Health Policy 110 (2013) 115–130

Contents lists available at SciVerse ScienceDirect

Health Policy

journal homepage: www.elsevier.com/locate/healthpol

Review

Effects of pay for performance in health care: A systematic review of systematic reviews

Frank Eijkenaar[a,*,1], Martin Emmert[b,1], Manfred Scheppach[b], Oliver Schöffski[b]

[a] Erasmus University Rotterdam, Institute of Health Policy and Management, Burgemeester Oudlaan 50, 3000 DR Rotterdam, The Netherlands
[b] Friedrich-Alexander-University Erlangen-Nuremberg, Chair of Health Business Administration, Lange Gasse 20, 90403 Nuremberg, Germany

ARTICLE INFO

Article history:
Received 13 September 2012
Received in revised form 9 January 2013
Accepted 11 January 2013

Keywords:
Pay-for-performance
Systematic review
Effects
Effectiveness
Cost-effectiveness
Unintended consequences

ABSTRACT

Background: A vast amount of literature on effects of pay-for-performance (P4P) in health care has been published. However, the evidence has become fragmented and it has become challenging to grasp the information included in it.
Objectives: To provide a comprehensive overview of effects of P4P in a broad sense by synthesizing findings from published systematic reviews.
Methods: Systematic literature search in five electronic databases for English, Spanish, and German language literature published between January 2000 and June 2011, supplemented by reference tracking and Internet searches. Two authors independently reviewed all titles, assessed articles' eligibility for inclusion, determined a methodological quality score for each included article, and extracted relevant data.
Results: Twenty-two reviews contain evidence on a wide variety of effects. Findings suggest that P4P can potentially be (cost-)effective, but the evidence is not convincing; many studies failed to find an effect and there are still few studies that convincingly disentangled the P4P effect from the effect of other improvement initiatives. Inequalities among socioeconomic groups have been attenuated, but other inequalities have largely persisted. There is some evidence of unintended consequences, including spillover effects on unincentivized care. Several design features appear important in reaching desired effects.
Conclusion: Although data is available on a wide variety of effects, strong conclusions cannot be drawn due to a limited number of studies with strong designs. In addition, relevant evidence on particular effects may have been missed because no review has explicitly focused on these effects. More research is necessary on the relative merits of P4P and other types of incentives, as well as on the long-term impact on patient health and costs.

1. Introduction

In many countries, healthcare delivery is suboptimal. For example, adherence to professional medical guidelines is often low [1–3], while costs of care continue to rise. Pay-for-performance (P4P) has become a popular approach to increase efficiency in health care. In addition to the United States where P4P has become widespread, P4P programs are being implemented in many other countries, including in the United Kingdom, Canada, New Zealand, Taiwan, Israel, and Germany [4]. In P4P, care providers receive explicit financial incentives based on their scores on specific performance measures that may pertain to clinical quality, resource use, and patient-reported outcomes.

Along with the dissemination of P4P, the literature on the effects of P4P has expanded rapidly over the past

* Corresponding author. Tel.: +31 10 408 9183; fax: +31 10 408 9094.
E-mail addresses: eijkenaar@bmg.eur.nl (F. Eijkenaar), Martin.Emmert@wiso.uni-erlangen.de (M. Emmert), manfred.scheppach@gmx.de (M. Scheppach), oliver.schoeffski@wiso.uni-erlangen.de (O. Schöffski).
[1] Both authors contributed equally to this work.

0168-8510/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.healthpol.2013.01.008

15 years. Although this is a desirable development, the evidence has become fragmented. Several systematic reviews have synthesized available evidence, but they all had different foci (e.g., only including experimental studies, only focusing on preventive services, not addressing other potential P4P effects besides impact on incentivized performance, etc.) and hence different conclusions. Consequently, it is challenging to comprehend this evidence and to extract success factors and pitfalls when it comes to implementing P4P.

In this paper, we summarize the existing literature on P4P effects in a broad sense by conducting a systematic review of published systematic reviews. The paper adds to the literature by synthesizing key findings from these reviews. The goal is to provide a structured, comprehensive overview of the evidence on P4P effects and mediating factors. We achieve this by addressing the following six questions: to what extent has P4P been (1) effective and (2) cost-effective? (3) Which unintended consequences of P4P have been observed? To what extent has P4P (4) affected inequalities in the quality of care and (5) been more successful when combined with non-financial incentives? (6) Which specific design features contribute to (un)desired effects? To our knowledge, no prior study has provided such an overview. The results will be of interest for policymakers who intend to implement a P4P-program as well as those who have already done so.

The next section provides a theoretical background on the relevance of these questions. Next, after describing the search strategy and inclusion and exclusion criteria, the results are presented for each question separately. In the discussion, the results are compared with findings from recent studies not included in any of the identified reviews (if available and relevant). We end with discussing the implications of our findings for research and policy.

2. Theoretical background

Effectiveness. Both economic theory and common sense support the notion that payment for health care should be determined, at least in part, based on meaningful indicators of quality or value [6]. Given notable deficiencies in the quality and efficiency of care, that healthcare providers (be they individual physicians, physician groups, or institutions) are responsive to financial incentives and that improving performance requires changes in their behavior, that many current base payment methods (e.g., fee-for-service, capitation) do not explicitly stimulate good performance, and that performance measurements have become more accurate, it seems natural to tie a portion of providers' compensation to their performance. However, although the idea underlying P4P is simple, in practice there are many potential pitfalls, as outlined below.

Cost-effectiveness. P4P can be considered cost-effective when improved quality is achieved with equal or lower costs or when the same quality is achieved with lower costs. Even in case P4P leads to cost increases it may still be viewed as cost-effective, as long as quality improvements are large enough [7]. Yet designing and implementing a successful P4P-program is complex [8]. Engaging providers, reaching consensus about program design, collecting and validating data, calculating payments, and maintaining and evaluating the program likely involve high transaction costs. This raises the question whether P4P can be cost-effective.

Unintended consequences. In theory, P4P may have several unintended consequences. First, when differences in casemix between providers are not taken into account, providers have an incentive to select healthy/compliant patients and to avoid severely ill/noncompliant patients, especially for outcome and resource use measures. Moreover, even sophisticated risk-adjustment models may fail in preventing selection because providers are likely to have superior information about their patients than included in these models [9]. Other strategies, such as allowing providers to exclude noncompliant patients from performance calculations [10], may be necessary. Second, P4P may cause providers to focus disproportionately on aspects of care that are incentivized and possibly neglect other important aspects that are not [11]. A broad set of measures (including e.g., clinical quality, patient satisfaction, continuity of care) seems therefore important. However, this is often not feasible in practice. Third, P4P may crowd out providers' intrinsic motivation to provide high quality care, especially when the definition of performance is not shared. P4P could then play a trivializing role regarding the non-financial motivation [12], which may have several undesired effects. Finally, to maximize income, providers may manipulate data so that their performance looks better than it is in reality ("gaming").

Inequalities. P4P may narrow, widen, or maintain inequalities regarding access to/receipt of high-quality care [13]. Inequalities may widen if P4P encourages risk selection or results in reduced income for providers serving minority populations [14]. Providers in deprived areas will typically have lower performance and be less likely to receive incentive payments compared to providers in affluent areas, for example because their patients are less likely to adhere to treatment [15]. By adversely affecting the income of providers practicing in deprived areas, P4P may reduce both the number of providers working in such areas and their ability to invest in performance improvement. Widening inequalities can be prevented by rewarding improvement in performance, adequate risk adjustment, inclusion of measures that are more important for minority patients, or directly rewarding reductions in inequalities [14–16].

Non-financial incentives. Non-financial incentives such as public reporting (PR) and timely performance feedback to providers may complement P4P incentives. PR and P4P both reward providers for good performance, but the financial incentive in PR operates indirectly via consumer choice [13]. Performance feedback and reminders make treatment patterns and performance issues salient and can activate providers to adjust their practice style. Feedback may also create a reputational incentive if reports include information on peer performance.

Program design. The design of P4P has important consequences for the incentives that physicians experience and how they respond to them [17]. Seemingly important design elements are the number and type of included performance measures, risk adjustment, the entity targeted

(individual physicians, groups, institutions), the type (rewards, penalties) and size of the incentive, frequency of payment, and type (absolute, relative, improvement) and number of performance targets [7,17,18]. In summarizing the literature, we attempt to infer about preferred design in practice by identifying patterns in the results.

3. Methods

For this review, we adhered to guidelines from the Cochrane Collaboration [19], the Institute for Quality and Efficiency in Health Care [20], the Hannoveraner Konsensus [21], and the NHS Economic Evaluation Database [22]. We searched five databases: Medline, Embase, ISI web of knowledge, the Cochrane Database of Systematic Reviews, and Scopus. Twenty-five expressions were entered in each database (see supplementary material). We also searched the Internet via Google, contacted experts, and reviewed reference lists.

Articles written in English, Spanish, or German published between January 2000 and June 2011 were eligible for inclusion. Two authors independently reviewed all titles generated by the search procedure and constructed a preliminary list of articles. These articles were subjected to abstract review and full texts of potentially relevant articles were obtained. Two authors independently assessed their eligibility for inclusion. Overview articles that were not systematic reviews and articles not covering at least one of the six domains were excluded. In addition, we excluded reviews that: only aimed to identify studies evaluating the effect of implicit financial incentives and/or excluded studies evaluating the effect of explicit financial incentives, only focused on financial incentives for patients, did not include empirical studies with original quantitative or qualitative data on P4P effect(s), are entirely overlapped by a subsequent review from largely the same authors, and/or did not (consistently) report the methodological design of included studies. The last criterion was applied because it would otherwise be impossible to assess the validity of reported results.

To determine the methodological quality of included reviews, we applied the checklist of the German Scientific Working Group, which contains 18 distinct criteria [23]. The items are grouped under five categories: research question, search procedure, evaluation of information, synthesis of information, and presentation of results. A total score is obtained by adding up points and dividing by the maximum number of points. Two authors independently carried out the scoring.

Two authors independently extracted relevant data from identified reviews using the same abstraction form containing the following elements: search period, number of studies, type of studies, sector and country in which studies were conducted, and a summary of the main results for each of the six domains. Because of the heterogeneity between studies, meta-analysis was not possible and results are presented narratively. To get an impression of the strength of the evidence, we assigned included primary studies to one of the following five categories: "level I" (systematic reviews, RCTs), "level II" (quasi-experiments, controlled before-after studies, time-series studies with before-after data), "level III" (uncontrolled before-after studies, controlled after studies), "level IV" (uncontrolled after studies, cross-sectional studies), and "others" (qualitative studies and studies that use statistical modeling to examine the effect the program will potentially have under certain assumptions using clinical data from previous studies). In some cases, the abstract or full text of individual primary studies was retrieved to verify the study design.

The findings from identified reviews are compared and expanded with findings from several recently published primary studies that are not included in any of the reviews but that do provide relevant information. These studies were not identified from an additional systematic review, but from our knowledge of the current evidence base on P4P effects. Although there may be more studies than the ones we discuss, comparing our results with the findings from additional studies we are aware of provides additional insight in the effects of P4P and enables us to draw stronger conclusions.

4. Results

The initial search identified 2004 articles (Fig. 1). After review of titles and abstracts, 487 studies remained for a detailed reflection. Reference tracking, Internet searches, and expert consultation yielded 28 additional articles. Of the 515 articles subjected to full-text review, 493 articles met at least one exclusion criterion, leaving 22 articles for inclusion in the review (the full list of excluded articles is available from the authors upon request).

Table 1 presents their main characteristics. The reviews vary considerably by inclusion criteria and focus. For example, some reviews focus only on one (subset of) condition(s) or on one specific sector. Others only include studies with a particular design (e.g., RCTs) while still others had no restrictions at all. The result is a wide range in the number of included studies. While most reviews only included studies from the US and the UK, studies conducted in other countries have increasingly been identified (10 in total). Most studies were conducted in primary care, although an increasing number of studies have evaluated P4P in other sectors (e.g., acute inpatient care). Evidence mainly comes from observational studies and many authors have therefore noted that results must be interpreted with caution. Table 2 and the following sections present the key findings for each of the six domains.

4.1. To what extent has P4P been effective?

Twenty reviews provide evidence on the effectiveness of P4P. We present the results according to the design of included studies: randomized controlled trials (level I) and non-randomized studies (levels II–IV).

Randomized controlled trials have largely investigated the impact of P4P on preventive services such as cancer screening and immunizations. Most reviews rely on the same core set of relatively dated studies conducted in US primary care settings. Dudley et al. found that among 10 dependent variables studied in eight RCTs, six showed a significant relationship with the incentive [24]. For example, one RCT found no difference between intervention and

Table 1
General characteristics of systematic reviews of the literature on effects of P4P.

Reference	Score[a]	Search period	Studies	Type of studies[b]	# of studies per level	Countries[c]	Sector[d]	Evidence on[e]
Alshamsan et al. (2010) [14]	93%	1980-November 2008	22	UBA(5), UA/CS(17)	Level III: 5 Level IV: 17	UK(21), US (1)	Mostly PC (QOF)	I, DF
Armour & Pitts (2003) [37]	73%	1966-December 2001	6	RCT(2), TS(1), UA/CS(3)	Level I: 2 Level II: 1 Level IV: 3	US	PC (5), H (1)	E, CE, UC, DF
Briesacher et al. (2009) [28]	87%	1980-August 2007	4	RCT(1), UA/CS(3)	Level I: 1 Level IV: 3	US	NH	E, CE
Chaix-Couturier et al. (2000) [38]	87%	1993-May 1999	2	RCT	Level 1: 2	US	PC	E, NFI, DF
Christianson et al. (2007) [57]	87%	1988-June 2007	44	R(7), RCT(7), QE(4), CBA(6), TS(2), UBA(4), CA(1), UA/CS(11), Q(2)	Level I: 14 Level II: 12 Level III: 5 Level IV: 11 Others: 2	US(27), UK(7), SP(1), AU(1), TW(1), NA(7)	H(6), NH(1), PC(36), PC + SC(1)	E, CE
Christianson et al. (2008) [12]	87%	-August 2007	27	RCT(2), QE(4), CBA(4), TS(1), UBA(4), CA(1), UA/CS(10), Q(1)	Level I: 2 Level II: 9 Level III: 5 Level IV: 10 Others: 1	US(18), UK(7), AU(1), SP(1)	H(6), PC(20), PC/SC(1)	E, CE, UC, I, DF
Dudley et al. (2004) [24]	93%	1980–2003	8	RCT	Level 1: 8	US	PC (8)	E, DF
Eldridge & Palmer (2009) [58]	60%	1990–2008	27	QE(1), CA(1), UA/CS(25)	Level II: 1 Level III: 1 Level IV: 25	8 Dev. countries	Not reported	E
Emmert et al. (2011) [7]	93%	2000-April 2010	9	RCT(3), CBA(3), UBA(3)	Level I: 3 Level II: 3 Level III: 3	US (8), TW(1)	H (5), PC (4), NH (1)	CE, NFI, DF
Frølich et al. (2007) [59]	93%	1980-June 2005	8	RCT	Level I: 8	US	Not reported	E, DF
Giuffrida et al. (2000) [60]	100%	1966-October 1997	2	RCT, TS	Level I: 1 Level II: 1	US, UK	PC	E, CE
Kane et al. (2004) [35]	93%	1966-October 2002	9	RCT(6), TS(1), UBA(2)	Level I: 6 Level II: 1 Level III: 1	US (8), UK (1)	Prevention	E, CE, DF
Mehrotra et al. (2009) [33]	87%	1996-June 2007	8	QE(2), CA(1), UA/CS(4), Q(1)	Level II: 2 Level III: 1 Level IV: 4 Others: 1	US	H	E, CE, UC, NFI
Petersen et al. (2006) [34]	100%	1980-November 2005	17	RCT(9), CBA(4), UA/CS(4)	Level I: 9 Level II: 4 Level IV: 4	Mainly US	Mainly PC	E, CE, UC
Rosenthal & Frank (2006) [26]	73%	-Late 2003	6	RCT(4), QE(1), UBA(1)	Level I: 4 Level II: 1 Level III: 1	US	PC	E, UC, NFI
Sabatino et al. (2008) [31]	80%	-September 2004	3	RCT, QE, UBA	Level I: 1 Level II: 1 Level III: 1	US	Prevention (cancer)	E
Schatz (2008) [27]	67%	2006–2007	22	RCT(7), CBA(6), UBA(7), UA/CS(2)	Level I: 7 Level II: 6 Level III: 7 Level IV: 2	US (19), UK (3)	Ambulatory care	E, UC, NFI, DF
Scott et al. (2011) [36]	93%	2000-August 2009	6	RCT(3), QE(1), TS(2)	Level I: 3 Level II: 3	US (5), GER (1)	PC	E, UC

Table 1 (*Continued*)

Reference	Score[a]	Search period	Studies	Type of studies[b]	# of studies per level	Countries[c]	Sector[d]	Evidence on[e]
Sorbero et al. (2006) [29]	73%	1995-April 2006	15	RCT(7), QE(2), UBA(6)	Level I: 7 Level II: 2 Level III: 6	US	PC (physicians)	E, NFI, DF
Steel & Willems (2010) [32]	78%	-January 2010	34	TS(4), UBA(8), CS/UA(17), M(1), Q(4)	Level II: 4 Level III: 8 Level IV: 17 Others: 5	UK	PC (QOF)	E, CE, UC, I
Town et al. (2005) [25]	73%	1966–2002	6	RCT	Level I: 6	US	PC (prevention)	E, CE, NFI, DF
van Herck et al. (2010) [30]	100%	1990-July 2009	128	RCT(10), QE(4), CBA(17), TS(6), UBA(30), UA/CS(57), M(4)	Level I: 10 Level II: 27 Level III: 30 Level IV: 57 Others: 4	US(63), UK(57), IT(1), SP(2), AG(1), AU(2), GM(2)	PC(98), H(17), H/PC(13)	E, CE, UC, I, NFI, DF

[a] Represents the total methodological quality score. See supplementary material for scores on individual items.
[b] R, review; RCT, randomized controlled trial; QE, quasi-experiment; CBA, controlled before-after study; UBA, uncontrolled before-after study; TS, time series with before-after data; CA, controlled-after study; UA/CS, uncontrolled-after study/cross-sectional survey; M, modeling study; Q, qualitative study.
[c] AG, Argentina; AU, Australia; GM, Germany; IT, Italy; NA, not applicable; SP, Spain; TW, Taiwan; UK, United Kingdom; US, United States.
[d] PC, primary care; H, hospital; HP, health plan; MG, medical group; NH, nursing home; IC, intensive care; QOF, quality and outcomes framework.
[e] E, effectiveness; CE, cost-effectiveness; UC, unintended consequences; I, inequalities; NFI, non-financial incentives; DF, design features.

control groups in cancer screening rates after 18 months, while another found that relatively small payments improved immunization rates by four percentage points. Overall, the effect size among the positive studies was moderate at best. Town et al., focusing on prevention, classified only one of eight outcomes as significantly improved [25]. They classified two studies as ineffective that found that increased immunization rates were largely due to better documentation, whereas Dudley et al. classified them as effective. Nonetheless, all authors (including also [26,27]) essentially reached the same conclusion: results are mixed and inconclusive and there is insufficient evidence to support the use of P4P to improve the quality of preventive and chronic care in primary care. Another review, focusing exclusively on nursing home care, identified an RCT (published in 1992) that found small beneficial effects on access and quality [28].

Most *non-randomized studies* showed improvement in selected quality measures. P4P appears to have had a small positive impact on the quality of care for diabetes and asthma, but not for heart disease [12,29]. Schatz reached a similar conclusion [27]: among 15 studies (6 level II, 7 level III, 2 level IV), 10 found positive and four found mixed results. More positive results were found among level III/IV studies than among level II studies. The most comprehensive review was conducted by van Herck et al., who identified 111 studies [30]. Of these, 30 reported an effect size, which ranged from negative to absent to (very) positive. The three studies finding negative effects also found

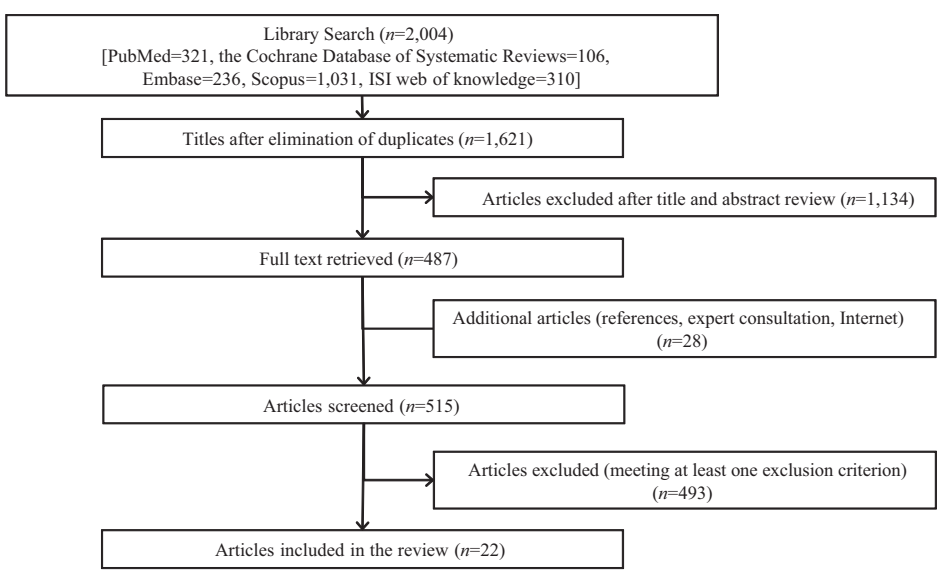

Fig. 1. Search flow and results.

Table 2
Key findings of identified systematic reviews of the literature on effects of P4P.

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Alshamsan et al. (2010) [14]				Some evidence that P4P reduces inequalities among socioeconomic groups, but disparities persisted with respect to age, sex and ethnicity. Evidence on long-term effects weak.		Some evidence that low achievement in year $t-1$ leads to high achievement in year t, so using measures with low baseline performance and/or adopting a tiered series of targets may yield the largest benefits.
Armour & Pitts (2003) [37]	1 study: financial risk for referrals decreased primary care visits; risk for cost of outpatient tests reduced # of outpatient tests; bonuses/withholds for productivity did not change resource use. 1 study: large reductions in # of admissions and visits. Regarding preventive care: increases in immunization rates (1 study), but not in cancer screening rates (1 study).	2 studies: bonuses/withholds for reduced resource use may lead to reduced outpatient expenses and utilization. 1 study: reduced outpatient expenditures by 5% as a result of bonuses/withholds	1 study: physicians at risk for cost of outpatient tests substituted primary care visits for outpatient tests, which increased visits per enrolee per year by 5%.		1 study: P4P with semi-annual feedback had no effect.	Regarding quality, studies with absolute targets effective while study with relative targets ineffective. Regarding resource use, 1 study found a greater reduction in resource use when directed at individuals whose contracts include withhold than when directed at groups. 1 study: lack of association between bonuses/withholds and change in resource use may have been result of a delayed rewards. 1 study: null effect possibly a result of limited physician awareness, and limited time frame of study.
Briesacher et al. (2009) [28]	1 RCT: improved access and outcome quality. Modest or no effect found in 3 observational studies.	1 study: improved access and outcome quality against a 5% increase in cost.				
Chaix-Couturier et al. (2000) [38]	1 study: improved immunization rates. 1 study: P4P + feedback did not increase cancer screening rates.				1 study: P4P with semi-annual feedback had no effect on cancer screening rates.	Study using 2 absolute targets effective in improving immunization rates; study using 2 relative targets ineffective in improving screening rates.
Christianson et al. (2007) [57]	Evidence base for justifying and designing P4P is thin. Few significant impacts reported, and only in selected measures.	1 study: positive ROI of P4P. 1 study: cost per QALY gaining between $13,000–30,000.				
Christianson et al. (2008) [12]	Most studies found improvement in selected quality measures, but the direct effect of P4P is largely unclear due to lack of control groups and concurrent other improvement efforts.	See Christianson et al. (2007)	2 studies: no evidence of teaching to the test. 1 study: P4P did not impair GPs' intrinsic motivation. 2 studies: initial improvements may reflect better documentation.	1 study: better record keeping for oldest patients and patients in most affluent areas; improvement for women larger than for men. Still lower recording for women, older patients, and patients in more deprived areas.		2 studies: engagement of providers contributes to better results. 2 studies: lack of impact may have been due to providers being unaware of the incentives. No evidence on ideal payment size. 2 programs using targets: most dollars awarded to high-quality providers at baseline.

Dudley et al. (2004) [24]	Results are mixed and inconclusive. Among 7 studies focusing on physicians, 5 of the 9 variables showed a significant relation to the incentive while 4 showed no change.				Individuals: 5 positive and 2 null results; groups: 1 positive and 2 null results. No relation between pay size and response. 2 studies with relative targets: no effect. Among 5 studies with enhanced FFS, 4 were positive and 1 insignificant, while in the 4 studies using bonuses there were only 2 positive results.
Eldridge & Palmer (2009) [58]	Lack of evidence on the effects of any type of P4P in any low-income country setting, mostly due to the absence of control groups.				
Emmert et al. (2011) [7]	Among the 7 studies, 5 showed improved quality of care can be achieved with higher costs.	P4P can potentially be cost-effective, but results are not convincing.		Feedback and/or PR additional to P4P did not lead to better or worse results.	Weak evidence that larger payments increase (cost-)effectiveness. The 3 studies with a high payment frequency were all relatively successful.
Frølich et al. (2007) [59]	8 RCTs showed mixed results; the potential to improve quality through P4P remains unknown.				See Dudley et al. (2004)
Giuffrida et al. (2000) [60]	Target payments associated with higher immunization rates, but the increase was significant in only 1 study.	1 study: additional cost of $3 per extra immunization.			
Kane et al. (2004) [35]	Literature is scarce. 4 studies were positive, 5 had no effects. Effect size is moderate.	1 study: additional cost of $3 per extra immunization.			Effects larger for groups. No dose–response relationship. 2 studies involving relative targets and low awareness found no effects.
Mehrotra et al. (2009) [33]	Of the 8 studies, most lack a control group and the best evidence comes from one program. Evaluation of this program (3 studies) found a 2–4%-point improvement beyond improvement seen in control hospitals.	1 study found an estimated cost per QALY of $12,967–30,081, a range generally considered cost-effective. Yet this study lacked a control group and trend data.	Difference in improvement between intervention and control hospitals on excluded measures not significant. 1 measure: intervention hospitals improved more. No insight in spillovers on other unincentivized conditions.	3 positive studies: PR may have contributed to improvements.	

Table 2 (*Continued*)

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Petersen et al. (2006) [34]	5 of 6 physician-level and 7 of 9 group-level incentives found partial (5) or positive effects (2). 2 RCTs with group-level incentives found no effect. 1 of the 2 'payment-system level' studies found a positive effect on access.	1 study: a combination of incentives to improve access to nursing home care and outcomes saved an estimated $3000 per stay.	4 studies: evidence of unintended effects, including selection and improvements in documentation rather than quality.			5 of 6 physician-level and 7 of 9 group-level incentives found partial (5) or positive effects (2). 2 RCTs with group-level incentives found no effect.
Rosenthal & Frank (2006) [26]	The empirical foundations of P4P are weak. Most studies found no effect with 2 positive findings.		Although not found in the context of P4P, several studies suggest unintended consequences are possible, including gaming and selection.		1 study: no effect of feedback only and of P4P + feedback. 1 study: feedback no effect and P4P may lead to better documentation. 1 study: P4P improved process quality, but P4P + access to a patient registry and counseling had no effect.	
Sabatino et al. (2008) [31]	1 positive result (but no control group) and 2 null results; insufficient evidence to confirm effectiveness.					
Schatz (2008) [27]	RCTs: 3 null, 3 positive (2 better documentation), 1 mixed. Non-randomized studies: 10 positive, 4 mixed, 1 null. Often unclear if effects are due to P4P.		Possible positive spillover effect found in 1 study.		Very weak evidence that incentives such as feedback and PR contribute to P4P success.	Positive studies typically used larger bonuses and measures more amenable to change.
Scott et al. (2011) [36]	Modest effects for typically only 1 out of several measures. Risk of bias due to methodological limitations.		One study: no evidence of positive/negative spillovers on unincentivized aspects of performance.			
Sorbero et al. (2006) [29]	4 RCTs had mixed results, while 3 reported no effect. 2 quasi-experiments found mixed results while observational studies tend to report positive results for at least one performance aspect.				Performance monitoring can boost the P4P effect. P4P should be implemented as part of a multifaceted strategy to performance improvement.	Lack of effects may be due to small payments; weak evidence that at least 5% of revenues is required. Low awareness may have contributed to limited effect. Physician engagement, pilot testing, accurate and reliable data, ongoing evaluation, and physician support were reported as being essential.

Steel & Willems (2010) [32]	Overall achievement increased since QOF, but post-QOF performance was roughly in line with the trend predicted from pre-QOF year. For some measures there is evidence for performance slightly above the predicted trend.	Evidence of cost-effectiveness for 12 measures with direct therapeutic effect.	No evidence that excluded conditions were neglected more after QOF than before. No evidence of reduced intrinsic motivation, but reportedly less attention to patients' concerns, unincentivized care, and continuity of care.	Changes in inequalities were small, variable, and dependent on the measure, achievement before QOF, and the demographic variable. Differences among age groups attenuated for some conditions; no changes in sex-related inequalities; reduced differences between most/least deprived areas on national level but not necessarily on local levels; mixed findings for ethnicity with reductions for some measures after QOF.		
Town et al. (2005) [25]	1 of 8 outcomes showed a significant effect. 1 significant difference found for feedback + bonus compared to control group. 1 study: P4P resulted in improved documentation.	1 study: $3 per extra immunization, which was deemed cost-effective as flu vaccines have been shown to save $117 in direct medical expenses in elderly.			1 study: feedback alone group was not different from control group. No difference between feedback + bonus vs. feedback only	Neither type of payment nor type of preventive service drives lack of effect. Limited effectiveness may be due to small rewards. Complex rules for rewards are less effective.
van Herck et al. (2010) [30]	5% improvement overall, but much variation. Negative results found in 3 studies, but together with positive results on other measures. Positive effects found especially for immunizations, diabetes, asthma, and smoking cessation. P4P most often failed to improve acute care.	4 studies on cost-effectiveness: all positive, though interpretation is difficult.	Mixed evidence of teaching to the test and gaming; very few studies have addressed such effects.	No negative effect on age, ethnic, and socioeconomic inequalities. Evidence from 28 studies suggests reductions in inequalities in the quality of care across groups rather than increases.	3 non-US studies: positive results when P4P is part of a larger quality improvement strategy including e.g., PR. Evidence from US ($n = 28$) more mixed.	More improvement for process measures than for outcomes; larger effects for measures with more room for improvement; involvement of providers, exception reporting, risk adjustment, and extensive communication appear to have contributed to positive effects; provider awareness important; relative targets generally less effective than absolute targets; no dose–response relationship; programs relying on new money had more positive effects than programs using existing funds; targeting individual physicians or small teams often more effective than targeting large provider groups or hospitals.

Note: FFS, fee for service; GP, general practitioner; P4P, pay for performance; PR, public reporting, QALY, quality adjusted life year; QOF, quality and outcomes framework; RCT, randomized controlled trial; ROI, return on investment; UK, United Kingdom; US, United States.

positive results on other measures. Overall, P4P seems to have led to 5% improvement in performance, although there is much variation [30]. For example, better results were achieved for immunizations than for cancer screening [31].

One review exclusively focused on the impact of the UK Quality and Outcomes Framework (QOF) [32], a large national P4P-program that pays bonuses to primary care practices of up to 30% of their income for reaching targets on about 130 measures. Overall, results from 28 studies (4 level II, 8 level III, 15 level IV, 1 modeling) show that achievement was high in the first year (2004–5) and has increased since. Large improvements were demonstrated in the period 2005–8 especially for diabetes, but also for hypertension, heart disease, and stroke. However, in most cases the trend showed a gradual improvement with little change after the QOF was implemented. For diabetes and asthma, a small but significant above-trend increase was found; another study (level II) found both slightly lower and slightly higher achievement than that predicted by the underlying trend. In addition, most studies (all level IV) found no relationship between target achievement and outcomes such as hospital admissions and mortality.

Several reviews discuss studies that assessed the impact of P4P in hospitals [12,30,33]. van Herck et al. found that compared to primary care services, P4P has more often failed to improve acute care [30]. Mehrotra et al. provide a detailed analysis of the effects of hospital P4P-programs in the US [33]. The most rigorous evidence (2 level II, 1 level III) comes from a single program, the Hospital Quality Incentive Demonstration (HQID). This program, which ran from 2003 to 2009, incentivized 266 hospitals to perform well on 33 clinical measures (largely processes) pertaining to six conditions. Overall, a 2–4% point increase was found beyond the improvement seen in control hospitals. No impact was found on mortality, despite that for some conditions 30-day risk-adjusted mortality was included as a performance measure. Finally, three level IV studies in the US nursing home sector showed small (e.g., improved patient satisfaction) or no effects [28].

4.2. To what extent has P4P been cost-effective?

Twelve reviews provide evidence on P4P cost-effectiveness, although only six explicitly focused on it (Table 2). Emmert et al. made a distinction between full and partial economic evaluations [7]. Full economic evaluations consider both (program) costs and quality and explicitly link both to each other (e.g., by calculating cost-effectiveness ratio's). Partial evaluations may allow for inferences about cost-effectiveness if the impact is described on both costs and quality. However, results have lower significance than those of full evaluations because no clear connection is made between the two effects. Partial evaluations also include simple costs comparisons without an analysis of the impact on quality.

Emmert et al. identified three *full evaluations* (2 level I, 1 level III, all from the US), which all found improvements in quality against increases in costs. For example, one study calculated a cost per QALY gained of $13,000–30,000 for inpatient heart treatment, while another found an intervention cost of $3 per additional immunization. van Herck et al. identified one additional full evaluation (level I) that demonstrated cost-effectiveness of a P4P-program for smoking cessation in Germany, but only when combined with training for GPs and free medication for patients [30].

Regarding *partial evaluations*, two studies (level I and II) found quality improvements and cost increases. The level I study – estimating the efficiency of a P4P-program in the nursing home sector designed to improve access and patient outcomes – found that the program saved $3000 per stay, but average costs to Medicaid rose by 5%, in part due to program costs. Another study (level II) found cost savings and improved quality, while still another level II study likely demonstrated P4P inefficiency in reducing 30-day mortality for four acute care conditions in US hospitals. Two simple cost comparisons (both from the US) showed a positive financial impact. Other reviews [12,28,34,35] discuss studies that were also identified by Emmert et al. [7] and reached similar conclusions. Steel and Willems found one additional study providing evidence of cost-effectiveness for 12 measures included in the QOF [32]. Although this highlights the potential of P4P to be cost-effective, no economic evaluation of the entire QOF (as a P4P-program) has been conducted.

Based on these results, most authors conclude that P4P has the potential to be cost-effective, but that convincing evidence is lacking. Although van Herck et al. conclude that "cost-effectiveness (. . .) is confirmed by the few studies available" [30], the evidence seems not sufficient to draw this conclusion, also because studies typically fail to include an appropriate cost and/or effect range (or fail to report about it in detail).

4.3. Which unintended consequences has P4P had?

Nine reviews provide evidence on unintended consequences, including risk selection, spillover effects, gaming behavior, and effects on providers' intrinsic motivation.

Three reviews provide weak evidence from three studies that P4P could lead to *risk selection* (Table 2). However, two studies (level I and IV) were conducted in the context of PR [26]. The third study (level II) investigated a performance-based contracting system for providers of substance abuse treatment and found that the likelihood of a patient in the program being in the most severely ill group increased in the control group and decreased in the intervention group.

Spillover effects have been discussed in six reviews (9 studies in total). The findings provide a mixed picture. Four reviews [12,27,33,36] discuss the results of three evaluations of large P4P-programs for GPs and hospitals. Two studies (level II, US HQID; level II, UK QOF) found no differences in trends in unincentivized and incentivized measures; the third study (level II, US, primary care) found no change in unincentivized performance while some incentivized measures improved. Another study from the QOF (level III) found that unincentivized measures improved when they were part of a condition for which there were incentives for other measures [32]. However, performance for two unincentivized conditions was not significantly improved despite that achievement was much

lower than for the incentivized conditions (which did continue to improve). In addition, qualitative studies found that providers are often concerned about less time for holistic care, deterioration of unincentivized care, and reductions in continuity of care [32]. Finally, van Herck et al. discuss two additional studies (level II) from the QOF [30]. The first showed a positive effect on unrewarded aspects of an included condition, a deterioration of unrewarded aspects of two other included conditions, and a reduction in continuity of care. The second study, focusing on four chronic conditions, found that the effect on recording of incentivized risk factors by GPs was larger for targeted patient groups (i.e., patients with an included condition) than for untargeted groups. The study also found evidence of substantial positive spillovers onto unincentivized factors for the targeted groups (an increase of 10.9 percentage points).

Four reviews discuss findings related to *gaming behavior* (Table 2). Most of these include an early study (level I) that found that US nursing homes tended to claim they were admitting extremely disabled patients, who then 'miraculously' recovered over a short period [34]. One review discusses 'exception reporting' in the QOF [30], which allows GP practices to exclude (noncompliant) patients from performance calculations but also provides opportunities to increase income by excluding patients for inappropriate reasons. One study (level IV) found low rates of exception reporting in the first year, but it was the strongest predictor of performance; a small number of practices may have achieved high scores by excluding large numbers of patients. A follow-up study (level IV) again found little evidence of widespread gaming; there seemed to be good clinical reasons for the exception reporting rates, which were still low in the second year.

Regarding effects on providers' *intrinsic motivation* and perceived professionalism, results of five qualitative studies are summarized in two reviews [12,32]. Two studies found that P4P did not impair providers' intrinsic motivation and that it had no effect on the quality of professional life, although providers did express more support for targets aligned with professional priorities. However, three other UK studies suggest that P4P may result in a loss in autonomy and that it may undermine providers' sense of professionalism. Providers have reported concerns about "a dual agenda in consultations, with less time for holistic care, patients' concerns and non-incentivized care, and a perceived loss in continuity of care" [32].0

4.4. To what extent has P4P affected inequalities?

Four reviews provide information on the impact on inequalities (Table 2). Most studies have addressed the impact on *socioeconomic inequalities*. Alshamsan et al. identified 18 studies, most of which examined cross-sectional associations (level IV) between quality of chronic care and an 'area deprivation score' after QOF implementation [14]. Most studies found lower quality in deprived areas compared to affluent areas before or shortly after the QOF, but differences were typically small and appear to have narrowed over time. One study (level IV) investigated a long-term effect of a more limited P4P-program in the UK in the early 1990s and demonstrated that the initial widening of inequalities in cervical cancer screening coverage had almost disappeared after 5 years. However, two level III studies found that after the QOF, medical records of patients living in affluent areas were more likely to include important risk factors (e.g., smoking status) than those of patients living in deprived areas, a difference that was not evident before. Steel and Willems found indications of narrowing inequalities between the most and least deprived areas in England, but also showed that large differences remained in individual measures and that the poorest performing practices remain concentrated in the most deprived areas [32]. Summarizing results from 28 studies (mainly from the QOF) van Herck et al. conclude that the evidence points to a reduction in inequalities across socioeconomic groups rather than an increase [30].

Alshamsan et al. identified nine studies (5 level III, 4 level IV) investigating the impact of the QOF on *age, sex and ethnic inequalities* for stroke, heart disease, and diabetes [14]. Although P4P does not appear to have widened inequalities, existing inequalities have persisted; women, older patients, and those from some minority ethnic groups continued to receive lower quality of care after QOF implementation than men, younger patients, and the white British group, although some gaps attenuated. Steel and Willems had similar findings [32]. For example, both before and after the QOF higher achievement was found for men for nearly all heart disease measures and three of eight diabetes measures. An additional study from Scotland observed lower recording for women, older patients, and patients in more deprived areas after the QOF [12].

4.5. Has P4P been more successful when applied with non-financial incentives?

Five level I studies (all conducted in US primary care settings) provide information on the merits of combining P4P with *performance feedback* to providers. One RCT found no effect of combining P4P with feedback on cancer screening rates [37,38]. Another RCT showed that neither feedback alone nor 'feedback + P4P' improved childhood immunization rates [26]. Town et al. discuss the results of three additional RCTs [25]. In one trial, results from the 'P4P + feedback' group were significantly different from the control group, but not from the 'feedback only' group. In the second study, screening rates of the group that only received feedback did not differ significantly from the group receiving feedback and a $50 bonus, and the third study also could not demonstrate superiority of 'P4P + feedback' over 'feedback only'. In contrast, Schatz did find some weak evidence that feedback contributes to P4P success [27], while Sorbero et al. found that performance monitoring can have the overall effect of improving performance [29]. Finally, van Herck et al. found that overall, P4P appears to have had a large positive effect when it is part of a larger quality improvement strategy that also includes structured feedback and PR, although the evidence is not conclusive and not convincing as studies typically lack a control group [30].

Some reviews found evidence that *public reporting* can be more effective when used together with P4P. One level

II study found that US hospitals subjected to PR and P4P improved between 2.6% and 4.1% more in process quality for certain inpatient diagnoses than hospitals subjected only to PR [33]. Mehrotra et al. also identified two other studies (level II and III) assessing the impact of the HQID, which combined P4P and PR. Studies indicated a 2–4% point improvement beyond the improvement seen in controls. Although the effects of P4P and PR could not be disentangled, the authors suspect that PR contributed to these findings, perhaps even more than P4P.

4.6. Which specific design features have contributed to desired effects?

Several design features seem important in reaching desired effects, although no studies have investigated their effect directly. These features relate to the type of measures, targeted entity, type and number of targets, type and size of the incentive, payment frequency, and provider engagement. Regarding *type of performance measures*, two reviews conclude that P4P will be more effective if desired behaviors are very specific and easy to track, and that complex rules for determining rewards are less effective [25,35]. Schatz adds to this by finding that the use of measures that are amenable to change was associated with positive results of five studies [27]. Larger effects were found for process measures than for outcomes, as well as for measures with more room for improvement [30]. Finally, results suggest that accurate/reliable data and adequate risk adjustment are vital and contribute to positive effects [29,30].

Regarding the *targeted entity*, the results suggest that P4P may be more effective when directed at individuals or small teams than when directed at (large) groups. Armour and Pitts found an early study (level IV) in which incentives directed at individual physicians had greater impact on resource use in HMOs than when directed at groups of physicians, which may have been a result of a greater incentive for individuals to use resources prudently because the risk is not shared [37]. In addition, Dudley et al. (only including RCTs) found five positive and two null results among studies in which the target was individuals and one positive and two null findings among studies in which the target was a group [24]. Furthermore, Petersen et al., identifying evidence mainly from US primary care settings, show that five of the six physician-level studies (2 level I, 1 level II, 2 level IV) found positive effects while seven of nine group-level studies either found partial (five: 1 level I, 2 level II, 2 level IV) or positive effects (two: level I). Two institutional-level studies (level I) found no effect [34]. Finally, van Herck et al. found that programs targeting individuals or small teams were often more effective than programs targeting large groups or hospitals [30].

Regarding *targets*, studies tend to find more positive effects when absolute targets are used rather than relative targets. For example, Armour and Pitts found that the two RCTs evaluating programs with absolute targets both found a positive impact while the RCT using relative targets found no effect [37]. Dudley et al. had a similar result [24]: the two studies with relative targets found no effect, while four of five studies in which absolute performance was rewarded found positive results. A more recent review also found programs using absolute targets to be more effective, although the relationship is not straightforward, in part due to the limited number of studies evaluating relative targets [30]. The number of targets also seems relevant. Alshamsan et al. found strong negative associations between scores in the previous year and improvement under the QOF [14]. This suggests that adopting a tiered series of targets, as in the QOF, may contribute to positive effects. Only using high targets may not motivate low performers, which may result in most rewards being awarded to providers already performing well before P4P [12].

Regarding *type and size of the incentive*, very little evidence is available on the relative effectiveness of bonuses and penalties. The only evidence is provided by van Herck et al., who found that programs based on "new money" seem to have generated more positive effects than programs that relied on reallocation of existing funds [30]. Regarding incentive size, Christianson et al. only found one study (level II, US Medicaid) showing that health plans that saw the largest improvements in the timeliness of well-baby care paid the largest rewards [12]. Others also failed to find a consistent 'dose–response' relationship [24,30,35]. Three reviews speculate that the limited effects may have been due to small rewards [25,27,29].

Regarding *payment frequency*, Emmert et al. found that programs in which there was little delay between care delivery and payment were all relatively successful, although the relationship is not straightforward [7]. A level I study, comparing the effect of quarterly versus annual payments for individual primary care physicians in a multispecialty group practice in California for nine preventive and chronic care measures, found no difference between the trial arms, but this may (also) have been a result of the small rewards; performance did not improve in both arms [30].

Finally, regarding *provider engagement*, better results have been achieved in programs designed collaboratively with providers (e.g., providers are involved in the selection/definition of performance measures and targets) and in which there was direct and extensive communication with providers regarding performance measurement and distribution of rewards [12,30]. In several studies that failed to find an effect of P4P (largely level I and II), many providers were actually unaware of the incentives [12,29].

5. Discussion

This paper provides an overview of the empirical literature on effects of P4P, as identified by 22 systematic reviews. Our aim was to synthesize the available (but fragmented) evidence captured by published reviews, and to structure the results according to six substantive domains. Regarding *effectiveness*, most studies have focused on prevention and chronic care provision in primary care. Results of the few studies with strong designs are mixed, justifying the conclusion that there is insufficient evidence to support or not support the use of P4P. Non-randomized studies have typically found improvements in at least one measure, although results from studies with relatively strong designs (level II) were generally less positive than results

from studies with weaker designs (levels III and IV). Overall, the impact of physician P4P has been estimated at 5% improvement in incentivized performance measures. The reviews further highlight P4P's potential to be *cost-effective*. Yet most studies use narrow cost and effect ranges. In addition, the evidence largely pertains to relatively small programs. Two recent articles not included in the reviews (level III and II) provide additional evidence that P4P can potentially be cost-effective. Walker et al. found that QOF payments were potentially a cost-effective use of resources for most GPs for most of the nine evaluated measures, but QOF administration costs, which are substantial, were not taken into account [39]. Cheng et al. examined the long-term effects of a national program for diabetes in Taiwan and found that compared to controls, P4P patients received more diabetes-specific exams/test and had fewer hospitalizations [40]. Although total costs were higher in year 1, continuously enrolled patients spent significantly less than controls in subsequent years.

Regarding *unintended consequences*, the reviews identified one study finding evidence of risk selection. Several other studies provide additional evidence. A qualitative study from California found that the inability to exception report led some physicians to deter noncompliant patients [41]. In addition, Wang et al. (level II) found that physicians referred more severely ill patients to higher-cost facilities under a performance-based incentive system in rural China [42], and Chen et al. (level III) showed that older patients and patients with greater disease severity/comorbidity were more likely to not be included in the diabetes P4P-program in Taiwan than younger and healthier patients [43]. Chang et al. (level II) had a similar finding [44]. There is some evidence of (negative) spillover effects, with some studies finding reductions in continuity of care and less improvement for excluded conditions than for included conditions. Two recent studies (level II and III) back this finding: Campbell et al. found a reduction in continuity of care after QOF implementation [45] and Doran et al. found that although incentivized and unincentivized aspects improved, improvements associated with financial incentives seem to have been achieved at the expense of small detrimental effects on unincentivized measures [46]. Evidence on gaming behavior and negative effects on providers' intrinsic motivation is virtually absent, although a recent study (level III) revealed that GPs in the UK probably gamed the system of exception reporting to some extent [47].

Although many *inequalities* in chronic disease management have not been addressed in the literature and the long-term effect on inequalities remains largely unknown, P4P seems to have narrowed socioeconomic inequalities in the UK (no evidence is available for other countries). A study by Doran et al. (level III) confirms this finding [48]: inequalities in age, sex, and ethnicity have largely persisted, although there were small attenuations for some measures. Lee et al. (level II) had a similar result by showing that the QOF was associated with a decrease in inequalities in some outcomes between ethnic groups, but that clinically important inequalities have persisted [49].

The evidence on the extent to which *non-financial incentives* can enhance the P4P effect is limited. There is some evidence that feedback alone improves performance, and that P4P does not add much when feedback is already provided. Conversely, while PR alone can stimulate quality improvement activity in hospitals [50], findings from the HQID in the US indicate that more favorable results can be achieved when P4P is added to PR. However, this only seems to hold for the short-term impact on process quality. A recent study (level II) on the long-term effect of the HQID showed that participation in the program was not associated with larger declines in mortality than those reported for PR-only hospitals [51].

Results further highlight the importance of *program design*. Although the evidence is only suggestive, several patterns emerged. P4P seems to have been more effective when:

- measures are used that have more room for improvement and are easy to track;
- directed at individual physicians or small groups;
- rewards are based on providers' absolute performance;
- the program is designed collaboratively with providers;
- larger payments are used. This is underscored by a recent US study that found that an increase in payments triggered an increase in behavioral response (level II) [52].

We are aware of one other overview of systematic reviews examining the effects of financial incentives, which was published by Flodgren et al. in 2011 [61]. There are several important differences with our review. First, Flodgren et al. searched for reviews published until January 2010, while we searched until June 2011. Two additional reviews were published between these two dates [7,32]. Second, Flodgren et al. used other inclusion criteria, resulting in only four included reviews. In addition to a different search period, this large difference with our review can be explained by the fact that Flodgren et al. required that reviews reported numerical data on outcomes, which was not required in our review. An important consequence of this requirement, however, is that several reviews that included studies investigating other effects besides impact on incentivized behaviors (e.g., cost-effectiveness, unintended consequences, impact on inequalities) were excluded. Although these reviews indeed do not consistently report numerical data, they do provide relevant information on other P4P effects for which evidence is scarce already. Another explanation for the difference in the number of included reviews is that, judging from their search strategy, Flodgren et al. did not specifically aim to identify reviews investigating the effect of financial incentives for institutions, leading them to miss the Mehrotra et al. review on P4P in the hospital setting [33] and the Briesacher et al. review on P4P for nursing homes [28] (both are not on their list of excluded reviews). Of the four reviews included by Flodgren et al., we excluded three because they did not contain studies on explicit financial incentives or were entirely overlapped by another review that provides more details. Regarding the remaining review that was included in both overviews [34], Flodgren et al. reached a similar conclusion as we did.

There are some limitations associated with our review. First, although evidence is available on a wide variety of

effects, most domains are only partially covered due to a limited number of studies with strong designs (e.g., cost-effectiveness) or a concentration of studies on a single program (e.g., effectiveness of hospital P4P and the impact on inequalities). In addition, for some domains (especially unintended consequences and design features) relevant evidence has probably been missed because no review has explicitly focused on identifying studies investigating effects for those domains. For these domains, strong conclusions are therefore not possible. Second, the included reviews lack important information on the context in which studies were conducted, such as the base payment system (e.g., P4P payments may be smaller under capitation than under FFS because of lower opportunity costs of improving performance), essential infrastructure (e.g., data collection systems), and health system features. Regarding the latter, the UK QOF (implemented in a single-payer system) appears to have generated more positive results than the more fragmented P4P initiatives in the US, but it remains unclear if this is a result of differences in the organization of care purchasing (the competitive nature of the US healthcare system and the resulting overlap in provider networks may result in conflicting incentives for providers) or of other factors such as the much larger potential bonuses that can be earned under the QOF compared to the typical P4P-program in the US. Third, research on the effects of P4P continues to be concentrated in the US and the UK. Although an increasing number of studies from other countries have been published in the last 5–10 years, it is difficult to generalize our findings to other high-income countries or any low- or middle-income country. Finally, we did not systematically verify the information reported in the reviews by consulting individual studies, which may have introduced bias (e.g., resulting from inaccurate reporting of findings from individual studies within reviews). However, because of the considerable overlap among reviews, we were able to check for potentially inaccurate representations of the evidence by comparing review authors' reports and interpretations. We encountered virtually no conflicting reports and interpretations, so the reviews' representation of the evidence is likely to be sufficiently adequate and the bias arising from our approach limited.

5.1. Implications for research and policy

Notwithstanding these limitations, our findings have several implications. First, although many studies have found improvements in selected quality measures and suggested that P4P can potentially be effective, at this point the evidence seems insufficient to recommend widespread implementation of P4P. Convincing evidence is still lacking (especially for inpatient care), despite the fact that P4P has been widely applied for many years now. In part, this lack of evidence may result from the fact that it is difficult to assess the impact of the financial part of "real-world" P4P-programs. Financial incentives are often introduced simultaneously with other improvement initiatives (e.g., non-financial incentives like public reporting) and thus as only one component of an improvement strategy. In many cases, the objective is solely to improve performance, not to test the impact of financial incentives per se. However, to facilitate evidence-based policy-making on P4P, it is crucially important that improvement strategies are implemented in the context of rigorous evaluation, using convincing control groups to disentangle the effects of the different components. This would also provide more insight in the relative merits of P4P and non-financial incentives; although different types of incentives have shown to be potentially effective when used in isolation, the literature remains almost silent on if and how they should be used together.

Second, thus far P4P evaluations have mainly focused on testing the short-term impact on clinical processes (e.g., screening for cancer, periodically performing eye exams for diabetes patients) and, to a lesser extent, intermediate outcomes (e.g., HbA1c levels of diabetes patients). However, the ultimate goal of P4P will typically be to improve patient health outcomes in the long run. Therefore, future evaluations should also assess the long-term impact on health outcomes such as complication rates, hospital readmission rates, mortality, and quality of life. Valuable information will likely become available in the coming years. In the US, the Center for Medicare and Medicaid Services is currently employing a large national P4P-program for hospitals, which will be thoroughly evaluated [53]. In addition, a large hospital P4P-program in England is currently being evaluated over a 5-year period [54]. These evaluations also include assessments of patient health outcomes and costs (including the costs of program administration), which is urgently needed given the limited data that are available on P4P cost-effectiveness.

Third, although evidence is limited, P4P may have several unintended effects, underscoring the importance of ongoing monitoring and more insight in how specific design features may help in mitigating incentives for undesired behavior. We still know very little about the appropriate amount and mix of performance measures that would minimize the risk of providers focusing disproportionately on incentivized performance. Similarly, although risk-adjustment methods for health outcomes have become increasingly sophisticated, there is still a lot to learn about how they can be applied transparently; a specific method may be very effective in leveling the playing field, but incentives for selection will persist if providers perceive it as a black box and therefore reject to support it. Furthermore, 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 defining, 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 [8]. In this respect, it is also important that program evaluations include qualitative studies to monitor the impact on providers' intrinsic motivation. More generally, insight is required in which design features contribute to desired effects. Our results indicate that program design matters, yet few studies have specifically addressed design features (e.g., the effect of varying the size of the incentive while holding other factors constant). Research is necessary to confirm our findings and to assess their influence in various contexts. In this respect,

it is critically important that studies consistently report information on the specific setting in which the program was implemented and the study was conducted.

Fourth, although it is reassuring that P4P does not seem to have widened inequalities, most studies relied on cross-sectional data from the UK and many inequalities have persisted. An explanation for these persisting inequalities may be that, with some notable exceptions [e.g., 16,55], most P4P-programs are not designed to address inequalities or lack important features that would enable them to reduce inequalities [13]. Rewarding improvement in performance and/or directly rewarding reductions in inequalities are good options to improve current programs. A recent evaluation of the HQID (level II) found that a change in design from rewarding only top performance to rewarding top performance, good performance, and improvement resulted in a significant redistribution of available funds toward hospitals caring for more disadvantaged patient populations, although significant gaps remained for incentive payments per discharge [56].

Finally, an important lesson is that improving performance via P4P is not straightforward. Important preconditions need to be fulfilled, including active provider engagement and support, adequate risk adjustment, a transparent information system for collecting performance data and for monitoring for undesired behavior, and a design that is tailored to the specific setting of implementation. Given that the interest in P4P worldwide is more likely to increase than decrease in the coming years, policymakers and researchers should give high priority to gaining more insight in how these and other preconditions can be fulfilled to ensure that P4P will yield as much value for money as possible.

Acknowledgments

We would like to thank Wynand van de Ven, Richard van Kleef, and two anonymous reviewers for their helpful comments on earlier drafts of this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.healthpol.2013.01.008.

References

[1] McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. New England Journal of Medicine 2003;348(26):2635–45.
[2] Grol R. Successes and failures in the implementation of evidence-based guidelines for clinical practice. Medical Care 2001;39(8 Suppl. 2):II46–54.
[3] Steel N, Maisey S, Clark A, Fleetcroft R, Howe A. Quality of clinical primary care and targeted incentive payments: observational study. British Journal of General Practice 2007;57(539):449–54.
[4] Eijkenaar F. Pay for performance in health care: an international overview of initiatives. Medical Care Research and Review 2012;69(3):251–76.
[6] Rosenthal MB. P4P: rumors of its demise may be exaggerated. American Journal of Managed Care 2007;13(5):238–9.
[7] Emmert M, Eijkenaar F, Kemter H, Esslinger AS, Schöffski O. Economic evaluation of pay-for-performance in health care: a systematic review. The European Journal of Health Economics 2012;13(6):755–67.
[8] Eijkenaar F. Key issues in the design of pay for performance programs. The European Journal of Health Economics 2013;14(1):117–31.
[9] Dranove D, Kessler D, McClellan M, Satterthwaite M. Is more information better? The effects of "report cards on health care providers. Journal of Political Economy 2003;111(3):555–88.
[10] Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. The New England Journal of Medicine 2008;359:274–84.
[11] Holmstrom B, Milgrom P. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. Journal of Law, Economics, & Organization 1991;7:24–52.
[12] Christianson JB, Leatherman S, Sutherland K. Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence. Medical Care Research and Review 2008;65(6):5S–35S.
[13] Chien AT, Chin MH, Davis AM, Casalino LP. Pay for performance, public reporting, and racial disparities in health care: how are programs being designed? Medical Care Research and Review 2007;64(Suppl. 5):283S.
[14] Alshamsan R, Majeed A, Ashworth M, Car J, Millett C. Impact of pay for performance on inequalities in health care: systematic review. Journal of Health Services Research & Policy 2010;15(3):178–84.
[15] Casalino LP, Elster A, Eisenberg A, Lewis E, Montgomery J, Ramos D. Will pay-for-performance and quality reporting affect health care disparities? Health Affairs 2007;26(3):w405–14.
[16] Blustein J, Weissman JS, Ryan AM, Doran T, Hasnain-Wynia R. Analysis raises questions on whether pay-for-performance in Medicaid can efficiently reduce racial and ethnic disparities. Health Affairs 2011;30(6):1165–75.
[17] Mehrotra A, Sorbero MES, Damberg CL. Using the lessons of behavioral economics to design more effective pay-for-performance programs. American Journal of Managed Care 2010;16(7):497–503.
[18] Conrad DA, Perry L. Quality-based financial incentives in health care: can we improve quality by paying for it? Annual Review of Public Health 2009;30:357–71.
[19] Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons, Ltd.: West Sussex, England; 2008.
[20] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden, Version 3; 2009. Available from: http://www.iqwig.de/download/IQWiG_Methoden_Version_3_0.pdf [24.11.09].
[21] Graf von der Schulenburg JM, Greiner W, Jost F, Klusen N, Kubin M, Leidl R, et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation – dritte und aktualisierte Fassung des Hannoveraner Konsens. Gesundheitsökonomie und Qualitätsmanagement 2007;12(5):285–90.
[22] Craig D, Rice S. NHS economic evaluation database handbook; 2009. Available from: http://www.york.ac.uk/inst/crd/pdf/nhseedhandb07.pdf [11.08.09].
[23] Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Vergleich von Bewertungsinstrumenten für die Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum, 1st ed. Cologne; 2010.
[24] Dudley R, Frolich A, Robinowitz D, Talavera J, Broadhead P, Luft H. Strategies to support quality-based purchasing: a review of the evidence. Technical Review 10. AHRQ publication No. 04-0057. Rockville, MD: Agency for Healthcare Research and Quality; July 2004.
[25] Town R, Kane R, Johnson P, Butler M. Economic incentives and physicians' delivery of preventive care – a systematic review. American Journal of Preventive Medicine 2005;28(2):234–40.
[26] Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? Medical Care Research and Review 2006;63(2):135–57.
[27] Schatz M. Does pay-for-performance influence the quality of care? Current Opinion in Allergy and Clinical Immunology 2008;8(3):213–21.
[28] Briesacher B, Field T, Baril J, Gurwitz J. Pay-for-performance in nursing homes. Health Care Financing Review 2009;30(3):1–13.
[29] Sorbero MES, Damberg CL, Shaw R, Teleki S, Lovejoy SL, DeChristofaro AH, et al. Assessment of pay-for-performance options for medicare physician services: final report. RAND Health, Working Paper WR-391-ASPE; 2006.
[30] van Herck P, Smedt Dde, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Services Research 2010;10:247.

[31] Sabatino S, Habarta N, Baron R, Coates R, Rimer B, Kerner J, et al. Interventions to increase recommendation and delivery of screening for breast, cervical, and colorectal cancers by healthcare providers systematic reviews of provider assessment and feedback and provider incentives. American Journal of Preventive Medicine 2008;35(Suppl. 1):S67–74.
[32] Steel N, Willems S. Research learning from the UK quality and outcomes framework: a review of existing research. Quality in Primary Care 2010;18(2):117–25.
[33] Mehrotra A, Damberg C, Sorbero M, Teleki S. Pay for performance in the hospital setting: what is the state of the evidence? American Journal of Medical Quality 2009;24(1):19–28.
[34] Petersen L, Woodard L, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? Annals of Internal Medicine 2006;145(4):265–72.
[35] Kane RL, Johnson PE, Town RJ, Butler M. Economic incentives for preventive care. Evidence Report/Technology Assessment (Summary) 2004;(101):1–7.
[36] Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database of Systematic Reviews 2011;9:CD008451.
[37] Armour BS, Pitts MM. Physician financial incentives in managed care: resource use, quality and cost implications. Disease Management and Health Outcomes 2003;11(3):139–47.
[38] Chaix-Couturier C, Durand-Zaleski I, Jolly D, Durieux P. Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. International Journal for Quality in Health Care 2000;12(2):133–42.
[39] Walker S, Mason A, Claxton K, Cookson R, Fenwick E, Fleetcroft R, et al. Value for money and the quality and outcomes framework in primary care in the UK NHS. British Journal of General Practice 2010;60(574):e213–20.
[40] Cheng S, Lee T, Chen C. A longitudinal examination of a pay-for-performance program for diabetes care: evidence from a natural experiment. Medical Care 2012;50(2):109–16.
[41] McDonald R, Roland M. Pay for performance in primary care in England and California: comparison of unintended consequences. Annals of Family Medicine 2009;7(2):121–7.
[42] Wang H, Zhang L, Yip W, Hsiao W. An experiment in payment reform for doctors in rural China reduced some unnecessary care but did not lower total costs. Health Affairs 2011;30(12):2427–36.
[43] Chen T, Chung K, Lin I, Lai M. The unintended consequence of diabetes mellitus pay-for-performance (P4P) program in Taiwan: are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? Health Services Research 2011;46(1 Pt 1):47–60.
[44] Chang R, Lin S, Aron DC. A pay-for-performance program in Taiwan improved care for some diabetes patients, but doctors may have excluded sicker ones. Health Affairs 2012;31(1):93–102.
[45] Campbell SM, Kontopantelis E, Reeves D, Valderas JM, Gaehl E, Small N, et al. Changes in patient experiences of primary care during health service reforms in England between 2003 and 2007. Annals of Family Medicine 2010;8(6):499–506.
[46] Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. BMJ 2011;342:d3590.
[47] Gravelle H, Sutton M, Ma A. Doctor behaviour under a pay for performance contract: treating, cheating and case finding? The Economic Journal: The Journal of the Royal Economic Society 2010;120(542):F129–56.
[48] Doran T, Fullwood C, Kontopantelis E, Reeves D. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. Lancet 2008;372(9640):728–36.
[49] Lee JT, Netuveli G, Majeed A, Millett C. The effects of pay for performance on disparities in stroke, hypertension, and coronary heart disease management: interrupted time series study. PLoS ONE 2011;6(12):e27236.
[50] Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. Annals of Internal Medicine 2008;148(2):111–23.
[51] Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcomes. New England Journal of Medicine 2012;366(17):1606–15.
[52] Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. Rand Journal of Economics 2010;41(1):64–91.
[53] Department of Health and Human Services. Hospital inpatient value-based purchasing program. Final rule. Federal Register 2011;88(88):26490–547.
[54] NHS North West. Advancing quality; 2012 www.advancingqualitynw.nhs.uk/index.php
[55] Balicer RD, Shadmi E, Lieberman N, Greenberg-Dotan S, Goldfracht M, Jana L, et al. Reducing health disparities: strategy planning and implementation in Israel's largest health care organization. Health Services Research 2011;46(4):1281–99.
[56] Ryan AM, Blustein J, Doran T, Michelow MD, Casalino LP. The effect of phase 2 of the premier hospital quality incentive demonstration on incentive payments to hospitals caring for disadvantaged patients. Health Services Research 2012;47(4):1418–36.
[57] Christianson JB, Leatherman S, Sutherland K. Financial incentives, healthcare providers and quality improvements: a review of the evidence. London: The Health Foundation; 2007.
[58] Eldridge C, Palmer N. Performance-based payment: some reflections on the discourse, evidence and unanswered questions. Health Policy and Planning 2009;24(3):160–6.
[59] Frølich A, Talavera JA, Broadhead P, Dudley RA. A behavioral model of clinician responses to incentives to improve quality. Health Policy 2007;80(1):179–93.
[60] Giuffrida A, Gosden T, Forland F, Kristiansen IS, Sergison M, Leese B, et al. Target payments in primary care: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews 2000;4:CD000531.
[61] Flodgren G, Eccles MP, Shepperd S, Scott A, Parmelli E, Beyer FR. An overview of reviews evaluation the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. Cochrane Database of Systematic Reviews 2011;7:CD009255.