. Y., Kung, K., Griffiths, S. M., Carthy, T., Wong, M. C., Lo, S. V., Chung, V. C., Goggins, W. B., & Starfield, B. (2010). Comparison of primary care experiences among adults in general outpatient clinics and private general practice clinics in Hong Kong. *BMC Public Health*, *10*(1), 397. https://doi.org/10.1186/1471-2458-10-397

Yu, H.-C., Tsai, W.-C., & Kung, P.-T. (2014). Does the pay-for-performance programme reduce the emergency department visits for hypoglycaemia in type 2 diabetic patients? *Health Policy and Planning*, *29*(6), 732–741. https://doi.org/10.1093/heapol/czt056

Yuan, B., He, L., Meng, Q., & Jia, L. (2017). Payment methods for outpatient care facilities. *The Cochrane Database of Systematic Reviews*, *3*, CD011153. https://doi.org/10.1002/14651858.CD011153.pub2

Annexes

Annex 1. General characteristics of the included studies

Reference	Last Search	# of studies	AMSTAR Rating	Income Level	Countries	Public/ Private	PHC areas
Allen et al. (2014)	NA	Not Given	-	High	GBR, USA, DEU	Public	General
Alshamsan et al. (2010)	2009	Unclear		High	GBR	Unclear	NCD
Boeckxstaens et al. (2011)	2009	27	-	High	GBR	Unclear	General
Carter et al. (2017)	2015	14	-	High	CAN	Unclear	General
Cattel et al. (2019)	2017	111		High	USA, DEU, ESP, NLD	Both	General
Christianson et al. (2009)	2007	7	3/9	High	AUS, FRA, ISR, NOR, ESP, SWE, GBR, USA,global	Unclear	General
Conrad et al. (2015).	NA	NA	-	High	USA	private	General
Das et al. (2016)	2014	8	7/10	Low Lower-middle	BDI, COD, EGY, PHL, RWA	Both	MCH
Eijkenaar et al. (2013)	2011	22	-	High Upper-Middle Low	USA, GBR, ARG, other	Unclear	General
Forbes et al. (2017).	2016	8	-	High	GBR	Unclear	General
Gertler et al. (2014)	NA	NA	-	Upper-middle	ARG	Unclear	MCH
Gillam et al. (2012)	2011	95	-	High	GBR	Unclear	General
Gillam et al. (2015)	2014	7	-	High	GBR	Unclear	General
Gupta et al. (2019)	2018	10	-	High	AUS, CAN, ITA, SWE, TWN, GBR	Unclear	Diabetes
Houle et al. (2012)	2012	30	9/10	High	CAN, DEU, GBR, USA	Unclear	NCD Diabetes
Kandpal (2016)	NA	7	-	Low Upper-middle	ARG, CMR, COD, AFG, RWA, ZMB, ZWE	Unclear	MCH
Khim et al. (2018)	NA	NA	-	Lower-middle	KHM	Public	General
Kolozsvári et al (2014).	NA	57	-	Upper-middle High	European Union	Unclear	General
Kondo et al. (2016)	2014	41	6/10	High	AUS, CAN FRA,ITA, KOR, NLD, TWN, GBR, USA	Unclear	General
Korda et al. (2011)	2011		-	High	USA	Both	General

Langdown et al. (2014)	2012	11	6/10	High	GBR	Unclear	General
Latham et al. (2015)	NA	NA	-	High	GBR, AUS, TWN, CAN	Unclear	NCD (Diabetes)
Lin et al. (2016)	2013	44	6/10	Upper-middle High	ARG, FRA, IRL, NDL, TWN, GBR, USA	Unclear	General
Mabuchi et al. (2018)	NA	NA	-	Lower-middle	NGA	Public	MCH
Markovitz and Ryan (2019)	2015	58	-	High	USA, UK, CAN	Both	General
Mauro et al. (2019)	2018	18	-	High	USA, TWN, NLD; AUS, CAN, FRA	Unclear	Screening
Mendelson et al. (2017)	2016	69	8/10	High	AUS, CAN FRA,ITA, KOR, NLD, TWN, UK (GBR), USA	Unclear	General
Odutolu et al. (2016)	NA	NA	-	Lower-middle	NGA	Both	MCH
Ogundeji et al. (2016)	NA	NA	-	Lower-middle	NGA	Unclear	MCH
Patel et al. (2018)	2017	13	5/9	Low Lower-middle Upper-middle	AFG, ARG, BGD, BDI, KHM, CMR, COD,EGY, PHL, RWA, ZMB, ZWE	Both	MCH
Paul, E., & Renmans, D. (2018)	NA	NA	-	Low Lower-middle	BEN, BDI, CMR, KHM, COD, RWA,TZA, UGA	Unclear	General
Peckham et al. (2010)	NA	2	-	High	GBR	Unclear	General
Petrosyan et al. (2017)	NA	NA	-	Upper-middle	ARM	Both	MCH NCD
Renmans et al. (2016)	2016	35	-	Low Lower-middle	BEN, BDI, CMR, KHM, COD, RWA,TZA, UGA	Both	General
Saddi, F. C., & Peckham, S. (2018).	NA	NA	-	Upper-middle	BRA	Unclear	General
Scott et al. (2011) [ref]	2009	6	10/10	High	USA, GBR, DEU	Both	General
Scott et al. (2018)	2015	80	-	Low Upper-middle High	USA, GBR, CHN, CAN, ITA, AUS, FRA, PHL, RWA	Unclear	General

So & Wright (2012)	2010	20	1/10	High	CAN	Unclear	General
Soranz & Pisco (2017)	NA	NA	-	Upper-middle High	BRA, PRT	Unclear	General
Soranz et al. (2017)	NA	NA	-	Upper-middle	BRA	Unclear	General
Tao et al. (2016)	2013	27	6/9	High	CAN, GBR, USA	Unclear	NCD, Preventive care
Van Herck et al. (2010)	2009	128	7/10	Upper-middle High	USA, GBR, AUS, DEU, ARG, ITA	Unclear	General
Wekesah et al. (2016)	2015	73	6/10	Low Lower-middle	AGO, BFA, BEN, BDI, CIV, ETH, GHA, KEN, MWI, MLI,MOZ, NGA, RWA, SEN, SOM, ZAF, TZA, UGA, ZMB	Unclear	MCH
Wiysonge et al. (2017).	2016	15	-	High	USA, CAN, AUS, ARE(UAE), TWN, Western Europe	Unclear	General
Yuan et al. (2017)	2016	21	10/10	High Upper-Middle	AFG, BDI, COD, CHN, RWA, TZA, GBR, USA	Unclear	General

Annex 2. Key findings from included studies

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
Allen et al., 2014	A systematic review of the evidence base for the effects of financial incentives on the quality of health care provided by primary care physicians found seven relevant studies. Six of these seven studies showed modest and inconsistent positive effects on quality of care for some primary outcome measures, and one found no effect whatsoever. The systematic review noted that study design led to substantial risk of bias for the majority of these studies (particularly self-selection into schemes by physicians). UK - Quality and Outcome Frameworks (QOF). Performance increased in incentivized areas of quality following the introduction of the QOF, but this appears to have been largely due to a step-increase in quality immediately after the QOF introduction, as performance increases were not found in later years of the QOF. Data from the first year of the QOF showed that practices in less deprived areas delivered a higher quality of care. Practices with lower QOF performance have a greater chance to improve on quality than practices with already high performance. By the third year of the QOF, the performance gap between the most and least deprived had disappeared.	Analyzing a period of time from 2000 to 2006, rates of recording were found to have increased for all the various groups of patients used, suggesting positive spillover effects on quality. The effect sizes did differ, however, and was largest for incentivized indicators for patients with targeted diseases. Increases in recording rates for risk factors that were not incentivized	More positive effects were also found for schemes that adopted absolute and not relative targets, potentially suggesting that "room for improvement" and benchmarking should be an important consideration for the design of P4P schemes. a higher degree of provider engagement and the collaborative design of schemes was found to correlate with better results.
Alshamsan et al., 2010	QOFs has been associated with improvements in the management of diabetes in primary care (for the quality indicators included in the QOF, particularly, in the process aspect of quality). However, these improvements do not appear to have been uniform across all patients' groups. Not all groups appear to have benefited equally from this policy, including women and people from certain ethnic minority groups, and many people with diabetes are still not meeting established treatment targets.	Not all groups appear to have benefited equally from this policy, including women and people from certain ethnic minority groups, and many people with diabetes are still not meeting established treatment targets. The impact of exception reporting on diabetes management is not clear, however findings from 312 primary care practices in Scotland shows that older patients with stroke and patients with co-	Longer term evaluation and monitoring of QOF is needed to gain a more complete assessment of its impacts. An increase in the threshold for achievement of existing targets may have been more appropriate, as it may lead to greater overall net benefits to patients.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
		morbidity were more likely to be exception reported.	Impact of exception reporting on diabetes management requires further and ongoing evaluation. Improvements in providing patients with continued, personal and coordinated care, elements that may need to be reflected in QOF in the future.
Boeckxstaens et al., 2011	Equal access to care for all patients is an essential prerequisite to equal health care. However, none of the selected publications compares the profile of users versus non-users of care, making it impossible to assess the impact of QOF on access to care. This is probably influenced by the context of the UK's health system where access to primary care services is almost universal because only a very small minority of patients is not registered. Despite the universality of the system, some specific population groups still find it difficult to register with a GP. For instance, homeless people often do not know that they have to register or are scared off by the complexity of the registration procedure. Financially-driven quality improvement systems using purely biomedical indicators may lead to the loss of important aspects of health care quality such as trust and high-quality empathic communication. It has been suggested that QOF might have changed the nature of the practitioner-patient consultation with, for instance, a decline in personal/relational continuity of care between doctors and patients. We can state that the introduction of QOF has benefited the aged and males. Regarding ethnicity and deprivation, it is almost impossible to draw general conclusions. At the level of total QOF score, ethnicity appeared to be of no influence. For deprivation, small but significant residual differences were observed after the introduction of QOF favoring less deprived groups. However, after correcting for practice characteristics, the influence of deprivation was no longer observed, indicating that the small but existing differences between socio-economic groups are mainly due to differences at the practice level. Practices in affluent areas are possibly better trained and better surrounded.	QOF type drivers may influence the nature of the doctor patient interaction shifting the focus to disease-oriented care especially when mainly disease oriented economic incentives are included in the care process, hereby possibly counteracting patient centered and comprehensive care. It has been suggested that QOF might have changed the nature of the practitioner-patient consultation with, for instance, a decline in personal/relational continuity of care between doctors and patients.	

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	According to the inverse equity hypothesis formulated in 2000 affluent groups in society preferentially benefit from new interventions, leading to an initial increase in inequality. Deprived groups only begin to benefit once affluent groups have extracted maximum benefit. Health inequalities ultimately diminish because deprived groups start with a lower baseline level of health and health care uptake and have higher potential gains.		
Carter et al., 2016	Small and sometimes non-significant improvements in processes of care as measured by the delivery of screening and prevention services and chronic disease management.		Incentive payments should be carefully designed with the overarching payment model in mind.
Cattel et al., 2019	The results from AQC initiatives: Significant, positive effect on pediatric preventive care quality measures tied to P4P (+1.8% for Children with special needs (CSHCN) and +1.2% for non-CSHCN; $p < .001$). No significant changes for measures not tied to P4P. Significant improvements of some measures (e.g., 3.1% for low-density lipoprotein cholesterol testing [$p < .001$] and 2.5% for cardiovascular disease [$p < .001$]), but no differential change for others. After 1 year: Improved quality for chronic conditions in adults ($p < .001$) and pediatric care ($p = .001$) but not for adult preventive care. After 2 years: Improvements in measures for chronic care management (+3.7%; $p < .001$), adult preventive care (+0.3%; $p = .008$), and pediatric care (+0.3%; $p < .001$). Over 4 years period: Measures of chronic disease management increased by 3.9%, and unadjusted performance in adult preventive care and pediatric care increased by 2.7% and 2.4% (p values are unavailable) compared to the healthcare Effectiveness Data and Information Set (HEDIS) national average. The five outcome measures for patients with diabetes, patients with coronary artery disease, and patients with hypertension improved compared to the national and regional HEDIS scores (size of the effect and p values unavailable). Process measures improved +1.2% per year more among individuals living in areas with lower versus higher socioeconomic status ($p < .001$). No significant differences in outcome measures.		Consistent with the recommendation by Roland and Campbell (2014) that P4P needs to be combined with other improvement strategies to produce sustained improvements, implementing VBP while disregarding other relevant factors is unlikely to materially affect value.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Significant improvements in 8 of the 14 HEDIS measures for preventive care, chronic care, and acute care primary care services for the group of Nationwide Children Hospital physicians compared to incentivized physicians ("traditional" P4P). ORs favored the intervention group mainly in the immunization measures (range of OR of 0.34 with CI of [0.31, 0.37] for hepatitis vaccine to 0.86 with CI of [0.78, 0.95] for meningococcal vaccine).		
Christianson et al., 2009	The findings from studies on the effect of payer initiatives that reward providers for quality improvements or the attainment of quality benchmarks are mixed. Relatively few significant impacts are reported, and it is often the case that payer programs include quality improvement components in addition to incentive payments, making it difficult to assess the independent effect of the financial incentives. Very little research has been done on the impact of direct payments to hospitals to improve quality. The published research to date in this area is too limited to draw conclusions with confidence. Though relatively more attention has been paid to preventive services, there is limited evidence that targeted interventions employing financial incentives to improve the delivery of preventive services are effective. The few studies in this area with strong research designs find small, if any, effects of payments to providers that are intended to improve quality. The accumulated body of research described in this chapter is not yet sufficient to assess the relative significance of identified barriers to the effective design and implementation of P4P initiatives.		There are large P4P programs underway in the US and the UK with more evaluations likely to appear in the peer-reviewed literature in the near future. Because of the variation in the way these programs have been designed and implemented, synthesizing their findings to provide useful guidance for decision-makers will be challenging. It will be especially important to have comprehensive reporting of results in future studies (not limiting results to a subset of quality measures rewarded by payers), accompanied by complete descriptions of study context and possible confounding factors. In the meantime, policy-makers can support, and learn from, process evaluations of ongoing P4P efforts with particular attention to accurate documentation of costs as well as continued tracking of outcomes.
Conrad et al. 2015	N/A		Facilitating factors: Several factors are facilitating POP: the history of collaboration and innovation in Oregon and particularly in Salem; leadership of Physicians Choice Foundation, Performance Health Technology, and WVP Health

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			Authority; the legislation's support of CCOs and its mandate to pay providers based on quality and reduction in cost growth; the contributions of local physicians and other providers to the program; and the substantial financial investments in developing the POP software and Performance Health Technology's systems for collecting claims data. Barriers: Competing priorities created with the Oregon legislature's and the federal government's authorization of CCOs, and other local health care reforms; turnover of a POP leader who was instrumental in building and maintaining cohesive relations between the independent practice association and the local medical society; inability of all providers to submit claims electronically; and POP's complexity and consequent difficulty in explaining the program to medical practices.
Das et. al 2016	P4P improved physicians' knowledge to manage under-five diarrhea and pneumonia (coefficient 1.6; $p < 0.001$). There was a small improvement in patient reported health measure for under-five (coefficient 7.37; $p = 0.001$).		
Forbes et al. 2017	QOF may be associated with a modest reduction in emergency admission rates in long-term conditions, a modest increase in consultation rates in severe mental illness, and modest improvements in certain limited aspects of the care of diabetes.		In the context of a demoralized primary care workforce, it is important also to consider ways other than financial incentives to motivate primary care teams to deliver high-quality care.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	No clear evidence that these changes have led to any effect on mortality. Because of the design of the studies, it is not possible to be sure that any of the positive effects seen are causally related to the QOF. Trend of increasing emergency hospital admission rates (which increased overall by 34% between 2004 and 2010) was modestly lower for conditions incentivized in the QOF compared with conditions that were not incentivised in the QOF, by 3% in the first year rising to 8% in 2010. The difference was mainly driven by relative reductions in emergency admission rates for coronary heart disease. No evidence to suggest that the QOF influences, positively or negatively, other aspects of care, such as integration or coordination of care, holistic or personalized care, or self-care, nor any evidence of its effects on patients' quality of life, experience, or satisfaction.		
Gertler et al., 2014	The results show a significant increase in the number of prenatal care visits and the quality of prenatal care measured by an increase in the share of mothers who receive the tetanus toxoid vaccine and a reduction in the number of births delivered by caesarian. Improved prenatal care appears to be translated into improved birth outcomes as we observe a significant increase in average birth weight and a reduction of the share of low birth weight babies. Plan Nacer (Brazilian P4P scheme) can reduce neonatal mortality both by preventing low birth-weight and by increasing survivorship of risky low-birth-weight babies.	There does seem to be some evidence of negative spillovers in birthweight or in quality of prenatal care (i.e tetanus and cesarean section). However, we do find negative and statistically significant spillover effect for the number of prenatal care visits.	
Gillam et al., 2011	The QOF has helped consolidate evidence-based methods for improving care by increasing the use of computers, decision support, clinician prompts, patient reminders, and recalls. t has resulted in better recorded care, enhanced processes, and improved intermediate outcomes for most conditions, notably diabetes. These improvements decreased after the first year of the QOF, however, and subsequent increases have followed secular trends.		Policy makers draw must, of course, take account of the different historical and organizational contexts in which their health system operates. Some indicators for which performance has reached a ceiling may need to be retired, although performance may not be maintained,

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Performance improvements for those conditions that were not included in the QOF were significantly lower than for incentivized indicators, and these differences increased over time. Efficiency. There is limited evidence that increasing the quality of ambulatory care may reduce admission rates and hence costs for some conditions. Equity. inequalities in processes of care comparing the most and least deprived areas have narrowed. The QOF has encouraged greater consistency of care irrespective of deprivation, but the practitioners' option to exclude (exception report) hard-to-reach patients from the population used to determine payment may limit its impact on health inequalities. Conflicting findings, but some consistent themes have emerged. It has been associated with an increased rate of improvement of quality of care during the first year of implementation, returning to preintervention rates of improvement in subsequent years. There have been modest reductions in mortality and hospital admissions in some areas, and where they have been assessed, these modest improvements appear cost-effective. The QOF has led to narrowing of differences in performance in deprived areas compared with areas not deprived. It has strengthened team working. The effect of the QOF in unincentivized areas has been disappointing. The costs of administering the scheme are substantial, and some staff are concerned that primary care has become more biomedical in focus and less patient centered. The QOF has strengthened team working and promoted a diversity of new roles, especially for nurses. Indeed, the QOF may have diminished the workload of general practitioners, enabled them to concentrate on more complex care. The QOF has been described as scientific bureaucratic medicine, where indicators and guidelines are perceived as threatening professionalism in various ways. The fear expressed by some that adherence to single disease-based guidelines might override respect for patient autonomy, lead clinicians to ignore comorbidities, promote a mechanistic approach to chronic disease management, or reduce clinical practice to a series of dichotomized decisions		and new indicators should be introduced after piloting. Consideration should be given to improving different dimensions of quality, including user experience and equity. Costs should be monitored and balanced against benefits. Wherever possible, schemes should be designed in collaboration with health service researchers to evaluate the benefits of minor differences in system design. Payment for performance is still an imperfect approach to improving primary care, and should be considered as only one option alongside alternative quality improvement methods.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	at the expense of personal aspects of care, has not been borne out by the research to date.		
Gillam et al., 2015	Evidence for the effectiveness of financial incentives is inconsistent. A recent Cochrane review of seven studies in primary care found that financial incentives were effective for some outcomes in some settings but concluded that there was "insufficient evidence to support or not support the use of financial incentives to improve the quality of primary health care". Similarly, previous systematic reviews have concluded that P4P contracts do affect physician behavior and increase the range of primary care services provided but that their impact is often limited. The modest effects of financial incentives tend to be measured in terms of improvements in the processes of chronic disease management. Review of the QOF (UK) found that quality of care for incentivized conditions during the 1st year of the framework improved at a faster rate than the preintervention trend but subsequently returned to prior rates of improvement. One study in the UK found that an externally imposed system of incentives did not appear to damage the internal motivation of GPs. The authors attributed this to the fact that the indicators within the QOF aligned with what GPs themselves considered good clinical care objectives. Another study found that GPs felt that, while professional autonomy had decreased and workload increased, they were paid more and their job satisfaction levels had increased under the QOF. Nurses also report that their specialist skills have been enhanced. There is some evidence that P4P can reduce health inequalities resulting from socioeconomic disadvantage: the gap in median achievement comparing practices from the most deprived and least deprived quintiles in the UK narrowed from 4.0% to 0.8% between 2004 and 2007. On the other hand, achievements incentivized under the QOF have not reduced premature death in the population and inequalities have persisted. There is some evidence that P4P can reduce health inequalities resulting from socioeconomic disadvantage: the gap in median achievement comparing practices from the most deprived and least deprived quintiles in the UK		The actual effect of financial incentives appears to depend on factors such as the age and sex of physicians, previous experience of financial incentives, the uptake of continuing professional education, the payment method, the type and severity of the conditions targeted through incentives, the volume of activity, and the location and type of organization. Research conducted in the USA found that the size and structure of incentives do seem to be important in promoting effective physician activity. Incentives have to be large enough to influence behavior and designed in such a way that they cannot be "gamed". The size of incentive may be less important in improving care processes than giving public recognition for scoring well on quality measures.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	narrowed from 4.0% to 0.8% between 2004 and 2007. On the other hand, achievements incentivized under the QOF have not reduced premature death in the population and inequalities have persisted.		
Gupta et al., 2019	P4P schemes aimed at optimizing the delivery of primary care medical services among patients with diabetes, when tied to physician performance metrics, can have important effects in limiting disease progression and severity for multiple morbidities over the long term. This may be attributed to enhanced clinical practices and counselling for patient self-management. High powered incentives. In Taiwan in the study of 396,838 patients found that P4P increased physician continuity of care among patients with diabetes, and in turn was associated with lower risk of mortality, other author reported significantly lower risks of cancer-specific mortality in newly diagnosed cancer patients. Another study found that the national diabetes P4P scheme reduced the 5-year risk of all-cause mortality and diabetes-related mortality among patients having survived cancer. (propensity score- matching used to control for selection bias in both studies) In UK, also a context of high-powered incentives, did not show a reduction in premature mortality rates associated with P4P in primary care, the authors acknowledged the limitation of their spatial analysis in terms of a lack of accounting for the quality of local secondary care services. In Sweden, the introduction of high-powered incentives in one county was associated with significantly greater target achievement for patients' hemoglobin A1c, blood pressure, and LDL cholesterol compared to a reference county. In low powered: Little evidence of improved primary care access or continuity, and mixed associations with the risk of diabetes-related hospitalization (Italy, Canada). Limited uptake of a low-powered P4P scheme in Denmark was attributed to the weak incentive structure to effectively promote behavior change among physicians		Cautiously treat P4P programs to incentivize NCD care while implementing and sustaining universal health coverage. The first issue should be how to measure and monitor from the onset quality of care and patient outcomes against specific targets and goals. These performance metrics need to be transparent, valid, and consensus-driven but not overly cumbersome. The second issue should be the size of extra payments. Studies to date, albeit limited, show that modest physician incentives yield limited to negligible health gains in patients. However, evidence on effects of larger payments for disease-specific P4P remains inconclusive, notably in terms of unintended diversion of resources from other public health concerns.
Houle et al., 2012	A recent Cochrane review on the effect of financial incentives for primary care physicians included 7 studies and concluded that "there is insufficient evidence to support or not support the use of financial incentives to improve the quality of primary health care."	Patient perception of continuity of care declined after P4P implementation in the UK (where rapid access to care rather than continuity with the same physician was incentivized), which raises concerns	Although P4P seems to be useful in business settings and may serve as a means to signal which elements of care are valued within a participating health care organization, the current

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Studies on the Effect of P4P on Preventive Care or Screening (n 10) Randomized, Controlled Trials: Statistically significant improvements were found in immunization rates with P4P versus fee-for-service (FFS), the absolute effect sizes in both trials were small. In contrast, Grady and coworkers found no improvement in mammography referral or performance rates for women seeing P4P physicians. Although uncontrolled before–after studies suggested that P4P improves adherence to quality-of-care indicators for chronic illnesses (such as the ordering of laboratory tests in patients with diabetes, measurement and achievement of target blood pressure, adherence to prescribing guidelines for patients with heart failure), higher-quality studies with contemporaneous control groups or analyses that considered secular trends failed to confirm these benefits. Vamos and colleagues reported statistically significant improvements in achievement of blood pressure and total cholesterol targets in individuals with diabetes but reduced achievement of glycosylated hemoglobin targets in the year after P4P introduction versus trends before P4P.	given the known negative effect of care fragmentation on patient satisfaction and outcomes. In addition, the potential negative effect of P4P remuneration schemes on the job satisfaction of clinicians should be considered; at least 1 study has documented reduced satisfaction among physicians in a P4P program as a result of increased administrative responsibilities. The potential to change health care provider focus from quality of care to quality of record-keeping, and the potential for gaming through such methods as exception reporting (that is, exclusion of patients from denominators to improve percentage target achievement), falsifying of data, and measurement fixation has also been raised. Exception reporting was not widespread in the UK after implementation of their primary care P4P program (median, 6%), they did find that the rate of exception reporting was the strongest predictor of target achievement and that 1% of all practices excluded more than 15% of their patients from target calculation denominators. Furthermore, as P4P schemas emphasize selected target indicators, it is unknown whether P4P-remunerated clinicians may preferentially avoid caring for patients with complex multisystem disease in whom hitting a target for one of their	evidence for P4P targeting individual practitioners is insufficient to recommend wholesale adoption in health care systems at this time. Performance incentives arose from the principal agent theory in economics and have been shown in some instances to affect behavior (for example, annual bonuses tied to sales or cost-savings in the business sector, although the benefits tend to be specific to the remuneration scheme and the setting. The optimal P4P scheme for health care remains an unresolved question, although our review provides some insights. For example, the targets chosen for incentive payments should not be too narrow because even the studies with positive results have shown improvement only for incentivized targets, with no spillover effect for non-incentivized targets. In addition, careful consideration must be taken in deciding whether to base incentives on process or outcome measures because process measures are more easily modifiable by the professional and may therefore be more achievable, but they may not always translate into improvements in clinical outcomes. The size of the financial incentive relative to the effort required is another consideration, although we

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
		conditions would be more difficult than in patients with single-system disease	found evidence that even small incentives (worth less than 5% of annual income) seemed sufficient to modify practice in some settings and that much larger incentives were ineffective in other settings. Furthermore, programs must consider whether to reward absolute or relative changes in performance and whether comparisons are made against one's peers or an individual's past performance. The potential to change health care provider focus from quality of care to quality of record-keeping, and the potential for gaming through such methods as exception reporting (that is, exclusion of patients from denominators to improve percentage target achievement), falsifying of data, and measurement fixation has also been raised.
Kandpal 2016	The first impact evaluation of the Plan Nacer (see Gertler et al.,2014). The second impact evaluation focuses on the Misiones province and uses a randomized field experiment to provide key evidence on the sustainability of effects of RBF incentives. The evaluation estimates the effect of large (three-fold) but temporary increase in financial incentives for health care providers on the initiation of prenatal care in the first trimester of pregnancy. Results show the rate of early initiation of prenatal care was 34% higher in the treatment group than in the comparison group while the incentives were being paid, and that this effect persisted 12 months after the incentives ended. Results, however, also suggest that the quality of care may have remained a constraint to improving health outcomes as the increase in early initiation of prenatal care did not have any effect on birth outcomes. Nonetheless, the study also finds that large-but-temporary incentives can be more cost-		

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	effective at motivating provider performance and changing clinical practice than permanent incentives when providers face fixed costs to changing clinical practice routines		
Khim et al., 2018	There was observed a significant improvement in service delivery in P4P districts for 3 important primary care indicators - monthly number of new cases visited outpatient facilities, proportion of 2nd ANC visits attended, children under 1 immunized - following Special Operating Agencies (SOA) introduction, with some notable exceptions. Although, the results must be interpreted as being associated with a broader public administration reform in the health sector focusing particularly on the management of provincial and district health service delivery and not simply as a P4P or PBF intervention. Nonetheless, salary supplements and/or performance-based incentives played a significant role in the Special Operating Agencies (SOA)-contracting outcomes. We conclude that much of the improvement in service delivery outcomes was not the result of the P4P contracting intervention alone but was also influenced by context and circumstances nationally and in the 3 study districts. In addition to routine monthly fluctuations, interruption in critical management functions (local and provincial) may possibly be the cause of an otherwise unusual drop in service delivery in many cases during the inaugural month of SOA implementation in 2009 (outpatient consultations, immunization, and ANC in Chamkaleu and ANC and newborn deliveries in Cheungprey).		Best practice in contracting requires that monitoring be implemented by an independent agent, but under the SOA reform, all 3 monitoring teams (central MOH, PHD, and SOA) were internal to the MOH and there is evidence of partisan behavior and inherent difficulties in applying penalties for poor performance within the bureaucratic system and a lack of autonomy of district officials. Contextual factors, such as public sector governance and regulation, are integral to success of the reform. The SOA model of internal contracting could be further strengthened by improved monitoring, linking incentive payment to performance, improving the governance arrangements, and providing a clearer purchaser-provider split under the MOH.
Kolozsvári et al., 2014			Ten countries were found and listed where primary care quality indicators are used and combined with financial incentives. The number of quality indicators varies from 1 to 134, the highest in the UK, the lowest in Italy. In 8 countries QI can influence the finances/salary of family physicians with a bonus of 1-25% of their total income. Besides the nation-wide

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			systems, there were local experiments and different regional systems mentioned in the Netherlands and in Italy, respectively. Are quality indicators needed for a better primary care? What indicators? The quality of the incentivized fields might improve; the non-incentivized activities could be neglected. How many indicators? Implementation of too many indicators can lead to increased bureaucracy and box ticking instead of spending time with patients. In the UK (134 indicators), there are opinions, that the indicator system should be simplified to decrease the GPs administrative workload. P4P schemes have become increasingly popular innovations in primary care and have generated questions about their effect on improving quality of care, although in some countries were not linked to QIs. There is no sufficient evidence that contradicts or supports the quality improvement effect of financial incentives. The effectiveness of P4P is inconclusive, though some reviews reported significant effects. A participatory P4P program might stimulate quality improvement in clinical care and improve patient

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			experiences with GP's functioning and the organization of care. P4P schemes need to take more account of broader definitions of quality, as whilst they can have a positive impact on incentivized clinical processes, it is not clear that this translates into improving the experience and outcome of care. Too low incentives are not likely to be effective, too high incentives can cause unintended consequences (e.g. data manipulation, "gaming"/cheating).
Kondo et al., 2016	The heterogeneity across health systems and organizations and the challenges related to the evaluation of complex interventions such as P4P preclude from drawing firm conclusions. Measures linked to quality and patient care were positively related to improvements in quality and greater provider confidence in the ability to provide quality care, while measures tied to efficiency were negatively associated. Perceptions of program effectiveness were related to the perception that measures are aligned with organizational goals. More statistically stringent methods of creating composite quality scores was more reliable than raw sum scores. The cost effectiveness of P4P varies widely by measure. Under both the QOF and in the VHA, removing an incentive from a measure had little impact on performance once a high-performance level had been achieved. Increasing maximum thresholds resulted in greater increases by poorer-performing practices.		Measures targeting process-of-care or clinical outcomes that are transparently evidence-based and viewed as clinically important may inspire more positive change than programs using measures targeted to efficiency or productivity, or that do not explicitly engage providers from the outset. Incentive structure needs to carefully consider several factors, including incentive size, frequency, and target. Incentivized measures must be congruent with institutional priorities, must address the needs of the institution at the local level, and must be designed to best serve the local patient population. P4P programs should have the capacity to change over time in response to ongoing measurement of data and provider input, should be

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			flexible and should be evaluated on an ongoing and regular basis. Improvements associated with measures achieving high performance can be sustained after the measure has been de-incentivized. Consistent evaluation of the performance of and adjustments to incentivized measures will allow institutions to shift focus and attention to areas in greatest need of improvement.
Korda et al., 2011	Endorsed by the Institute of Medicine (2007). here have been questions about the lack of standard measures used to reward providers and concern that financial incentives may widen health disparities as providers seek to maximize patient care revenue by selecting "easier," less complex and less socioeconomically diverse patients. P4P does not encourage integration across providers. Possible shortcomings and unintended program consequences include inappropriate measures and objectives, competing or uncoordinated efforts, insufficient or inappropriate incentives, and excessive focus on the reward. MedPAC recommends that the P4P system be budget neutral, with the incentive pool funded by setting aside 1% or 2% of budgeted payments. The evidence on performance-based incentives, such as pay-for-performance arrangements, is less convincing. There is little demonstrable return on investment (i.e., evidence of net savings) from such programs. Because the U.S. health care system is characterized by a large number of overlapping contracts among payers (i.e., health plans and government programs) and providers, financial incentives introduced by any one payer must account for a relatively large percentage of total reimbursement to justify any quality improvement effort with substantial fixed costs. There is no empirical evidence suggesting how large a payment gradient needs to be to stimulate quality improvement.	Possible unintended consequences of P4P arrangements include gaming, where participants find ways to maximize measurable results without actually accomplishing the desired objective; skimming of healthier patients for treatment by physicians; and the multi-tasking problem, where compensation based on available measures may distort effort away from unmeasured objectives. Among other limitations of pay-for-performance are: defining and unifying measures across the vast number of reporting initiatives, risk adjustment for clinical outcome measures, resource burdens on smaller versus larger hospitals, and the need for data on the effectiveness of pay-for-performance in improving care processes and outcome. There is no empirical evidence suggesting how large a payment gradient needs to be to stimulate quality improvement.	MedPAC recommends that the P4P system be budget neutral, with the incentive pool funded by setting aside 1% or 2% of budgeted payments. Incentives also must be clearly communicated, understood, and transparent to physicians and other providers.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Successful incentive arrangements for integrated care models address care coordination and recognize the contributions of all members of the interdisciplinary care team. Rewarding members of the team equitably encourages smooth communication and collaboration to share the effort required for high value care.		
Langdown et al., 2014	QOF has led to an improvement in health outcomes for some conditions including Diabetes, although the results are mixed for others such as CHD. Also, despite a surge of improvement during its introductory period for some conditions, levels of achievement reached a plateau in later years which may be due to a ceiling effect caused by the maximum threshold levels set for each indicator as practices were not incentivized to improve health outcomes beyond the various clinical target threshold levels (e.g. 85% of practice population). The studies highlight that the QOF is currently limited in what it measures in terms of health outcomes. Only one indicator is solely focused on the achievement of an actual health outcome (i.e. the number of Epilepsy patients which have been seizure free in the last 15 months), whereas the remaining intermediate outcomes relate to targets which are an indirect measure of one's health, e.g. cholesterol,5 mmol/l. The QOF points available are also weighted towards particular conditions such as Diabetes (88 points), and CHD- secondary prevention (69 points); compared to scores available for COPD (30 points) and Depression (31 points).	Non-incentivized activities did decline over the longer term in comparison to the temporal trend which existed following the introduction of the QOF, with the exception of Diabetes for which there was no significant change from the trend. The size of the incentive must be large enough to influence the clinician's behavior. This may imply that the relationship between improved health outcomes and incentivized activities under the QOF may be closer related to the number of QOF points available rather than whether a clinical activity is incentivized or not, particularly given that the scheme is voluntary.	The scheme can be used to inform practices of their population's health needs; however, the incentives operate in a way that rewards practices for 'high-workload activities' rather than influencing practices to proactively address health needs and provide preventative services. The evidence also demonstrates that although more practices are achieving higher or maximum QOF points, the ceiling placed on indicator thresholds do not incentivize practices to address the needs of all their population.
Latham et. al., 2015	Nationally implemented incentive program coupled with integration of pay-for-performance elements into primary care physicians' salaries may be effective in improving the quality of diabetes care. However, even broadly implemented incentive programs such as QOF have demonstrated mainly the effects on process and intermediate clinical diabetes outcomes. More evidence is required to understand whether these improvements are sustained and translate into better long-term outcomes such as reduced hospitalizations for diabetes-related complications. In the UK, incentive models have spurred some improvements in process outcomes and achievement of cholesterol, blood pressure and A1C targets. Still, the evidence is mixed, and its interpretation is hampered by methodologic challenges and confounders. Broad implementation and uptake of QOF means there is no adequate control group. Interpreting trends in	Based on the studies conducted in 2009, 2011 concerns have also emerged that patients from disadvantaged and vulnerable populations may be disproportionately excepted from QOF because their diabetes may be more challenging to manage. Patients with longstanding diabetes or multiple comorbidities were also more likely to be excluded from the A1C indicator. The same study of 2011 and other earlier studies found that QOF does not address	

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	quality improvement is also complicated by the fact that the quality of diabetes care was improving prior to the introduction of QOF in 2004. quality of care improved over and above the pre-incentive trend in the first year after the introduction of QOF by about 14.2%. However, during the second and third years, this difference diminished, and there was increasing variability according to patient demographic characteristics. An interrupted, time series analysis of a subset of family practices found that the introduction of QOF accelerated improvements in diabetes quality of care between 2003 and 2005, but this trend had slowed by 2007. The same study found that continuity of care was reduced after the introduction of QOF, while no changes in reported access to care were observed. Australia. physician encounters with patients with diabetes concluded that participation in PIP increased the likelihood of a physician's ordering an A1C test. Taiwan. Several cross-sectional studies have indicated that patients enrolled in Taiwan's incentive program were more likely to receive guideline-recommended tests and examinations. A longitudinal study of the effect of the Taiwanese program on hospitalization rates found that patients enrolled in the incentive program were less likely to be hospitalized after 3 years of care compared with non-enrolled patients. demonstrated that older patients and those with higher comorbidity and severity of disease are more likely to be excluded from this program. In Ontario, researchers studying completion of recommended diabetes management practices before and after the introduction of the incentive billing code found minimal improvements to monitoring practices. Selectively rewarding primary care physicians may discourage teamwork and coordinated care with other members of the healthcare team.	ethnic disparities in diabetes care adequately.	
Lin et al., 2016	clinical effects of P4P for most diseases has a certain improvement, medical costs will also increase. Thirty-six studies identified have showed the impact on the management of diseases. Thirteen focused on the preventive care 10 of which reported the positive results in vaccine injection or screening of diseases such as cervical cancer screen. Twelve focused on the hypertension of which 11 presented positive results. All from 14 related to coronary heart disease showed positive	Medical unfairness is still rather serious, patient satisfaction has no significant improvement. In Taiwan confirmed the results that primary practices with lower baseline level of medical quality tended to exclude patients with severe condition, so as to	Small practices demonstrated better results compared to bigger practices: when related to the process indicators of P4P, such as physicians prescriptions of examination ($P < 0.001$) or drugs ($P = 0.001$), the quality of primary care in smaller

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	results. All of the 26 studies about the management of diabetes reported significant improvement. mostly on process clinical outcomes. In addition, if the medical indicators of P4P are categorized into process indicators (clinical behavior of physicians, like ordering a test) and endpoint indicators (biochemical test/physical examination/history taking result of patients, like blood pressure level), positive results were not reported in both. It was found (five studies) that patients recruited in the practices with lower health baseline level and poorer compliance can benefit more from the new improved primary care. In addition, the practice with better quality of service before improved less than the practices with worse baseline before ($P > 0.05$). only one RCT drew the conclusion that the baseline level was unrelated to the improvement of the quality of primary care ($P = 0.22$). Related to process indicators quality of care in smaller practices improved more, although one study showed no difference and one study showed opposite results. A total of 20 studies reported various impacts on equity. Factors which influenced the equity of the health care included genders of patients or physicians, ages of patients and physicians, socioeconomic status of patients, ethnic of patients, comorbidity or severity. What is worse, doctors preferred to treat patients with milder disease condition or better socioeconomic status, which not only intensifies the inequity, but also is likely to exaggerate the improvement of clinical performance.	show great promotion in clinical performance apparently.	practices improved more. In addition to this, another study found that except the management of chronic obstructive pulmonary disease ($P = 0.1$), the management of diabetes ($P = 0.004$), hypertension ($P < 0.001$), and coronary heart disease ($P = 0.01$) all improved more in smaller practice. Studies demonstrated that patients recruited in the practices with lower health baseline level and poorer compliance can benefit more from the new improved primary care. In addition, the practice with better quality of service before improved less than the practices with worse baseline before ($P > 0.05$). Ceiling effect. After practitioners and medical institutions had achieved the upper limit of P4P indicators, their improvement for medical care quality would soon reach a plateau, which was called the ceiling effect. Earlier UK experience showed that when ceiling effect happened, not only did the quality of medical service cease to improve, but also other medical indicators, unrelated to payment, saw a drop to a certain degree.
Mabuchi et al., 2018	During the pre-pilot phase in 33 PHCCs in Adamawa, Nasarawa and Ondo states which started in December 2011, the PBF created large variations in performance among the participating PHCCs. For example, coverage of institutional delivery was around 10% of catchment population before the PBF in all target PHCCs, high-performers achieved 80–90% coverage while low-performers struggled with 20–30% coverage.		(i) Contextual and health system factors particularly staffing, access and competition with other providers; (ii) health center management including community engagement, performance management and staff

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	The examples of high-performing PHCCs in Nigeria provide a clear picture of how primary health centers can improve their performance with sufficient levels of autonomy and support. It should be noted that the performance of high-performing PHCCs was equally very low and the difference with low performers was negligible before the PBF scheme.		management (proactive engagement by PHCCs to recruit patients, and community leaders' support to encourage PHCC use and regulate unauthorized providers thus reducing competition); and (iii) community leader support (involving local authorities and communities and adapting approaches to the local situation) interacted and drove performance improvement among the PHCCs. The performance and staff management activities (system of accountability, various measures to improve staff motivation and team work) were interlinked and mutually reinforcing as strong staff awareness of plans and targets motivated staff, and motivated collaborative teams appear to improve performance management and community engagement activities.
Markovitz and Ryan, 2017	Heterogeneity in the effects of P4P does not fundamentally alter current assessments about its effectiveness (that P4P has largely failed to realize substantial quality improvements). Discussion around heterogeneity and treating them as modifying or direct effects is important as there are important explanations of success or failure of P4P programs.		Serving poor patients and patients of color was associated with lower performance at baseline and over time under P4P. In the US larger practices outperformed independent practice associations and smaller ones, opposite was found under the QOF. Evidence of the direct and modifying effects of patient age, gender and

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			health status is mixed and inconclusive. Several organizational factors such as quality improvement engagement, higher staffing level and greater financial strengths yielded conflicting results. Low performing practices do not appear to give up, even when they are unrealistically far from reaching high performance targets.
Mauro et al., 2019	The studies discussed have demonstrated the heterogeneous effects of financial incentives on improving the delivery rates of health preventive services - cancer screening services. In particular, for breast cancer screening, most of the studies showed partial or no effects; one explanation could be that women may take a proactive role in breast cancer screening, making physician incentives less important. For cervical cancer screening, 6 studies showed positive effects, 3 partial effects, 5 no effect, and 1 negative effects. Wee et al. examined the proportion of Pap smears carried out among women 20 to 75 years old and found that patients cared by physicians with financial productivity incentives were significantly less likely than those cared by physicians without this incentive to receive Pap smears (74.6% vs 86.3%). Thus, it is important to note that even if cervical cancer screening is mostly performed during gynecologist consultations, GPs' roles are essential: most GPs declare that they routinely perform cervical cancer screening and that performing this act is part of their job. Few positive or irrelevant effects were found regarding colorectal cancer screening. In this context, many guidelines have a positive position on the effectiveness of screening. However, screening rates are still low in some countries, and many barriers are present. Overall, many factors influence the impact of financial incentives on cancer screening delivery rates. Among these		Breast cancer screening rates in France French P4P program (CAPI) has not changed significantly since the P4P program implementation. According to the authors' conclusions, this result may reflect the fact that the low-powered incentives implemented in France through the CAPI might not provide sufficient leverage to generate better practices in the field of prevention and screening.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	factors, the low-powered incentives might not provide sufficient leverage to generate better practices in the field of prevention.		
Mendelson et. al., 2017	Low-strength contradictory evidence that P4P programs may improve process-of-care outcomes over the short term (2 to 3 years). Most evidence from QOF programs. Evidence on the longer-term effects was limited. Biggest improvements seen in areas with poor baseline performance. No clear evidence that P4P improves patient health outcomes. Stronger study designs showed no effect on utilization outcomes (hospitalizations, emergency or ambulatory care-sensitive visits). No clear evidence on intermediate health (such as a laboratory value or blood pressure, etc.) outcomes. Eight of the studies (most of which found positive results) were conducted in Taiwan and should be interpreted with caution due to selection bias (patients enrollment in the scheme).	Very limited evidence assessing the extent of gaming, no consistent evidence of a negative effect on health disparities, and a small amount of evidence suggesting the potential for both positive and negative effects on unincentivized measures. Qualitative studies reporting that P4P programs are imposing a considerable burden and threatening clinical autonomy.	Importance of designing P4P programs using the principles of behavioral economics, in which such factors as payment size, timing, and frequency of payment have effect on behavior. Careful consideration of number of measures, use of incentives in the most needed areas, review measures regularly and discontinue after achieving sustained improvements.
Odutolu et al., 2016	PHC accountability varies significantly between the three NSHIP states, and its pattern is mirrored in differences in service utilization performance. Between 2013 and 2015, all three states recorded very significant increases in service utilization for the three focus indicators. Average coverage for institutional normal deliveries in the project states increased from 2% in 2013 to 33.1% in 2015. In the same period, the average coverage for utilization of modern family planning methods increased from 1.04% to 21.3%, and the average coverage for completely vaccinated children increased from 1.4% to 49.2%. PBF implementation contributed to the success recorded: On the one hand, that may be true because of an injection of much-needed funds at every level of the health system and because of the autonomy that comes with PBF, which allows institutions, all the way down to the PHC facilities, to take managerial decisions, including how to allocate funds, thus avoiding the inefficiencies of central bureaucracy. The Primary Health Care Under One Roof (PHCUOR) and PBF reforms have therefore mutually reinforced each other, jointly strengthening the (Nigeria State Health Investment Project - NSHIP states') health system as a whole		In order to ensure sustainability, however, political commitment will still need to be reinforced.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
Ogundeji et al 2016	P4P scheme did not result in health facility performance improvement as its implementation considerations e.g. delay in payment, ineffective communication, incomplete incentive payment, and skepticism in the division of bonuses (individual assessment tool) generally led to distrust and uncertainty in payment, possibly led to decreased health worker motivation. Findings are consistent with that of the review by Eijkenaar et al. (2013) which found that P4P schemes in which health service providers were not knowledgeable about the schemes were mostly ineffective or unsuccessful.		Poor motivation of health workers results from a combination of factors such as poor salaries, poor working conditions, inadequate infrastructure and limited opportunity for career development or training, lack of government ownership of this health financing mechanism, lack of understanding of the P4P scheme; delayed incentive payments. Factors that should be considered for scheme successful implementation are the following: • Make timely quarterly payments to each health facility for delivery of services as agreed in the P4P contract. • Ensure clear communication strategies about changes and difficulties encountered in the scheme to stakeholders, particularly to inform and keep the health workers up to date. • Include a criterion/a set of criteria that captures individual contribution of health workers in the individual assessment tool (basis by which individual health workers earn bonuses). For example, a criterion on outreaches or home visits could be included. • Provide clear and short guidelines to encourage the use of the individual assessment tool to allocate bonuses to the health workers. • Provide training and regular workshops for health workers and

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			equip health facility managers with materials to help improve their managerial skills, with a focus on setting priorities, and recognizing and meeting the needs of the health facility or how to motivate the health workers (whether it is infrastructure or hiring additional staff). • Make one-off investments in the poorer facilities by either the scheme implementers or the State governments, so as to bring the concerned health facilities to an acceptable standard for a more effective program.
Patel, S. et. al 2018	In Cambodia, P4P did not have a significant effect on antenatal care (3 percentage point increase) or vaccination (2.3 percentage point increase). The Plan Nacer in Argentina demonstrated a significant positive effect on increasing prenatal visits (6.8 percentage point increase) and provision of tetanus toxoid (5.6 percentage point increase), as well as a very significant reduction in neonatal mortality (74%) in the beneficiary group. There was also a positive spillover effect with an Overall 22% reduction in neonatal mortality (beneficiaries and non-beneficiaries) using the same clinics. In Misiones province of Argentina the strongest evidence for sustained impact from P4P was seen with a substantial 3-fold increase in incentives. Demonstrated a 7%-9% improvement in General Self-reported Health and age adjusted wasting over time in the P4P group. Provider clinical Mean Vignette score for child health increased by 9.7% points. Clinical outcomes for under-five children improved by 9% (Children underweight for height following discharge from hospital for diarrhea and pneumonia).	Interesting, there was also a positive spillover effect with an overall 22% reduction in neonatal mortality (beneficiaries and non-beneficiaries) using the same clinics.	The extent to which the P4P scheme actually had on the improved quality of care has to be viewed within the economic, policy and overall context of the country. The perception and acceptance of P4P programs by health workers needs careful consideration during planning and implementation. Lack of understanding can undermine the potential impact of P4P program by limiting the behavioral response of health workers. The overall number of indicators measured needs to be carefully considered and should cover all aspects of quality and not focus on structural quality. In addition, clear communication about the structure of P4P programs

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Authors estimated the large impact of higher quality care with 294 cases of wasting averted and 229 more children reporting at least good health. Positive effect of measuring quality without incentives was found, whereby the act of measurement and feedback in itself led to improvement from awareness and consequent motivation to perform better. Quality effects seen with incentives provided to individuals may also be possible through indirect financial incentives that operate at the system level. These effects on quality affected performance earlier and to a greater degree than measurement and feedback of performance alone.		to health workers will likely improve the acceptance of them. In this regard, careful thought should be given to select indicators that will be acceptable to providers but can also maximize the efficiency of spending. Adequate levels of incentives as health workers may not feel the added effort is worth the reward. Monitoring and verification is essential to ensure quantity and quality objectives are being met. Feeding performance data back to providers facilitates performance improvement. It is suggested that the 'easier' structural quality indicators are addressed first and then programs can move onto introducing process measures of clinical care. This will allow health providers to address less complex quality of care issues first, develop better understanding of RBF and quality of care, and then shift gradually toward more demanding measures of care under the RBF programs.
Paul et al., 2018	-		Context nature of incentive in Cambodia: PBF is more likely to succeed when income, training needs, and the desire for a sense of community service are addressed and institutionalized within the health system. Basic salary and the bonus amount affect the motivation:

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			In Cambodia financial incentives accounted to 42% of the average total income of a health worker and was associated with higher job motivation.
Peckham et al., 2011	In the UK, the evidence of whether the QOF rewards outputs that are expected to lead to good outcomes is contradictory, demonstrating both that meeting certain QOF indicators might improve health outcomes in some areas and a weak causal relationship between key clinical indicators and outcomes. Recent systematic reviews have concluded that P4P contracts do affect physician behavior and increase the number of primary care services provided – although often in complex and limited ways. The actual effect depends on factors such as the age and sex of physicians, previous experience of financial incentives, the uptake of continuing professional education, the type of payment method, the type and severity of the conditions targeted through incentives, the volume of activity and the location and type of organization. The size of incentive has also been found to be less of a factor in the use of care management processes for patients with chronic illnesses by physician organizations (POs) than are schemes that give public recognition for scoring well on quality of care measures, schemes which require POs to provide quality of care or outcomes data to outside organizations or those that reward high-quality scores with better contracts that assist in developing better organized quality provision.	A key concern that recurs in the literature is whether financial incentives generate dysfunctional physician behavior or negatively affect motivation, particularly in the light of well-established inverse care patterns at primary care level. Impact of externally structured incentives such as financial inducements is that they might 'crowd out' professional self-esteem and a sense of self-determination. This might have implications for the quality of care offered by practitioners. However, it has been noted that there is an equal chance of a 'crowding in' effect if practitioners feel like they have some ownership of incentives. Another potential problem created by external financial incentive schemes is that they could lead to the neglect of those non-incentivized areas of care which will continue to rely on the professionalism or moral motivation of GPs. There is some evidence of concern amongst GPs that non-incentivized areas like acute care, preventive care, care for specific groups such as children or older people and care for patients with multiple comorbidities would suffer as GPs chased targets. There is some concern that the QOF may lead to an exacerbation of health	The size and structure of incentives seems to be important in incentivizing effective physician activity. Incentives have to be large enough to influence behavior and designed in such a way that they cannot be played off so as to reward both process and improved outcomes. It is technically challenging to connect performance targets with health gain and most P4P schemes adopt a pragmatic approach and focus on processes (such as measuring blood pressure) and intermediate outcomes (controlled blood pressure) for which there is either evidence or professional consensus and which can be easily measured and rewarded. This means that treatment and secondary prevention is favored over primary prevention and can lead to the marginalization of some conditions. The actual effect depends on factors such as the age and sex of physicians, previous experience of financial incentives, the uptake of continuing professional education, the type of payment method, the type and severity of the conditions targeted through incentives, the volume of

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
		inequalities by allowing GPs to use the exception reporting system to exclude high-risk patients , or by not sufficiently rewarding the extra work required in delivering equal treatment to disadvantaged populations, maintaining inverse care patterns. Incentive payments may skew physician activity towards high-reward labor-intensive activities with relatively low health benefits, thereby marginalizing non-incentivized areas. This potential for 'gaming' may create a conflict of interest for physicians between maximizing revenue and ensuring good quality care. Financial incentives may also distort care by encouraging a focus on individual measures for care management instead of a more integrated approach which might be appropriate, particularly in areas of comorbidity. In addition, the use of targets and financial incentives can have unintended consequences on practitioner behavior, such as goal displacement and rule following, leading to the 'crowding out' of and reduction in focus on non-incentivized tasks.	activity and the location and type of organization.
Petrosyan et al., 2017	Introduction of RBF contributed to the improvement PHC service utilization: average number of visits to PHC facilities per person per year had increased from 2.0 in 2000 to 4.0 in 2013. RBF scheme played a role to improve the maternal and child health and NCD services in PHC facilities to meet annual targets.		Stable economic growth enabled the government to begin implementing much needed reforms of its social system, spending significant resources of its own in the process and building a sense of national ownership for programs such as the RBF program.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			An important enabler was a well-sequenced reform process that included the most politically important stakeholders. The pilot project was designed and implemented over a three-year period, with indicators added progressively, a practice that persisted even after the nationwide implementation and scale-up. Another enabler was the embedding of RBF in national regulatory frameworks and the provision of funds from the national budget. Both the piloting and subsequent scale-up of the OE (open enrolment) mechanism were brought about through legal decrees and the amendments of earlier rules and regulations. With respect to funding, not only did the State Health Agency provide funds for the initial piloting of the program but there was also a medium-term budgetary commitment for the RBF program through the MTEF, reflecting a degree of national ownership of the program. Finally, an important enabler to the subsequent scale-up and integration of RBF into the PHC system, as opposed to it remaining a vertical program, was its introduction as part of a larger reform of the primary care system. This reform included efforts to enhance financing for primary care, to introduce OE, to introduce

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			measures for improving quality of care, to strengthen family medicine, to renovate facilities and provide equipment, to develop a health information system as well as to prepare policies and procedures for nationwide extension of all aspects of PHC reforms.
Renmans et al., 2016	PBF scheme had no effect on neonatal mortality in Cambodia despite the rise in institutional deliveries.	Misreporting had decreased in Cambodia thanks to regular monitoring, random verification and web-based reporting.	In Cambodia, the M&E arrangements helped to limit rent seeking behavior and reduce absenteeism. Administrative burden was reported in Cambodia caused by time spent by health workers and managers on PBF activities, verification system, etc.
Saddi et al., 2018			Themes such as organizational capacity, staff engagement, professional stress, and work overload are also extensively considered. Organizational capacity issues have also been considered important to highlight the need for capacity building in African countries, for instance, and foster the successful delivery of performance programs. Researchers have taken into account the cognitive/ subjective aspects ("alternative logics") in performance measurement and claimed that focusing on what is measured induces potentially dysfunctional effort substitution and gaming behaviors. Moreover, performance indicators have been considered political instruments and used in diverse and complementary ways in the

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			construction of improvement frameworks and tools to measure and monitor policies. Undesired effects of P4P will often be a result of diminished intrinsic motivation. It is therefore important that providers are actively involved in designing the program, especially in developing and maintaining the aspects of performance to be measured. This increases the likelihood of provider support and alignment with their professional norms and values . . . In this respect, it is also important that program evaluations include qualitative studies to monitor the impact on providers' intrinsic motivations. Findings have also revealed, in this case related (possibly) more to middle- and low-income countries, that workers and managers were not fully aware of performance indicators and standards. Furthermore, frontline professionals have limited prospects for career progression, and there have been inadequate performance feedback and poor reward mechanisms
Scott et al., 2011	Six of the seven studies included in this review showed positive but modest effect on a minority of the measures of quality of care included in the study. There is insufficient evidence to support or not support the use of financial incentives to improve the quality of primary health care.		Implementation of financial incentive schemes should proceed with caution and should be more carefully designed before implementation. Studies should more consistently describe

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Although six out of the seven studies found a statistically significant and positive effect the majority of these were across only one out of a range of quality measures used in each study. There was significant heterogeneity across the studies in terms of the types of financial incentives used, the contexts in which they were implemented, and the types of outcome measures (also in uncertainty).		i) the type of payment scheme at baseline or in the control group, ii) how payments to medical groups were used and distributed, and iii) the size of the new payments as a percentage of total revenue.
Scott et al., 2018	Of all 44 schemes 46% of outcome measures were positive (these include wide range of outcome measures including expenditures and quality of care). Weaker study designs were more likely to show positive effects, suggesting that as study designs improve the likelihood of finding stronger effects will be lower. Schemes from the US had the same probability of finding an effect as non-US studies. The key innovation in the US has been the combination of rewards for P4P with rewards for reducing costs such as one- and two-sided risk sharing models, yet this seems no better than P4P alone in terms of the proportion of positive outcomes. Many shared savings models are in their early stages, and so more evidence is required to examine if this persists over time. A key finding is that schemes that reward for improvements in performance over time have a lower probability of being effective than those that do not. This is important to understand further as the dynamics of incentive schemes are complex. Schemes that did not reward for performance improvement included single threshold schemes but also other types of scheme such as value-based pricing of DRGs. The behavioral effects also depend on a range of more specific factors that could not be easily captured due to heterogeneity and small sample sizes, including the distance between measures/thresholds (i.e., the number of thresholds), whether payments are nonlinear (e.g., increasing) at each time point/threshold, and whether the thresholds are set high or low in the distribution of performance. We find weak evidence that schemes allowing incentive funding to be used for specific (but non-physician income) purposes leads to a higher probability of an effect compared with physicians being allowed to use incentive funding as income.	Evidence suggests that there was a reduction in expenditure growth for Medicare patients who were not covered, but who were enrolled with the same provider organizations participating in the Alternative Quality Contract	Schemes need to build rigorous evaluation into the implementation and roll out of schemes if knowledge is to improve.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	The size of the incentives as a percentage of revenue was not associated with the probability of an effect, contrary to expectations. Though the sample size was small (22 schemes), the scatterplot did not show a clear relationship between incentive size and effect and so increasing the sample size may not make a difference if future studies are similar. Each of the 80 empirical studies reported an average of 16.3 outcome measures, of which 7.4 were positive and statistically significant. The mean percentage of positive outcomes per study was 54%. Of the 25 schemes in the US, an average of 56% of outcomes was positive and statistically significant. This compares with 91% for the P4P schemes in Taiwan, 75% in Canada and Italy, and 48% in the UK. In US Six out of eight studies of the Alternative Quality Contract (two sided ran by private insurance) showed an impact of the scheme on both reducing spending and improvements in quality after 4 years of the scheme. The nine studies of the three schemes conducted in Taiwan of the National Health Insurance P4P scheme, eight showed a positive effect. Six studies evaluated the impact of the scheme on diabetes care. The study that showed a negative effect showed an increase in emergency admissions for diabetes patients. Further study found that patients in the program were more likely to receive guideline-recommended tests and examinations, and that this was also the case for patients not enrolled in the program but seeing the same physicians. Two studies examined tuberculosis treatment and found that the cure rate, length of treatment, and default rates improved. Finally, one study examined breast cancer screening and found patients had improved quality of care, higher 5-year survival rates, and lower rates of reoccurrence. Other studies of the Taiwan schemes have shown that there may have been substantial selection bias of patients enrolled in the program, such that any positive effects were likely to have been due to selection rather than the impact of the program. Thirteen of the 44 schemes (37/80 studies) reported using a design that provided incentives for performance improvement. The results from the regression show that the percentage of positive outcomes from these schemes		

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	was just over 20 percentage points lower compared with schemes that did not P4P improvement (attainment schemes with single threshold), and this is statistically significant. The schemes conducted in the US had around an 8 % lower percentage of positive outcomes compared with other countries, but this was not statistically significant. Serumaga 2011 evaluated the effects of P4P incentive on quality of care and outcomes among patients in the UK with hypertension in primary care. This study included patient utilization and patient health outcomes: the percentage of patients with blood pressure measured, the proportion of patients with controlled blood pressure, and the percentage of patients with hypertension-related adverse outcomes (myocardial infarction, stroke, renal failure, heart failure). It found that there was little or no change in levels and change trends of these outcome measures.		
So and Wright 2012	P4P can improve quality, the type, amount, and timeliness of the incentives all affected the magnitude of the behavioral change and the potential benefit of the strategy. Only three studies did not report improvement. Improvement may be sustained even after intervention, but at least one study suggested not only reduction in continuity in care once targets were achieved, but decline in the rate of quality of care improvement with time.	A potential unintended consequence of pay-for-performance was the increase in health inequalities with an incentive to select healthier patients and avoiding reducing income by serving minority populations. However, at least in the UK, minimal reductions in chronic disease management were observed. Sustainability of gains was another issue. Improvement may be sustained even after intervention, but at least one study suggested not only reduction in continuity in care once targets were achieved, but decline in the rate of quality of care improvement with time.	P4P, to be effective, needs to consider all aspects of quality of care, including reduction in disparities and improvement in access to care with a consideration of anticipated and potential unanticipated outcomes.
Soranz et al., 2017	Longitudinality and of access in PHC - increasing trend, but always below the upper limit of the targeted 90%. The lowest point observed in the first quarter of 2013 can be explained by the increased entry and arrival of new medical residents to the health units, generating redeployment of physicians among teams and units.		P4P indicators are reviewed every 2-3 years in order to avoid the gaming behavior described in the literature. Monitoring process indicators are a key step towards ensuring good outcomes. In PHC, it is fundamental

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	It showed an average score of 6.1 for users and 7.5 for health professionals (previous study in 2012), with significant differences between the evaluated units but with an average close to that expected by the management baseline. These scores are close to the recommended minimum of 6.6 to have a strong and quality PHC measured in the subject tool. In Brazil the proportion of hypertensive patients with a blood pressure record in the last 6 months remained between 60 and below 70% throughout 2012 and 2016, and greater investments are still required to improve this basic indicator. In Portugal, this goal is established as interval between 38 - 80%. Primary health care patients' referrals to other health system levels - good process of coordination of care was observed in Rio de Janeiro's PHC, since the upper limit of the goal has never been achieved.		to use information systems that allow the association of health indicators (structure, process and results) with the primary healthcare attributes.
Soranz et al., 2017 (2)	P4P has a relevant influence on clinical practice. With this payment, quality and quantity go hand in hand. It is about rewarding good practices and the associated workload. Brazil: Hospitalizations for PHC-sensitive conditions are an indirect measure of the clinical efficacy of primary healthcare for certain health problems. Compared with other capitals of the Brazilian Southeast and South, there was a significant decline in the proportion of sensitive conditions, placing Rio de Janeiro at the second lowest proportion of hospitalizations (10.5%) for sensitive conditions in 2014, behind Curitiba with 8.8% of sensitive conditions against total hospitalizations. Portugal: In the indicators of preventive health care (indicators of oncology surveillance, screening and vaccination plan) and disease prevalence, "B model" Family Health Units (USFs) evidenced a better performance, followed by "A model" USFs.		From the political standpoint, mixed payment models, with well-explained quantitative and qualitative objectives and increased desirability of group incentives are recommended as long as these indicators are updated and revised every one or two years. Pay-for-performance is a payment method and not an absolute guarantee of health gains. Teamwork with motivation of the professionals: motivated workers are the true engines of reform and change. A good leadership of a primary healthcare facility perceives its culture and uniqueness, creates a participatory climate with autonomy and responsibility, delegation, objective identification of action areas, monitoring, good working environment, promoting a good relationship between people.

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			Investing in the information system and computerization - the development of interoperability and individual electronic records of patients allowed monthly monitoring of teams and their indicators. Technical leadership – it requires a clinical governance policy that is viewed as a set of quality-based policies, strategies and processes that can ensure continuous improvement in the way the health facility unit cares and treats its patients, in the way it is accountable to the community and to the tutelage and efficiency in managing resources entrusted to it. The effective exercise of clinical governance is not achieved by decree. It is not a matter of achieving a goal, but of going a long way, which requires from the genuine start the will to change and openness to new models of thinking, managing and providing health care. Political leadership - clear and unequivocal support from the highest political officials, in particular the Minister of Health of Portugal and the Mayor of Rio de Janeiro.
Tao et al., 2016	Little scientific evidence supporting an association between reimbursement system and socioeconomic or racial inequity in access, utilization and quality of primary care.		
Van Herck et al., 2012	Clinical effectiveness The effects of P4P ranged from negative or absent to positive (1 to 10%) or very positive (above 10%), depending on the target and program. Negative	Negative effects, in terms of less quality improvement compared to non P4P use, which were first reported in the review	Our review has further contributed to the contextual framework from a health system, payer, provider and

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	results were found only in a minority of cases: in three studies on one target, each of which also reported positive results on other targets. It is noteworthy that 'negative' in this context means less quality improvement compared to non P4P use and not a quality decline. In general, there was about 5% improvement due to P4P use, but with a lot of variation, depending on the measure and program. For preventive care, we found more conflicting results for screening targets than immunization targets. Across the studies, P4P most frequently failed to affect acute care. In chronic care, diabetes was the condition with the highest rates of quality improvement due to P4P implementation. Positive results were also reported for asthma and smoking cessation. This contrasts with finding no effect with regard to coronary heart disease (CHD) care. The effect of P4P on non-incentivized quality measures varied from none to positive. However, one study reported a declining trend in improvement rate for non-incentivized measures of asthma and CHD after a performance plateau was reached. Finally, one study found positive effects on P4P targets concerning coronary heart disease, COPD, hypertension and stroke when applied to non-incentivized medical conditions (10.9% effect size), suggesting a spillover effect. This implies a better performance on the same measures as included in a P4P program, but applied on patient groups outside off the program. Access and equity of care (mainly UK) In general, P4P did not have negative effects on patients of certain age groups, ethnicity, or socio-economic status, or patients with different comorbid conditions. This finding is supported by 28 studies with a balanced utilization of cross sectional, before after, time series and concurrent comparison research designs. Equity has not suffered under P4P implementation and is improving in the UK. Coordination and continuity of care directing P4P toward the coordination of care might have positive effects. One time-series study reported no effect on non-incentivized access and communication measures. This study, however, did observe a patient self-reported decrease in timely access to patients' regular doctors, which might be a negative spillover effect.	paper by Petersen et al (2006), are rarely encountered within the 128 studies, but do occur exceptionally. Previous authors also questioned the level of gaming, and possible neglecting effects on non-incentivized quality aspects. The presence of limited gaming is confirmed in this review, although it is only addressed in a minority of studies. Its assessment is obscured by uncertainty of the level of gaming in a non P4P context as a comparison point. As the results show, a few studies included non- incentivized measures as control variables for possible neglecting effects on non P4P quality targets. Such effects were absent in almost all of these studies. The results of one study suggest the need to monitor unintended consequences further and to refine the program more swiftly and fundamentally when the target potential becomes saturated. It is too early to draw firm conclusions about gaming and unintended consequences. However, based on the evidence, there may be some indications of the limited occurrence of gaming and a limited neglecting effect on non-incentivized measures. Positive spillover effects on non-incentivized medical conditions are observed in some cases, but need to be explored further. One study found positive effects on P4P targets concerning coronary heart disease, COPD, hypertension and stroke when applied to non-incentivized medical	patient perspective. Program development and context findings, which related P4P effects to its design and implementation within a cyclical approach, enable us to identify preliminary P4P program recommendations. Incentive forms are dependent on its objectives and contextual characteristics. However, considering the context and goals of a P4P program, six recommendations are supported by evidence throughout the 128 studies: 1. Select and define P4P targets based on baseline room for improvement. This important condition has been overlooked in many programs, with a clear effect on results. 2. Make use of process and (intermediary) outcome indicators as target measures. See also Petersen et al (2006) and Conrad and Perry (2009), who stress that some important preconditions, including adequate risk adjustment, must be fulfilled if outcome indicators are used. 3. Involve stakeholders and communicate the program thoroughly and directly throughout development, implementation, and evaluation. The importance of awareness was already stated previously. 4. Implement a uniform P4P design across payers. If not, program effects risk to be diluted. However, one

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	Patient Centeredness With regard to patient-centeredness, two studies–one Spanish before-and-after study without a control group and one cross-sectional study in the US–found no and positive P4P effects, respectively, on patient experience. Another before-and-after study, this one from Argentina, reported that P4P had no significant effect on patient satisfaction, due to a ceiling effect.	conditions (10.9% effect size), suggesting a spillover effect. This study, however, did observe a patient self-reported decrease in timely access to patients' regular doctors, which might be a negative spillover effect.	should be cautious for anti-trust issues. 5. Focus on quality improvement and achievement, as also recommended by Petersen et al (2006). The evidence shows that both may be effective when developed appropriately. A combination of both is most likely to support acceptance and to direct the incentive to both low and high performing providers. 6. Distribute incentives at the individual level and/or at the team level. Previous reviews disagreed on the P4P target level. As in our review, some authors listed evidence on the importance of incentivizing providers individually. Others questioned this, because of two arguments: First, the enabling role at a higher (institutional) level which controls the level of support and resources provided to the individual. Secondly, having a sufficiently large patient panel as a sample size per target to ensure measurement reliability. Our review confirms that targeting the individual has generally better effects than not to do so. A similar observation was made for incentives provided at a team level. Statistical objections become obsolete when following a uniform approach (see recommendation 4). The following recommendations are theory based but at present show absent evidence (no. 1) or conflicting evidence (no. 2

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
			and 3): 1. Timely refocus the programs when goals are fulfilled, but keep monitoring scores on old targets to see if achieved results are preserved. 2. Support participation and program effectiveness by means of a sufficient incentive size. As noted by other authors, there is an urgent need for further research on the dose-response relationship in P4P programs. This is especially important, because although no clear cut relation of incentive size and effect has been established, many P4P programs in the US make use of a remarkably low incentive size (mostly 1 to 2% of income). Conflicting evidence does not justify the use of any incentive size, while still expecting P4P programs to deliver results. 3. Provide quality improvement support to participants through staff, infrastructure, team functioning, and use of quality improvement tools.
Wekesah et al., 2016	Nigeria: A co-financing program for maternal health between the government and the community resulted in a 60% increase in the utilization of maternal health services (from 26.7 to 85.6 %).		
Yuan et al., 2017	Two comparisons related to P4P: 1) P4P plus some existing payment method (capitation or input-based payment) compared to the existing payment method; (12 of 14 studies)	Four studies reported some unintended or adverse effects. Petersen 2013 found that after the P4P intervention had ended, there was a significant reduction in blood pressure control and appropriate	Carefully consider each component of their P4P design, including the choice of performance measures, the performance target, payment frequency, if there will be additional

Reference	Effectiveness	Unintended consequences & spillover effect	Implementation consideration
	2) P4P combined with capitation compared to FFS. (1 RCT study on antibiotic prescription) Thirteen studies (included in the effects analysis) found that adding P4P to an existing payment method probably slightly improved the care provided by health professionals (moderate-certainty evidence) and may have little or no effect on utilization of health services (immunization, ANC) or patient outcomes (low-certainty evidence)	response to uncontrolled blood pressure in the intervention group compared with the control group (low-certainty evidence).	funding, whether the payment level is sufficient to change the behaviors of health providers, and whether the payment to facilities will be allocated to individual professionals. Unfortunately, the studies included in this review did not help to inform those considerations. Electronic information system or resources to support the administrative cost of P4P was used by P4P programs in developed countries.