Marc Pomp
Economische Beleidsanalyse

Pay for performance and health outcomes: a next step in Dutch health care reform?

Marc Pomp

Background paper for the Council for Public Health and Health Care

(Raad voor de Volksgezondheid en Zorg)

November 2010

Table of contents

1. Introduction

Concerns about underperforming health systems are widespread. Such concerns may take three different forms:

- Too much spending on care that is of little value (Wennberg et al. 2007).
- Too little spending on care with high value (e.g. certain types of preventive care) (e.g. Cutler 2004).
- Productive inefficiency: the cost per unit of care is much higher than necessary given the existing state of knowledge (Wennberg et al. 2007).

This paper does not address the empirical basis for these concerns about the underperformance of health systems. Rather, the focus in this paper is on a particular type of *solution* for addressing these problems (or some of them) that has recently been gaining in popularity, notable in the US and the UK. This solution is known as *pay for performance*, often abbreviated as P4P. In a recent literature review, P4P is defined as follows:

> "Pay for performance is one of the newest methods of medical compensation, combining reimbursement with quality improvement. Health care providers receive a base payment and, with the achievement of certain quality benchmarks for process measures (care provided) or outcome measures (result of patient care), providers receive certain rewards."(Greene and Nash, 2009, p. 140).

According to Glickman and Peterson (2009), P4P is widely used by private health insurers in the US:

> "… more than half of commercial health plans in the United States currently use P4P incentives in their provider contracts. Many of these programs involve joint efforts among employers, health management organizations, pharmaceutical companies, physician groups, academia, as well as for-profit and not-for-profit organizations." (p. S300).

Moreover, Medicare, the US government health insurance program for the elderly, runs a number of P4P pilot projects.

Health economist David Cutler and others have suggested that P4P could also be used as an incentive for raising efficiency, rather than (or in combination with) quality.

> "Evidence on the impact of pay-for-performance [on efficiency, MP] is mixed, reflecting the paucity of large experiments using these methods and the focus of most programs on quality improvement, not cost efficiency. The documented improvement in quality that some programs achieve suggests that cost saving are feasible, however. Overall, payment reform shows a good deal of promise as a complement to improved information. (Cutler 2010, p. 29).

Despite its popularity, some observers argue that pay for performance has not lived up to expectations. For example, Meredith Rosenthal, a researcher who has written extensively on the subject, summarizes her views on as follows:

> "Earlier this decade, pay for performance took center stage as a tactic for realigning payment with value. Payers' experiences during this period, as well as several major studies, clarified the limitations of this approach - characterized by some as putting lipstick on a pig. Both the enthusiastic adoption and somewhat lackluster early results of pay for performance have given rise to a broader payment-reform movement, with proposals and pilots emerging from a wide variety of stakeholders and policy leaders." (Rosenthal 2008).

Moreover, critics have argued that P4P will undermine the intrinsic motivation of doctors:

> "The central premise of pay-for-performance is that if you pay people to do something, they will do it more often. This premise is so intuitively obvious it is rarely questioned, but the fact is, it isn't always true. A great deal of experimental evidence from both social psychology and econometrics suggests that when an activity is largely driven by internal motivations - such as professionalism or pride in the quality of work one achieves - adding an external (e.g., financial) motivator can actually backfire, often dramatically."(Wynia 2009, p. 885).

The main question to be asked in this paper is given by the title: what do we know about the impact of P4P on health outcomes? The reason for focusing on health outcomes is that improvements in process indicators do not always translate into improvements in health outcomes. Hence, ideally one would want to link P4P to health outcomes rather than health process. However, in most P4P programs, process indicators are used (in a few cases in combination with outcome indicators). As will be explained in further detail later in the paper, the reason for using process indicators rather than outcome indicators is that good data on health outcomes are often lacking. Moreover, with outcome indicators it becomes more important to adjust for patient characteristics, which poses additional demands on data

availability. Nevertheless, there seems to be a trend (in the US) towards using more outcome indicators in P4P programs.

The paper is structured as follows. The next section presents the theoretical framework that serves as a guide for the review of the literature in the sections that follow. Section 3 presents a descriptive overview of existing P4P programs. Section 4 summarizes the empirical literature on the effect of P4P on process and outcomes. Section 5 summarizes what is known about the unintended, possibly adverse, side effects of P4P. Section 6 presents conclusions. Based on the findings from the survey, section 7 discusses the case for a P4P-experiment in the Netherlands.

2. Theoretical framework: P4P and the economic theory of incentives

2.1. Economic theory and optimal incentives for quality

An important (theoretical) literature in economics deals with the problem of designing optimal incentives for quality. This literature is known as the principal/agent literature. A central (and realistic) assumption of this literature is that the principal (the business manager or, in health care, the patient or the health insurance firm) cannot perfectly observe the level of effort of the agent (the employee or, in health care, the medical professional) for providing good quality. Therefore, the principal must reward the agent for his effort on the basis of something else than effort. This 'something else' can be the quality and volume of final output, number of hours spent on the job, customer satisfaction, or a combination of these elements. A key insight of the principal/agent literature is that the harder it is to observe quality, the more important it becomes to select the right agents, i.e. those who are intrinsically motivated to produce high quality. Translated to health care: one of the purposes of training doctors is to select candidates who are intrinsically motivated to provide high quality care. As pointed out by Petersen et al. (2006) in their literature survey of P4P:

> "It is generally accepted that professionals are motivated by the satisfaction of doing their jobs well (intrinsic motivation). Indeed, it is doubtful whether some valued-but-difficult-to-observe dimensions of quality (such as empathy or listening in the medical encounter) would be provided at all if physicians were solely interested in income. Thus, physicians have both nonmonetary (that is, personal ethics, professional norms, regulatory control, clinical uncertainty) and monetary (from the payment system) incentives, all of which affect effort."

This observation suggests that financial incentives, including P4P, cannot be the only motivator for delivering good quality in health care. Indeed, there is (at least in theory) a risk that explicit financial incentives crowd out intrinsic motivation (Wynia 2009).

2.2. Health care: multiple principals, multiple tasks

In health care and in many other settings, agents work for more than one principal. For example, a doctor's principals include the patient but also the health insurance firm and perhaps the hospital management. To complicate matters further, principals often have more than one objective. Patient objectives include good clinical outcomes, clear and friendly

communication, timely treatment etc.; the health insurer will also care about costs. Not all of these objectives are easy to measure. This type of setting is known as the multiple principal/multiple task principal agent problem (Dixit 1997). Designing optimal incentives in this setting is even more difficult than in the standard principal/agent setting with one agent, one principal and one objective. A key insight of the literature on the multiple principal/multiple task principal agent problem is that rewarding some tasks (these are the "incentivized" tasks), will lead agents to shirk on tasks that are not rewarded ("non-incentivized" tasks). In the literature on education this is known as "teaching to the test"; translated to health care an equivalent expression would be "managing to the measure" (Roland 2004). This insight is important when interpreting the empirical literature on P4P: even if P4P leads to improvements in the incentivized measures, this does not imply that P4P raises welfare. For that to be true it must also be the case that shirking on non-incentivized tasks does not undo the positive effects of P4P (and, of course, that the benefits of the program outweigh the costs).

Another key insight from this literature is that agents often have more information on the effort and skills of other agents in their firm, team or hospital than the principal(s). By rewarding groups rather than individuals, agents are motivated to use this superior knowledge by monitoring each other, resulting in better performance for the group as a whole. For example, there is evidence that multidisciplinary teams produce better results for chronic disease (Peterson et al., p. 270). However, rewarding groups rather than individuals comes at a price: the power of financial incentives is strongest if agents are rewarded for their own performance, rather than for the performance of the group or the hospital. This is because the potential for free-riding on the efforts of others reduces the incentive to improve quality.

2.3. Policy implications

The principal/agent framework assumes (realistically) that there is asymmetric information between agents and principal. This immediately suggest a policy implication: improving the information for principals, e.g. by mandatory disclosure of quality data by health providers, will make it easier for principals to give the right incentives. Hence, one role for government policy would be to improve publicly available information on quality.[1] However, the role of

[1] This assumes that the market by itself will not generate information disclosure, an assumption that is supported by empirical evidence, see Jin 2005.

the government may go further if principals fail to act in their own best interest. E.g., a health plan that pays physicians for smoking cessation counseling may not attract new customers by doing so, even if the health gains from cessation counseling would be large. Put differently, principals (in this case patients) fail to choose health plans that improve their welfare by helping them to quit smoking. In such cases there may be an additional (paternalistic) role for government policy that goes beyond improving information. For example, the government could directly pay bonuses to physicians offering smoking cessation counseling, a form of P4P.

2.4. P4P: process or outcome?

A question that is not directly addressed in the principal/agent literature, but that is relevant to the design of many incentive mechanisms, is the following: what should be the basis for rewards? More specifically, should rewards be based on final results (e.g., the percentage of patients that stopped smoking) or on the observation that certain tasks have been performed (e.g., smoking counseling)? In brief, should one reward outcomes or process? There are advantages and disadvantages to each of these options. Often, data on process are easier to collect and more highly correlated with physician effort. Outcomes (e.g. the number of patients who actually stopped smoking) will not only depend on physician effort also on other factors beyond the control of the medical professional, such as socio-economic background, while process is within the physician's control. However, process indicators are not always closely correlated with outcomes. Mark Chassin (2006), an American internist and president of The Joint Commission, a standards-setting and accrediting body in US health care, argues that the best way out of this dilemma is to base payment on those process indicators that are closely correlated to outcomes:

> "In my opinion, the overriding principle that should govern these choices is to maximize the likelihood that improving performance on the measure will lead directly and substantially to improved health. It follows from this principle that some measures will be more appropriate than others for inclusion in payment incentive programs. For example, highly valid process measures are much more appropriate than outcome measures. The latter are problematic for providers to employ in improvement efforts because outcomes by themselves do not contain specific information about what providers must change in order to affect the outcomes. In fact, we have very few data that establish the relationship between specific processes that may be employed by

> doctors or hospitals and outcomes that might be used as quality measures. Furthermore, virtually all outcome measures must be carefully risk-adjusted, using detailed clinical data to produce meaningful comparisons among providers. This requirement substantially increases the cost and operational complexity of using such measures in payment incentive programs. Thus, highly valid process measures will be much more appropriate for incentive payment programs. To be a valid measure of quality, a process must have a proven relationship to outcomes that are important to patients. The best of these measures will assess processes that are as close as possible to the end of the causal pathway from process to outcome. Some commonly used quality measures fail this test. For example, dilated eye examinations in diabetic patients should be part of high-quality care. However, physicians must take many additional steps after an eye examination in order to achieve improved outcomes. By itself, the eye examination does not guarantee improved outcomes. Similarly, pap smear screening for cervical cancer or mammography screening for breast cancer are early links in long chains of care processes that lead to improved outcomes only if all the links in the chain hold firm. In contrast, prescribing aspirin to heart attack survivors or angiotensin converting enzyme (ACE) inhibitors to patients with heart failure requires only that the patient actually take the medication to achieve improved outcomes. Payment incentive programs should select their quality measures with this consideration in mind to maximize the likelihood that improvement on these measures will actually lead to improved health outcomes." (Chassin 2006, p. 123S).

As will become clear from the review of literature below, nearly all P4P existing programs do indeed focus (mainly) on process indicators, although more recently there has been a trend towards using (intermediate) outcomes indicators (see section 3.4). Moreover, there have been some attempts by researchers to estimate the effect on outcomes of improvements in process. These attempts will be summarized in section 5.5, which deals with the evidence on the cost effectiveness of existing P4P programs.

3. Existing P4P-programmes: why, which, how (much)?

This section provides a brief descriptive overview of existing P4P-programmes, including a few programs that have been discontinued. The focus is on the US and the UK, since these are the only countries where P4P has been introduced on a substantial scale. The following questions will lead the overview:

- What was the reason for introducing P4P, and which actors initiated P4P?
- Which conditions of behaviors do existing P4P programs target?
- How large are existing P4P programs (in terms of patients and money)?

3.1. A very brief history of P4P in health care

In the US, the first P4P programs in health care were implemented in the private sector in the late 1990s (Tanenbaum, 2009, p. 719-720). In part as a response to two influential reports by the Institute of Medicine, *To Err Is Human* and *Crossing the Quality Chasm*, large employers and insurers sought to address patient safety concerns by adding elements of P4P to their provider payment systems. Business coalitions, such as the Leapfrog Group, were responsible for large-scale P4P programs.[2]

Medicare, the government health insurance program for the elderly in the US, has also initiated a number of P4P-programs, although the sums involved are still small (about 4 mln US dollar in 2008, compared to overall Medicare spending of 400 bln US dollar). The best known Medicare program is the CMS/Premier Hospital Quality Incentive Demonstration (HQID) project, initiated by the Centers for Medicare and Medicaid Services (CMS) with Premier, Inc., a nonprofit hospital alliance. The project started in the early 2000s as a large three-year demonstration project known as the CMS/Premier Hospital Quality Incentive Demonstration (HQID) project, Under the terms of the demonstration, 270 participating hospitals reported data on performance indicators in five clinical areas with special importance to older people, for example heart failure. Hospitals were ranked annually for three years and Medicare paid those in the top 10 percent a 2-percent bonus in addition to the standard DRG payment amount. Those in the next 10 percent were paid a 1 percent bonus. In

[2] The Leapfrog Group maintains a compendium of ongoing P4P programs in the US, see http://www.leapfroggroup.org/compendium.

the third year of the project, hospitals scoring in the lowest 20 percent were subject to equivalent payment reductions. The results of the CMS/Premier demonstration project have been analyzed in a number of research papers, summarized in section 4 below.

In the UK, the government introduced a pay-for-performance scheme in 2004. Payment was based on 136 indicators for family practices. The program is known as the Quality and Outcomes Framework (QOF). Practices of family practitioners (groups of, typically, one to six physicians) entered into a contract with the government that will provide additional payments for high quality care in excess of £1 billion - more than 20 percent of the previous family practice budget. Roland (2004) points out that the scale of the change that came about was possible only because in 2000 the government of the United Kingdom decided to provide a substantial increase in health care funding.

3.2. Which conditions or behaviors do existing P4P programs target?

The survey of the empirical literature by Greene and Nash (2009) mentions the following conditions and behaviors that are targeted by existing P4P projects: diabetes, smoking cessation, asthma/copd, depression, hypertension, acute myocardial infarction, pneumonia, substance abuse, acute sinusitis, mammography, secondary prevention of cardiovascular disease, well-baby care. Of these conditions, diabetes is mentioned most often.

In most cases, indicators focus on process (including adherence to protocol) and structure (mainly ICT). Health outcomes were infrequently mentioned. Existing P4P programs differ in many dimensions, including the indicators used for measuring performance, the size of payment, the frequency of payment, the criteria for receiving payment (e.g., absolute performance or relative performance, level or improvement from previous year, thresholds that have to reached, and the use of risk-adjustment.

3.3. How large are existing P4P programs?

P4P programs of private health insurers (HMOs) in the US[3]

Precise data on the number of insured individuals affected by P4P are not available. However, Rosenthal et al. (2006) estimated that more than half of all private health insurers, representing more than 80% of all insured Americans, use a form of P4P. In a later paper, Rosenthal et al. (2007), provide a descriptive analysis of 27 major P4P programs in the US. One of their findings is that the amount of money at stake in these P4P-programmes tends to be small, typically less than $20 per enrollee, with a few exceptions. As various observers have pointed out, this may be one reason why the effects of P4P sometimes seem to be small or even non-existent (see below).

P4P in primary care in the UK

Table 1 presents a summary of the indicators used in the QOF and the maximum number of points that can be earned in treating each of these condition. A full list of indicators, including descriptions, is available at http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_Guidance_2009_final.pdf. This is a report of over 170 pages, making it impossible to reproduce all indicators here.

Table 1. Summary of indicators included in primary care P4P in theUK.

Disease area	Indicators	Points
Asthma	7	72
Cancer	2	12
Copd	8	45
Coronary heart disease	15	121
Diabetes	18	99
Epilepsy	4	16
Hypertension	5	105
Hypothyroidism	2	8
Mental health	5	41
Stroke	10	31
Total	*76*	*550*

Source: Campbell et al. (2009)

[3] As already mentioned above public PP programs in Medicare are still small.

The indicators used in the QOF focus on the management of chronic disease, practice organization, and patients' experiences with respect to care. Only 10 indicators focus on (intermediate) outcomes (see table 2). In total the maximum number of points that can be earned on outcome indicators is about 30% of the maximum number of points that can be earned on the QOF.

Table 2 Outcome indicators in the QOF

Indicator	Maximum points
STROKE 6. The percentage of patients with a history of TIA or stroke in whom the last blood pressure reading (measured in the previous 15 months) is 150/90 or less	5
STROKE 8. The percentage of patients with TIA or stroke whose last measured total cholesterol (measured in the previous 15 months) is 5mmol/l or less	5
DM 23. The percentage of patients with diabetes in whom the last HbA1c is 7 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months	17
DM 24. The percentage of patients with diabetes in whom the last HbA1c is 8 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months	8
DM 25. The percentage of patients with diabetes in whom the last HbA1c is 9 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months	10
DM 12. The percentage of patients with diabetes in whom the last blood pressure is 145/85 or less	18
DM 17. The percentage of patients with diabetes whose last measured total cholesterol within the previous 15 months is 5mmol/l or less	6
EPILEPSY 8. The percentage of patients age 18 and over on drug treatment for epilepsy who have been seizure free for the last 12 months recorded in the previous 15 months	18
BP 5. The percentage of patients with hypertension in whom the last blood pressure (measured in the previous 9 months) is 150/90 or less	57
CKD 3 (Chronic kidney disease). The percentage of patients on the CKD register in whom the last blood pressure reading, measured in the previous 15 months, is 140/85 or less	11
Maximum points in QOF on all outcome indicators	155
Maximum points in QOF overall	550

Source: NHS QOF Guidance 2009 (available online)

An important feature of the QOF program is so-called 'exception reporting': physicians are allowed to exclude patients from the calculation of their performance score if they can point to specific circumstances, e.g. a patient refuses treatment, does not attend for review, or a medication cannot be prescribed due to a contraindication or side-effect. Exception reporting

has probably contributed to the acceptance of the program by physicians. However, according to critics exception reporting may also have reduced the public health effectiveness of population targets by shifting the focus away from harder to reach patients (Gillam 2010).

In 2006, revisions to the scheme added seven new clinical areas, including dementia and chronic kidney disease, and two new indicators of patient access to care. Payments make up approximately 25% of family practitioners' income, and 99.6% of family practitioners participated in the pay-for-performance scheme, which is voluntary.

3.4. Recent developments in P4P

In *Beyond Pay for Performance – Emerging Models of Provider-Payment Reform*, Rosenthal (2007) describes recent trends in the design of P4P-programs, or as she puts it, "enhancement of existing pay-for-performance programs". These changes involve the scope, performance measures, and magnitude of funding. These changes appear to be motivated by two perceived shortcomings of earlier P4P-programs: 1. too little impact on provider behavior and 2. not enough demonstrable benefit, including both health outcomes and spending.

- One of the changes is nonpayment for treatment of preventable complications - the mirror image of pay for performance. For example, HealthPartners in Minnesota refuses to pay for "never events" (rare and preventable errors or complications).
- Another change is a move to larger bonuses . For examples, the Prometheus Payment model uses incentives equal to 10 to 20% of the case payment rate related to clinical quality, patient experience, and cost efficiency.
- Increasingly P4P programs are including outcomes in addition to processes as performance indicators on which payment is based. Rosenthal et al. (2007) find that for physicians and medical groups, the most commonly targeted outcome measures were intermediate outcomes such as HbA1c, LDL cholesterol, and blood pressure control. For hospitals, complications and in-hospital mortality rates were frequently targeted.
- In line with this move toward rewarding outcomes, risk adjustment is increasingly used to account for the variation in the populations served by different providers.
- As the previous example makes clear, P4P-sponsors are also using cost efficiency as a performance measure, apparently motivated by a desire to use P4P (also) for controlling costs. The Medicare Physician Group Practice Demonstration program is an example of this shared savings model of payment reform. In this program, participating group

practices agree to manage the care of a population of Medicare patients with the prospect of sharing in savings that accrue to Medicare. Savings are calculated as the difference between actual spending and the risk-adjusted spending trend in a given market. Once this difference surpasses 2 percentage points, savings are shared with the integrated physician groups involved, which can receive up to 80% of these savings by performing well on cost-efficiency and quality measures.

- A final trend is to focus P4P programs not (or not exclusively) on health care providers but (also) on patients. This is called pay for performance for patients, or P4P4P. One interesting example of such a program, focusing on smoking cessation, was recently evaluated by Volpp at all (2009)[4], who randomly assigned 878 employees of a multinational company based in the United States to receive information about smoking-cessation programs (442 employees) or to receive information about programs plus financial incentives (436 employees). The financial incentives were $100 for completion of a smoking-cessation program, $250 for cessation of smoking within 6 months after study enrollment, as confirmed by a biochemical test, and $400 for abstinence for an additional 6 months after the initial cessation, as confirmed by a biochemical test. Individual participants were stratified according to work site, heavy or nonheavy smoking, and income. The primary end point was smoking cessation 9 or 12 months after enrollment, depending on whether initial cessation was reported at 3 or 6 months. Secondary end points were smoking cessation within the first 6 months after enrollment and rates of participation in and completion of smoking-cessation programs. They find that the incentive group had significantly higher rates of smoking cessation than did the information-only group 9 or 12 months after enrollment (14.7% vs. 5.0%, $P<0.001$), and 15 or 18 months after enrollment (9.4% vs. 3.6%, $P<0.001$). Incentive-group participants also had significantly higher rates of enrollment in a smoking-cessation program (15.4% vs. 5.4%, $P<0.001$), completion of a smoking-cessation program (10.8% vs. 2.5%, $P<0.001$), and smoking cessation within the first 6 months after enrollment (20.9% vs.

[4] See also Volpp et al. 2008a and 2008bfor other examples of successful of P4P4P programs (in the area of medication compliance and weight loss respectively).

11.8%, P<0.001). Thus, in this study of employees of one large company, financial incentives for smoking cessation significantly increased the rates of smoking cessation.[5]

4. Did the intended quality improvements occur?

This section summarizes the empirical literature on the effects of P4P, focusing on the performance indicators that the programs seek to improve. The next section (section 5) summarizes what is known about side effects, such as neglect of quality aspects that are not targeted by the P4P program, and negative effects on the motivation of participants.

After summarizing general conclusions from the literature, two subsections take a closer look at the experience with P4P in smoking cessation and diabetes treatment. The first of these two conditions is an interesting candidate for P4P since smoking accounts for a very large share of avoidable mortality. The second condition warrants special attention since diabetes accounts for almost all the available evidence on outcome-P4P rather than process-P4P.

In order to gather all relevant papers on the effects of P4P, the most recent of the review articles listed in Appendix A was taken as the point of departure. Next, earlier review articles were scanned for papers not mentioned in later reviews (which indeed turned out to be the case). An additional Pubmed search uncovered 3 more recent evaluations that were not included in any of these surveys. This yielded a total of 27 papers reporting on the effects of a pay for performance program.[6] Relevant details of these 27 papers, including the main conclusions on the effects of P4P, are given in Appendix B.

[5] Of course, P4P and P4P4P can be combined, in which case both the health care provider and the patient receive financial incentives on the basis of certain indicators. The P4P-project analyzed in the paper by Rosenthal et al. (2009) and summarized below is an example of such a program.

[6] Greene and Nash (2009) include in their review of the empirical literature (part B of their paper) 36 papers. However, this lists also includes other surveys (e.g. the various reviews by Rosenthal and coauthors and the review by Petersen et al.) and papers without new empirical material.

4.1. Overall conclusions reached by other reviewers

Before discussing the results from the literature, it may be worthwhile to quote earlier reviewers' own conclusions from these studies. Greene and Nash (2009) reach the following overall verdict:

> "These studies determined the efficacy of financial incentive and P4P programs to date, and most have produced positive results. Many of the references recount improved performance and better patient outcomes. All articles cited demonstrate positive results or mixed results (including no conclusion), except 2 that reveal an increased probability of health care disparities with P4P implementation. Despite these early success stories, it is paramount to ensure that programs will not exacerbate health care disparities, especially within minority and underserved populations." (Greene and Nash, 2009. p. 148).

In contrast to this rather positive judgment, Rosenthal and Frank (2007) reach a much more skeptical conclusion in their review of the literature:

> "[T]he review not only demonstrates that the current enthusiasm for pay for performance in health care rests more on conceptual than empirical foundations but also points out the key questions that need to be answered by future research in this area." (Rosenthal and Frank, 2007, p. 138). And later on in the same paper: "Overall, past experience within the health care sector and elsewhere suggests that paying for quality is unlikely to be a silver bullet. Moreover, our interpretation of the literature is that unilateral, small-scale bonus arrangements will be insufficient to motivate substantial changes on the part of physicians and hospitals. Unfortunately, these are precisely the characteristics of most recent pay-for-performance programs in the U.S. health sector (Rosenthal et al. 2004). Finally, findings related to selection, gaming, and other forms of unintended consequences are a reminder that even in health care, agents behave strategically, and pay-for-performance programs need to designed carefully to be welfare improving." (p. 153).

Peterson et al. (2006) are also rather skeptical:

> "Most physicians and hospitals are paid the same regardless of the quality of the health care they provide, producing no financial incentives for quality and, in some cases, disincentives. Thus, there is increasing enthusiasm for the idea of linking payment to performance. Despite widespread implementation, we found few informative studies of explicit financial incentives for quality. This literature review suggests some

> positive effects of financial incentives at the physician level, the provider group level, and the health care payment system level."(Peterson et al.2006, p. 269)

The differences in the overall judgments of the various reviewers may partly be due to the fact that the Greene and Nash paper included more recent papers (Rosenthal and Frank include papers published until late 2003, Greene and Nash until June 2007). However, in later articles Rosenthal remains skeptic about P4P (see Rosenthal 2008). Apparently then, the literature allows different conclusions (or perhaps, no definitive conclusions) about the effects of P4P.

4.2. Overall conclusions from this review of the empirical literature

Turning to the results of the individual papers as summarized in Appendix B, the following observations can be made:

- With a few exceptions (highlighted in **bold** in Appendix B), the P4P-programmes analyzed in these papers focus on process, not outcomes. The recent trend towards focusing on outcomes and efficiency mentioned by Rosenthal (2007) has apparently not yet resulted in evaluations of these programs.
- A few papers do report on outcomes, even though the performance indicators used in the P4P-programme focused on process. In some cases there was a significant improvement in outcomes, in other cases no improvement in outcomes was found. This is in line with the existing evidence on the imperfect correlation between process and outcomes.
- Not all of the papers are based on a comparisons of a treatment group and a control group. This makes it impossible to separate the effects of the P4P program from other trends in quality. Since in many cases there is an upward trend in quality (this follows from studies that do include a control group), the results from studies without a control group will produce an upwardly biased estimate of the effects of P4P. Some papers (notably Campbell et al. 2009) attempt to correct for this by taking into account the trend in quality improvement before the introduction of P4P. Only the improvement in excess of trend is then attributed to the program.
- Even those papers that did include a control group may report results that are not representative for the population as a whole. For example, Glickman et al. (2007) use as control group other hospitals participating in a voluntary quality improvement program (the CRUSADE program). Thus what they measure is the added effect of P4P for the group hospitals that chose to participate in the CRUSADE-program. The results may not

extend to non-participating hospitals (CRUSADE hospitals may belong to a group of hospitals that put a heavier weight on quality than other hospitals). As another example, Chen et al. (2010) take as a control group physicians that chose not to participate in the P4P program under study. This raises serious concerns about the possibility of an upward bias in their estimate of the effect of P4P.

- Most papers conclude that P4P had a favorable effect on one or more indicators, with only four exceptions: the two papers with Hillman, the paper by Fairbrother and the paper by Pearson. Rosenthal and Frank (2006) point out that the sample sizes in the two Hillman papers and in the Fairbrother paper may have been too small to detect a statistically significant effect.
- Before concluding that the bulk of the evidence offers support for the effectiveness of P4P, the methodological shortcomings – in particular the lack of a control group, or the special characteristics of the sample – of many papers should be stressed. On the other hand, the size of the estimated effects may also be biased downward due to Hawthorne effects: members of a control group who knew that they were being studied may have improved their behavior purely for that reason.

4.3. How large are the effects of P4P?

Apart from statistical significance, a relevant question is whether the effects are large in terms of the observed quality improvement. Because of the different metrics in which the papers report the observed results (% improvement, % reaching target; % of targets reached, odds ratio's), this information could not easily be included in the appendix table. The overall conclusion that emerges from the various studies is that the size of the effect differs greatly both between and within studies. For example, Lindenauer et al. (2007) find that

> "..The difference in improvement between pay-for-performance hospitals and control hospitals varied with baseline performance, ranging from 1.2% for the composite measure of care for heart failure among hospitals with the highest baseline performance to 9.6% for the same measure among hospitals with the poorest baseline performance." (Lindenauer et al. (2007), p. 490).

An interesting case in this respect is the UK QOF, since this P4P program featured by far the largest bonuses. Did these large bonuses also produce large effects? Campbell et al. (2009) summarize their findings as follows:

> "As compared with the expected level of improvement based on the pre-introduction trend, the pay- for-performance scheme was associated with an improvement in the quality of care for diabetes of 7.5 percentage points in 2005 (95% CI, 4.7 to 10.4) and 6.9 percentage points in 2007 (95% CI, 3.8 to 10.0). For asthma, the increase in quality potentially attributable to pay for performance was 9.4 percentage points in 2005 (95% CI, 3.9 to 15.0) and 5.5 percentage points in 2007 (95% CI, −1.0 to 12.1)." (Campbell et al. (2009), p. 372-3). "

If these effects can really be attributed to P4P (which is not certain given the lack of a control group), one would say that these are substantial effects.[7]

4.4. A closer look at P4P and smoking cessation

Three papers focus on P4P programs aimed at smoking cessation (Amundson et al., 2003, Roski et al., 2003, Millet et al., 2007). All three papers describe P4P programs that offer rewards to doctors for documentation of smoking status and documentation of smoking cessation counseling. Amundson et al. (2003) do not report effects in terms of numbers of smokers. Of the other two papers, Roski et al. (2003) report that although performance on these two indicators improved significantly, there was no measured effect in terms of numbers of smokers. The most recent of these three papers, the paper by Millet et al. (2007), describes the effects of the UK QOF P4P program on smoking by people with diabetes. Primary care physicians could earn 8 points out of a total 550 for offering smoking cessation advice to people with diabetes. Since the maximum bonus (paid to doctors earning all 550 points) in the overall QOF program was equivalent to about 25% additional income, documentation of smoking status of, and smoking advice to people with diabetes cannot have yielded more that 1/3 of a percent in additional income. In this sense the bonus was small. However, in the QOF smoking cessation advice was also rewarded for other conditions. Summing up over all conditions targeted by the QOF, a total of 74 points could be earned for smoking documentation and cessation advice. This could have resulted in 3.25% in additional income. The researchers report a large increase in the number of patients with documented smoking cessation advice over the period mid-2003 to late 2005, from 48.0% to 83.5%. Importantly, they also find a substantial decline in the number of smokers from 20.0% in 2003 to 16.2% in

[7] Still, one of the designers of the QOF, Martin Roland, points out that ".. the government, in retrospect, paid out more than it needed to, to achieve the levels of quality, but nobody knows how much more." (Roland 2004, p. w417).

2005. Although this decline may partly be due to a secular downward trend in the number of smokers in dependent from the P4P-program, the observed fall in this study is probably larger than would have been expected on the basis of this secular trend.

4.5. A closer look at outcome-P4P for diabetes

Of the 27 papers listed in Appendix B, no less than 11 report on P4P in the area of diabetes. Of these, only one paper (Beaulieu and Horrigan 2005) analyzes P4P-projects that reward physicians not only on the basis of process but also on the basis of outcomes. More specifically, participating doctors were rewarded on the basis of 7 process indicators (screening and testing) and 3 outcome indicators (maintaining blood pressure <130/80, HBa1c < 7,5% and LDL <100mg/dl). The combined weight of these three outcome indicators in the reward scheme was 60%, so 60% of the bonus depended on these three outcome indicators. Beaulieu and Horrigan (2005) find statistically significant improvements in the outcome indicators HbA1c and LDL, but warn that "..self-selection by physicians into the pay pilot and the small sample size [36, MP] of participating physicians limit the generalizability of the results." (p. 1318).

The paper by Chen et al. (2010b) studies a P4P-program that rewarded physicians on the basis of two process indicators (testing for Hba1c and LDL), but the paper differs from other papers in reporting on the effect of the program on hospitalization, an intermediate outcome. They find a significant decline in hospitalization as a result of the P4P program (but recall the earlier remark about the non-random control group).

5. Did P4P have unintended side-effects?

This section addresses various possible unintended side effects of P4P:

- Lower quality in non-incentivized dimensions
- A increase in socio-economic differences in quality of care
- Lack of support among health care practitioners
- Adverse changes in patient – doctor interaction

5.1. Lower quality in non-incentivized dimensions?

As pointed out in section 2, positive effects of P4P programs on the incentivized indicators do not constitute definitive evidence that the program is welfare improving. To begin with, this depends also on the costs (or rather, cost-effectiveness) of P4P (on which more below). In addition, there may be negative side-effects on non-incentivized (and perhaps unmeasured) aspects of health care. Empirical research on this issue is severely handicapped by the fact that information on these non-incentivized aspects is often lacking. One reason may simply be that data on side effects were not collected, but a more fundamental problem is that some of the relevant quality aspects are intrinsically hard to measure (see again the quote by Petersen et al in section 2). Only one of the empirical papers contains evidence on the effects on non-incentivized aspects of care for which indicators were available. This is the paper by Campbell et al. (2009) on the QOF in the UK. For asthma and heart disease, they find that mean quality scores for aspects of care that were not linked to incentives dropped between 2005 and 2007, whereas mean scores for aspects of care that were linked to incentives continued to increase. (p. 375). They also find that continuity of care (being seen by the same doctor) declined after the introduction of P4P and that this decline was statistically significant. It is worth quoting their own interpretation of this finding:

> "This study suggests that continuity of care declined after pay for performance was introduced. One possible explanation is that practices focused on meeting rapid-access targets in which access to any doctor in the practice within 48 hours was linked to incentives but access to a particular physician was not, making it more difficult for patients to see their own doctor. This could be an unintended and perverse effect of the scheme and is a concern, since continuity is an aspect of family practice that patients value. Another explanation is that there were increases in the size of practices, and many practices introduced nurse-led clinics for management of individual chronic diseases.

> Although this may have been an important part of improving the quality of care, it may have made continuity of care harder to achieve." (Campbell et al. (2009), p. 376).

In a survey of managers of health plans in the US that operate P4P-programs, Rosenthal et al. (2007) find little support for negative effects:

> "Our previous analysis had led to hypotheses about the potential for unintended consequences such as impaired access for certain patient groups and payments rewarding primarily groups that had already attained high-quality care. However, we found little support for those two hypotheses: Most sponsors (representing 81 percent of enrollees) had not observed such effects, although none had looked for them rigorously. Two respondents, however, expressed concerns about patient dumping, despite having no clear evidence of this. Representatives of two other plans also voiced concern that payments had been made to providers who were already high-performing without the realization of quality improvement. Finally, one payer noted that a provider group in its network had dissolved in a dispute over how to distribute performance bonuses." (Rosenthal et al. (2007), p. 1679).

5.2. Wider socio-economic differences in quality of care?

Various observers have pointed out that P4P may lead to larger socio-economic differences in health care quality. According to Friedberg et al., 2010, this may happen for two reasons:

> "First, if providers believe that pay-for-performance programs inadequately account for patients' characteristics, such programs may prompt providers to avoid racial and ethnic minorities and those of low socioeconomic status. Second, these vulnerable patients are predominantly seen and treated by a relatively small number of providers. If these providers receive lower performance-based payments than others do, new resources will be steered away from the care of vulnerable patient populations, potentially exacerbating health care disparities." (Friedberg et al., 2010, p. 926).

In order to assess whether this is a serious risk with existing P4P-programs, Friedberg et al. (2010) simulate the financial effects of a virtual P4P-project in primary care in which higher performers earn higher bonuses. The program is modeled along the lines of existing Medicare P4P projects. They find that practices serving higher shares of vulnerable populations would receive less per practice compared to others, by estimated amounts of more than $7,000.

However, this finding stands in sharp contrast to the finding in the UK where an unanticipated benefit of QOF has been a reduction in socio demographic inequalities in the delivery of health care (Doran et al. 2008). Apparently then, the impact of P4P on socio-economic differences in health care is ambiguous.

5.3. How popular is P4P among health care practitioners?

Obviously the probability of success of a P4P-program will be greater if the program is supported by participating physicians (and other health workers included in the program). Various researchers have conducted surveys in order to find out whether this is the case with exiting P4P-programs. The general conclusion from this literature is that physicians are sometimes (but not always) supportive of P4P-program. On the one hand, the primary care P4P program in the UK (the QOF) seems to be fairly popular with doctors (although there are critics), see e.g. Roland (2006):

> "My sense is that GPs are pretty satisfied. We've done serial surveys of GP job satisfaction and have found that external changes, particularly reorganizations, don't go down well with practitioners and that they're generally associated with a reduction in job satisfaction. We don't yet have the results of surveys carried out since the pay-for-performance scheme, but recruitment to general practice now looks as if it's quite good." (Roland (2006), p. W414).

This is in line with the findings of Whalley et al. who conducted a mailed survey of English general practitioners reporting on job satisfaction, hours worked, income, and effect of P4P in 2004 (before P4P implementation) and in 2005 (after P4P implementation). The authors found greater job satisfaction, decreased hours worked, and increased incomes. In addition, respondents reported a decrease in autonomy and an increase in administrative and clinical workload. Overall, the study found physicians to be more positive about the effect of P4P following the program's implementation.

On the other hand, P4P programs in the US often seem to be resented by doctors. In an effort to explain these differences, McDonald and Roland point to the following differences in the programs:

> "Californian physicians were more likely to express resentment about pay for performance and appeared less motivated to act on fi nancial incentives, even in the

> program with the highest rewards. The inability of Californian physicians to exclude individual patients from performance calculations caused frustration, and some physicians reported such undesirable behaviors as forced disenrollment of noncompliant patients. English physicians are assessed using data extracted from their own medical records, whereas in California assessment mostly relies on data collected by multiple third parties that may have different quality targets. Assessing performance based on these data contributes to feelings of resentment, lack of understanding, and lack of ownership reported by Californian physicians." (McDonald and Roland 2009, p. 121).

In addition, it also seems quite likely that the fact that the QOF was financed entirely out of additional funds played an important role in its acceptance (if not popularity) among doctors. P4P programs in the US are partly self-financing or zero-sum, in the sense that P4P programs where (partly) paid out of existing budgets. Moreover, the amount of money available in the UK was rather large.

It should be added that the available evidence does *not* indicate that US physicians are opposed to financial incentives per se. This follows from a survey Casalino et al. (2007) among 556 primary care physicians. They report that almost three fourths of respondents favored financial incentives for quality. Opposition to P4P has more to do with design and implementation: only 30% of all respondents agreed that current measures of quality are accurate, 88% stated that performance measures are not correctly adjusted for patients' medical conditions, 85% responded that the measures are not adequately adjusted for patient's socioeconomic status, and 82% believe that performance measures could cause physicians to avoid high-risk patients.

5.4. P4P and the patient – doctor interaction

As already pointed out in section 5.1, Campbell et al. (2009) report that in the QOF in the UK, continuity of care (being seen by the same doctor) declined after the introduction of P4P. More specifically, patient evaluations of continuity of care were 4.1 percentage points lower than expected in 2005 and 4.3 percentage points lower in 2007; these differences are statistically significant. This finding is in line with the criticism voiced by Gillam (2010):

> "The framework promotes a mechanistic approach to managing chronic disease, reducing clinical practice to a series of dichotomised decisions. Both doctors and nurses are concerned about the "box ticking culture": the intrusive impact of computerised templates that turn people into codes, to the detriment of person focused care."(Gillam 2010).

However, diminished continuity of care may have been an unavoidable side effect of the improved division of labor between doctors and nurse practitioners, that is also observable during the study period. Therefore, this effect of P4P should probably not get a very large weight in the overall cost/benefit assessment of the QOF.

5.5. Is P4P cost effective?

Evidence that P4P leads to improvements in process or outcomes without negative side-effects does not necessary imply that P4P is worth the money. For this to be the case, the cost per unit of health produced must also be acceptable. What acceptable is depends in part on the perspective taken. For example, from the perspective of a private payer (e.g. health insurer), the benefits of P4P in terms of money saved or positive publicity must ultimately be sufficient to cover the cost of the program. As pointed out earlier, the business case for private sponsors (health insurers) is not proven, as is evidenced by the trend towards inclusion of cost reduction in addition to quality as a performance indicator. But there are exceptions to this overall assessment. Parke (2007) analyzed an ophthalmologic P4P program and determined its cost of implementation for a health plan compared with baseline (ie, the cost of expenditures the year before implementation). Total expenditures decreased by about 10%. This was achieved in spite of a 10% increase in provider pricing, yielding the conclusion that P4P reduces costs for a health plan by reducing the volume of services used. Similarly, Curtin et al. (2006) estimated the ROI for a diabetes P4P program. The cost of the P4P program was $1.15 million each year. In the first 2 years of the program, the ROIs were 1.6 to 1 and 2.5 to 1, saving the HMO more than $4 million.

From a social perspective, what matters is not only money saved but also (or primarily) the amount of extra health produced (for example, measured in QALYs) per unit f money spent. In other words, from a social perspective we would like to know the cost effectiveness of P4P program. This has been studied by a number of researchers. Nahra et al. (2006) estimated

QALYs gained by patients in a P4P-program operated by an American health insurer (Blue Cross Shield of Michican) for heart-related care in 85 participating hospitals. During the 4-year period from 2000 to 2003, the insurance company paid approximately $22 million in incentive payments and administrative costs for a P4P program. The increased medical care caused a estimated gain of at least 733.3 QALY, which translates to approximately $30 000 per QALY. However, the estimated health gains is based on a simple trend analysis, without a control group. Thus, this estimate should be treated with caution.

Researchers from the universities of East Anglia and York have attempted to estimated the cost effectiveness of the QOF in UK primary care (Walker et al., 2010). They use the following approach. First, literature reviews were carried out to identify relevant evidence on the effect of each of the clinical indicators used in the QOF on mortality (both process and outcome indicators). Next, the actual improvement observed on each of the indicators included is used as an estimated of the effect of the QOF. Note that no adjustment is made for trends in improvement that took place independent of the QOF as in the study of the QOF by Campell et al. (2009). This will lead to an overestimate of the actual health gains due to the program. The authors could find sufficient evidence on only 9 indicators to determine the cost effectiveness. Of these, only 1 focuses on outcomes: "blood pressure 150/90 in patients with hypertension in the past 9 months". Third, the cost per patient treated determined on the basis of the number of points earned for this indicator times payment per point under the QOF. The authors find that all nine indicators for which sufficient evidence could be found are cost effective (cost per QALY < 20 000 pound). Needless to say, these estimates must be considered as very preliminary, given the hypothetical nature of the health gains per indicator.

The authors also report the improvement in each of the 9 indicators required for cost effectiveness. In general, the required changes are quite small (see table 4), suggesting that the QOF probably has been cost effective.

Table 3 Cost effectiveness in the UK QOF

Indicator	**% improvement needed for P4P to be cost effective using 20 000 pound as a threshold**
BP5 The percentage of patients with hypertension in whom the last blood pressure (measured in last 9 months) is 150/90 mmHg or less	0.3
CHD9 The percentage of patients with CHD with a record in the last 15 months that aspirin, an alternative antiplatelet therapy, or an effective than anticoagulant is being taken (unless a comparator) contraindication or side-effects are recorded)	8.4
CHD10 The percentage of patients with CHD who are currently treated with a beta-blocker (unless a ontraindication or side-effects are recorded)	0.06
CHD11 The percentage of patients with a history of myocardial infarction (diagnosed after 1 April 2003) who are currently treated with an ACE inhibitor	19.8
CS1 The percentage of patients aged 25–64 years (in Scotland 25–60 years) whose notes record that a cervical smear has been performed in the last 3 to 5 years	0.12
DM15 The percentage of patients with diabetes with proteinuria or micro-albuminuria who are treated with ACE inhibitors (or A2 antagonists)	0.4
DM21 The percentage of patients with diabetes who have a record of retinal screening in the previous 15 months	0.4
LVD3 The percentage of patients with a current diagnosis of heart failure due to LVD who are currently treated with an ACE inhibitor or or A2 antagonist, who can tolerate therapy and for whom there is no contraindication	4.2
Stroke12 The percentage of patients with a stroke shown to be non-haemorrhagic, or a history of TIA, who contraindication or side-effects are recorded)	0.9

Source: Walker et al. (2010)

6. Conclusions

Is there convincing evidence that P4P works?

The vast majority of empirical studies on P4P report positive effects. However, many of these studies are plagued by methodological problems, in particular a lack of randomization between 'treatment' (participation in P4P) and 'non-treatment' (non-participation, the control group). Many empirical results in the literature are based on before/after comparisons, making it impossible to separate the effect of P4P from the influence of other factors not observed by the researcher. Moreover, even where randomization is used, the participants (treatment and control) may not be representative of all health providers. Thus, the evidence on P4P, while rather favorable, is not conclusive.

Very little is known about effects of P4P on health outcomes. There are a few papers reporting positive effects in the area of smoking cessation and diabetes care, but due to small samples sizes and/or lack of an appropriate control group, the results of these papers should not be seen as conclusive evidence in favor of P4P papers. Perhaps this gap in the existing knowledge will disappear when more data on outcomes-P4P become available as a consequence of the recent trend towards P4P focusing on outcomes in the US.

As a corollary of the last conclusion, little is also known about whether P4P delivers value for money, i.e. about the cost effectiveness of P4P. However, a recent analysis of the UK QOF program showed that even with small improvements the program is already cost effective, using a cost-effectiveness threshold of 20 000 pound per QALY.

Are the effects large enough to make a substantial contribution to improved health outcomes?

The literature does contain a few papers in which the (alleged) effects of P4P on the performance indicators used are quite substantial. However, whether this would translate into substantial potential health gains at the macro-level when P4P programs are introduced on a larger scale is unclear due to the above mentioned lack of knowledge about the effect of P4P on health outcomes.

Are there unintended side-effects?

There is some evidence of negative side-effects, e.g. in the sense of diminished continuity of care (i.e. being seen by someone else than your usual primary care physician). Observers have also pointed out that payment on the basis of outcomes requires adequate risk adjustment, which implies heavy data requirements.

Who should take the lead in P4P?

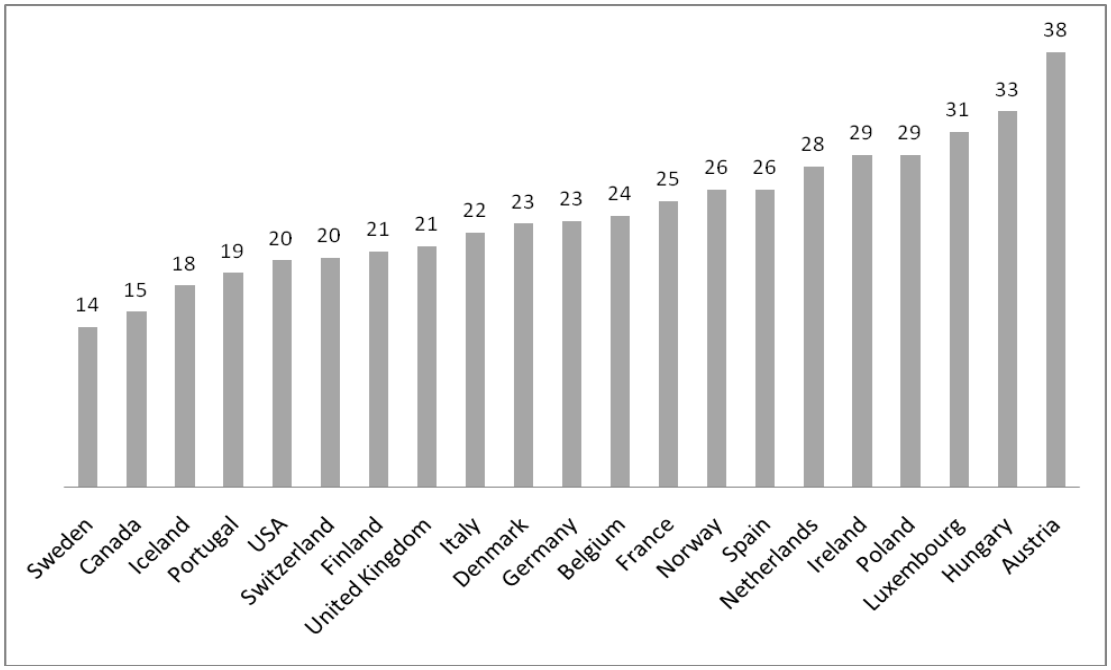
Although in the US private health insurers have taken the lead in P4P programs, the business case for private sponsors (health insurers) is not proven. This is consistent with a trend towards inclusion of cost reduction in addition to quality improvement as a performance indicator: apparently health insurers are not satisfied with the (lack of) financial returns on previous P4P programs. Moreover, at least initially, there are public good aspects of a P4P-program in the sense that such a program generates public knowledge about what works in P4P and what does not work, what are the side-effects, what is the most successful design, how large should the bonus be etc. Finally, there are likely to be positive externalities of a P4P program financed by a private health insurer, since customers of other health insurers will probably also benefit from any quality improvements that occur. These public good aspects and positive externalities imply that the private sector will (at least initially) underinvest in P4P-programs. For all these reasons, private health insurers may, from a social point of view, underinvest in P4P. Hence there is a case for government policy in financing P4P-programs, at least initially.

7. The case for a P4P-experiment in the Netherlands

7.1. Deficiencies in the Dutch health system

Although the Netherlands ranks high in terms of universal access to health services, in terms of *health behavior* there is much room for improvement. This is most strikingly the case for smoking: the prevalence of smoking is higher in the Netherlands than in most countries with comparable income levels (see figure below). Smoking accounts for roughly two years lost in life expectancy per capita; moreover, 15% of pregnant women continue smoking during pregnancy, resulting in low birth weight and higher infant mortality. Alcohol consumption deserves also attention: among Dutch adolescents drinking alcohol is much more common than among adolescents in most other countries. According to a European survey in 2007, no less than 24 percent of all Dutch children aged 16 used an alcoholic beverage ten times or more during the last 30 days, compared to 19 percent in Germany, 14 percent in the UK and 13 percent in France (Espad 2009). Also, in line with international trends obesity has reached epidemic forms in the Netherlands, with over 10 percent of the population being obese (body mass index > 30) (Bemelmans et al. 2004).

Percentage of adults smoking, 2007



Source: WHO

As is well known, health behavior is strongly correlated with education - the so called health–education gradient. Cutler and Lleras-Muney (2009) summarize the evidence on the health–education gradient as follows: "..controlling for age, gender, and parental background, better educated people are less likely to smoke, less likely to be obese, less likely to be heavy drinkers, more likely to drive safely and live in a safe house, and more likely to use preventive care." For a description of socio-economic differences in health in the Netherlands, see e.g. Mackenbach (2010), chapter 4.

Using data for the US and the UK, Cutler and Lleras-Muney (2009) find that the health–education gradient is *not* caused by differences in tastes or personality, but rather related to economic resources (income), knowledge, cognitive ability and social integration. Between them these factors explain about 60 percent of the gradient.

Given these fundamental factors underlying unhealthy behavior, it will be clear that pay for performance cannot be a silver bullet. Pay for performance will not solve the problems of low income, lack of cognitive ability or peer pressure. Moreover, pay for performance for health providers can by definition only improve the health behavior of those who use the services of health providers. People with an unhealthy lifestyle but (still) without health problems are outside the health system; therefore their behavior cannot be improved by P4P (or for that matter by any other policy that works through health providers).[8] Thus, P4P should be seen as one instrument in a much broader policy package aimed at changing unhealthy behavior. Such a broader package could include information and education, taxation (e.g. cigarettes are still relatively cheap in the Netherlands), further expansion of smoking bans, further restrictions on the sale of alcohol to adolescents and young adults, and expanded insurance coverage for smoking cessation and other behavioral interventions.

7.2. Experience with P4P in the Netherlands

In the Netherlands there has been very little experience with P4P and quality. One of the few P4P-projects that has been carried out is the so-called bonus/malus experiment for primary gatekeepers in the Tilburg region in the 1980s (van Tits, 1989). In this experiment, primary care doctors were rewarded for preventing unnecessary hospitalization and medication.

[8] Of course, this is not true if pay for performance were directly aimed at e.g. smokers, as was the case in the P4P4P program analyzed by Volpp et al. (2009).

Although the project was quite successful in achieving its objectives (reducing unnecessary hospitalization and medication), the aim of the project was not to improve quality, nor was the effect of the experiment on quality measured.

There is only one study on quality-P4P in the Netherlands (in Dutch, not peer reviewed). This is the study by Kirschner et al. (2009) on a P4P program in 2007/2008 for primary care doctors in the south of the Netherlands. In this P4P-program, participating practices (72) were rewarded on the basis of relative and absolute levels of achievement on process of care, management and organization, and patient satisfaction in the areas of diabetes, COPD, asthma, cardiovascular disease, flu vaccination and cervical screening. Practices could earn additional bonuses on the basis of absolute and relative levels of patient experience. The program was initiated by two health insurance firms. The average bonus per practice was about 7500 euro, or roughly 5% of total turnover. The study is based on a before/after comparison and did not include a control group. For the first four out of the six conditions mentioned, the researchers report significant improvements over this 2-year period. The improvements were substantial: the number of patients receiving adequate care increased by about 10 percentage points for each of these four conditions, and these changes were all statistically significant. The improvement in patient experience was almost 5% and again statistically significant. This does not constitute conclusive evidence in favor of P4P, since other quality-improvement initiatives (notably obtaining NHG-accreditation) were also in operation in these years. The lack of a control group makes it impossible to distinguish the effect of the P4P-program from the effects of these other initiatives. Nevertheless, the results of this study underscore the overall conclusion that emerges from the literature: P4P is a promising tool for improving the quality of healthcare.

7.3. Suitability of the Dutch health system for a P4P-program

From the above review of the literature it follows that the evidence on P4P is inconclusive. This is true both for the intended effects of P4P and for the unintended negative side effects. Having said this, it is also true that the available evidence is consistent with, on balance, positive effects of P4P. This is also the conclusion drawn in a recent OECD report:

> ‘However, even with limited evaluation in OECD countries, the initial results of P4P appear promising and have galvanised payers and providers to measure health care quality. There appears to be growing evidence that incentivising priority public health interventions like cancer screening works and also P4P works in getting physicians to

follow evidence based guidelines for chronic diseases like diabetes and heart failure (OECD 2010, p. 120).

Moreover, the literature suggests that the likelihood of positive effects of P4P is larger if:

- The institutional setting is favorable, in particular if the health care system is integrated rather than fragmented. Episode of care payment (in Dutch: "ketenzorg") seeks to achieve such integrated care. It follows that P4P programs are probably most successful for conditions for which episode of care payment has already been introduced, e.g. diabetes.
- P4P is financed out of additional money, since this will increase the willingness of health professionals to participate. A self-financing P4P-program is in principle possible (by simultaneously lowering base payments or by including negative incentives), but this will undermine support of health professionals for the program.
- Bonuses are large (although the size of the necessary bonus will depend on the effort needed to achieve an improvement).
- The indicators used are supported by professionals and their professional bodies, since this will increase support for the program.
- No additional data need to be collected since this will reduce the amount of additional paperwork.

In the light of these success factors, in a number of respects the Netherlands seems well suited for P4P: risk adjustment data are already being collected by health insurance firms (although it remains to be seen whether these data can be used for casemix corrections in P4P programs), and episode of care initiatives ("ketenzorg") are currently being introduced for a number of chronic diseases, including diabetes.

Whether there is a case for introducing P4P programs in the Netherlands also depends on the deficiencies in the existing health system. One condition where there are clear indications of such deficiencies in the Netherlands is smoking. Smoking cessation is therefore an interesting candidate for a P4P program, despite the fact that not all the evidence indicates that such a program will actually lead to less smoking. Another promising candidate for a P4P program is diabetes care, for which there is some favorable evidence on the effects of P4P on health outcomes such as LDL and HBa1c.

Before deciding on a national P4P program, it is advisable to start with a limited experiment. Given the (at least initial) public-good nature of P4P-programs, there is a case for government involvement in the design and financing of a P4P-experiment. Such an experiment should be designed in such a way that the effects of P4P can be measured in a meaningful way. Thus, participation in the program should be randomized and participants should be representative of the underlying populations. Of course, before implementing such an experiment many details need to be worked out, including size of the sample, size of the bonus, criteria for paying bonuses, whether or not to include bonuses paid to patients (P4P4P), choice of indicators used, randomization into treatment or control etc.

Such an experiment can be used to assess the effects of a P4P program, both on the incentivized indicators and on the non-incentivized indicators. Also, the effects on socio-economic differences in health outcomes should be monitored, given the ambiguous results on the effects of P4P in this area reported in the literature.

While this is not the place to work out the details of a possible P4P experiment, the literature does suggest that it would be a good idea to base the P4P-bonus not only on process indicators but (also) on outcome indicators. This is for the following two reasons. First, the evidence on P4P in the area of smoking cessation indicates that improvement in process (in this case, smoking cessation counseling) does not always lead to an actual increase in the number of successful quitters. Second, the evidence on P4P in the area of diabetes makes it clear that intermediate outcome indicators (HBa1c and LDL levels) can successfully be targeted by P4P.

8. References

Amundson, G., L. I. Solberg, M. Reed, E. M. Martini, and R. Carlson. 2003. Paying for quality improvement: Compliance with tobacco cessation guidelines. Joint Commission Journal on Quality and Safety 29 (2): 59-65.

Beaulieu ND, Horrigan DR. Putting smart money to work for quality improvement. *Health Serv Res.* 2005;40(5, pt 1): 1318-1334.

Bemelmans, W.J.E., R.T. Hoogenveen, T.L.S. Visscher, W.M.M. Verschuren, A.J. Schuit, Toekomstige ontwikkelingen in overgewicht: Inschatting effecten op de volksgezondheid, RIVM rapport 260301003/2004.

Bremer RW, Scholle SH, Keyser D, Houtsinger JV, Pincus HA., Pay for performance in behavioral health, Psychiatr Serv. 2008 Dec;59(12):1419-29.

Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M., Effects of pay for performance on the quality of primary care in England, N Engl J Med. 2009 Jul 23;361(4):368-78.

Campbell SM, McDonald R, Lester H., The experience of pay for performance in English family practice: a qualitative study., Ann Fam Med. 2008 May-Jun;6(3):228-34.

Casalino LP, Alexander GC, Jin L, Konetzka RT. General internists' views on pay-for-performance and public reporting of quality scores: a national survey. Health Aff (Millwood). 2007;26(2):492-499.

Chassin MR, Does paying for performance improve the quality of health care?, Med Care Res Rev. 2006 Feb;63(1 Suppl):122S-125S.

Chen JY, Kang N, Juarez DT, Hodges KA, Chung RS, Legorreta AP, Impact of a pay-for-performance program on low performing physicians, J Healthc Qual. 2010 Jan-Feb;32(1):13-21; Erratum in: J Healthc Qual. 2010 Mar-Apr;32(2):52 [2010a].

Chen JY, Tian H, Taira Juarez D, Hodges KA Jr, Brand JC, Chung RS, Legorreta AP, The effect of a PPO pay-for-performance program on patients with diabetes, Am J Manag Care. 2010 Jan 1;16(1):e11-9. [2010b].

Curtin K, Beckman H, Pankow G, Milillo Y, Green RA. Return on investment in pay for performance: a diabetes case study. J Healthc Manag 2006;51:365-6.

Cutler, D, and A. Lleras-Muney, Understanding Differences in Health Behaviors by Education, July 2009.

Cutler, D. Where Are The Health Care Entrepreneurs? The Failure of Organizational Innovation in Health Care, Harvard University and NBER, May 2010 (2010).

Cutler TW, Palmieri J, Khalsa M, Stebbins M. Evaluation of the relationship between a chronic disease care management program and California pay-for-performance diabetes care cholesterol measures in one medical group. *J Manag Care Pharm.* 2007;13(7):578-588.

Dixit A, Power of Incentives in Private versus Public Organizations, The American Economic Review, Vol. 87, No. 2, Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association (May, 1997), pp. 378-382.

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: 728-36.

ESPAD, The 2007 ESPAD Report: Substance Use Among Students in 35 European Countries, Stockholm 2009.

Fairbrother, G., K. L. Hanson, S. Friedman, and G. C. Butts. 1999. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. American Journal of Public Health 89 (2): 171-75.

Felt-Lisk S, Gimm G, Peterson S. Making pay-for-performance work in Medicaid. *Health Aff (Millwood).* 2007;26(4): w516-w527.

Gillam S, Should the Quality and Outcomes Framework be abolished? Yes, BMJ 2010;340:c2710.

Gilmore AS, Zhao Y, Kang N, et al. Patient outcomes and evidence-based medicine in a preferred provider organization setting: a six-year evaluation of a physician pay-for per formance program. *Health Serv Res.* 2007;42(6, pt 1): 2140-2159, discussion 2294-2323.

Glickman SW, Boulding W, Roos JM, et al. Alternative pay-for performance scoring methods: implications for quality improvement and patient outcomes. Med Care. 2009;47(10):1062-1068.

Glickman SW, Ou FS, DeLong ER, et al. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA.* 2007;297(21):2373-2380.

Glickman, SW, Fang-Shu Ou, MS; Elizabeth R. DeLong, PhD; Matthew T. Roe, MD, MHS; Barbara L. Lytle, MS; Jyotsna Mulgund, MS; John S. Rumsfeld, MD, PhD; W. Brian Gibler, MD; E. Magnus Ohman, MD; Kevin A. Schulman, MD; Eric D. Peterson, MD, MPH , JAMA. 2007;297:2373-2380.Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction, JAMA. 2007 Jun 6;297(21):2373-80.

Greene SE and DB Nash, Pay for Performance: An Overview of the Literature, American Journal of Medical Quality 2009; 24.

Grossbart SR. What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. *Med Care Res Rev.* 2006;63(1)(suppl):29S-48S.

Hillman, A. L., K. Ripley, N. Goldfarb, I. Nuamah, J.Weiner, and E. Lusk. 1998. Physician financial incentives and feedback: Failure to increase cancer screening in Medicaid managed care. American Journal of Public Health 88 (11): 1699-701.

Hillman, A. L., K. Ripley, N. Goldfarb, J.Weiner, I. Nuamah, and E Lusk. 1999. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. Pediatrics 104 (4): 931-35.

Jin, G.Z., 2005, Competition and disclosure incentives: an empirical study of HMOs, RAND Journal of Economics, vol. 36(1), pp. 93-112.

Kirschner K, J. Braspenning T. Gootzen, C. van Everdingen, J. Batenburg, W.Verstappen M., Klomp, R. Grol, Pay-for-Performance in de huisartsenpraktijk Een experiment in Zuid-Nederland IQ healthcare, 2009 (in Dutch).

Levin-Scherz J, DeVita N, Timbie J. Impact of pay-forperformance contracts and network registry on diabetes and asthma HEDIS measures in an integrated delivery network. *Med Care Res Rev.* 2006;63(1)(suppl):14S-28S.

Lindenauer PK, Remus D, Roman S, et al. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med.* 2007;356(5):486-496.

Mackenbach, J., Ziekte in Nederland: Gezondheid tussen Politiek en Biologie, 2010.

Mandel KE, Kotagal UR. Pay for performance alone cannot drive quality. *Arch Pediatr Adolesc Med.* 2007;161(7):650-655.

Mason A, Walker S, Claxton K, Cookson R, Fenwick E, Sculpher M. The GMS Quality and Outcomes Framework: are the quality and outcomes framework (QOF) indicators a cost-effective use of NHS resources? Joint executive summary reports to the Department of Health from the University of East Anglia and University of York, 2008.

McDonald, R and M Roland, DM, Pay for Performance in Primary Care in England and California: Comparison of Unintended Consequences, Annals of Family Medicine 7:121-127 2009.

Mehrotra A, Damberg CL, Sorbero ME, Teleki SS., Pay for performance in the hospital setting: what is the state of the evidence?, Am J Med Qual. 2009 Jan-Feb;24(1):19-28.

Millett C, Gray J, Saxena S, Netuveli G, Majeed A. Impact of a pay-for-performance incentive on support for smoking cessation and on smoking prevalence among people with diabetes. *CMAJ.* 2007;176(12):1705-1710.

OECD, Value for Money in Health Spending, Paris, 2010.

Pearson SD, Schneider EC, Kleinman KP, Coltin KL, Singer JA. The impact of pay-for-performance on health care quality in Massachusetts, 2001-2003. *Health Aff (Millwood).* 2008;27(4):1167-1176.

Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann Intern Med.* 2006;145(4):265-272.

Rodriguez, HP, MPH, Ted von Glahn, MS, Marc N. Elliott, PhD, William H. Rogers, PhD, and Dana Gelb Safran, ScD, The Effect of Performance-Based Financial Incentives on Improving Patient Care Experiences: A Statewide Evaluation, Gen Intern Med. 2009 December; 24(12): 1281–1288.

Roland M, Pay-for-performance: too much of a good thing? A conversation with Martin Roland. Interview by Robert Galvin, Health Aff (Millwood). 2006 Sep-Oct;25(5):w412-9. Epub 2006 Sep 5.

Rosenthal MB, Li Z, Robertson AD, Milstein A, Impact of financial incentives for prenatal care on birth outcomes and spending, Health Serv Res. 2009 Oct;44(5 Pt 1):1465-79. Epub 2009 Jul 13.

Rosenthal MB, Beyond pay for performance - emerging models of provider-payment reform. N Engl J Med. 2008 Sep 18;359(12):1197-1200.

Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. *JAMA.* 2005;294(14):1788-1793.

Rosenthal MB, Frank RG, What is the empirical basis for paying for quality in health care? *Med Care Res Rev.* 2006;63(2):135-157.

Rosenthal MB, Landon BE, Howitt K, Song HR, Epstein AM. Climbing up the pay-for-performance learning curve: where are the early adopters now? *Health Aff (Millwood).* 2007;26(6):1674-1682.

Rosenthal MB, Landon BE, Normand SL, Frank RG, Epstein AM., Pay for performance in commercial HMOs. N Engl J Med. 2006 Nov 2;355(18):1895-1902.

Roski, J., R. Jeddeloh, L. An, H. Lando, P. Hannan, C. Hall, and S. H. Zhu. 2003. The impact of financial incentives and a patient registry on preventive care quality: Increasing provider adherence to evidence-based smoking cessation practice guidelines. Preventive Medicine 36 (3): 291-299.

Shepard DS, Calabro JA, Love CT, McKay JR, Tetreault J, Yeom HS. Counselor incentives to improve client retention in an outpatient substance abuse aftercare program. *Adm Policy Ment Health.* 2006;33(6):629-635.

Tits, van, Experiment huisartsenhonorering, Medisch Contact 24/8 (1988), p. 255-7.

Volpp KG, John LK, Troxel AB, Norton L, Fassbender J, Loewenstein G, Financial incentive-based approaches for weight loss: a randomized trial, JAMA. 2008a Dec 10;300(22):2631-7.

Volpp KG, Loewenstein G, Troxel AB, Doshi J, Price M, Laskin M, Kimmel SE, A test of financial incentives to improve warfarin adherence, BMC Health Serv Res. 2008b Dec 23;8:272.

Volpp KG, Troxel AB, Pauly MV, Glick HA, Puig A, Asch DA, Galvin R, Zhu J, Wan F, DeGuzman J, Corbett E, Weiner J, Audrain-McGovern J., A randomized, controlled trial of financial incentives for smoking cessation, N Engl J Med. 2009 Feb 12;360(7):699-709.

Walker S, Mason AR, Claxton K, Cookson R, Fenwick E, Fleetcroft R, Sculpher M., Value for money and the Quality and Outcomes Framework in primary care in the UK NHS., Br J Gen Pract. 2010 May;60(574):213-220.

Wennberg, JE, ES Fisher, JS Skinner and KK Bronner, Extending The P4P Agenda, Part 2: How Medicare Can Reduce Waste And Improve The Care Of The Chronically Ill, Health Aff (Millwood). 2007; 26(6): 1575–1585.

Whalley D, Gravelle H, Sibbald B. Effect of the new contract on GPs' working lives and perceptions of quality of care: a longitudinal survey. Br J Gen Pract. 2008;58(546):8-14.

World Health Organization (WHO), Smoking Statistics (downloadable excel spreadsheet)

Wynia, MK, The Risks of Rewards in Health Care: How Pay-for-performance Could Threaten, or Bolster, Medical Professionalism, J Gen Intern Med. 2009 24(7): 854-859

Young GJ, Meterko M, Beckman H, et al. Effects of paying physicians based on their relative performance for quality. J Gen Intern Med. 2007;22(6):872-876.

Appendix A Search strategy

I employed the following strategy for retrieving the relevant literature. First, I searched Pubmed for recent surveys of the literature on P4P. As keywords I used "Pay for performance" and "P4P". This yielded the following results (most recent article listed first):

- Greene SE, Nash DB., Pay for performance: an overview of the literature, Am J Med Qual. 2009 Mar-Apr;24(2):140-63.
- Mehrotra A, Damberg CL, Sorbero ME, Teleki SS., Pay for performance in the hospital setting: what is the state of the evidence?, Am J Med Qual. 2009 Jan-Feb;24(1):19-28.
- Bremer RW, Scholle SH, Keyser D, Houtsinger JV, Pincus HA., Pay for performance in behavioral health, Psychiatr Serv. 2008 Dec;59(12):1419-29
- Rosenthal MB, Landon BE, Howitt K, Song HR, Epstein AM., Climbing up the pay-for-performance learning curve: where are the early adopters now?, Health Aff (Millwood). 2007 Nov-Dec;26(6):1674-82.
- Rosenthal, MB, Pay for Performance and Beyond, Expert Review of Pharmacoeconomics and Outcomes Research, Volume 7, Number 4, August 2007, pp. 351-355(5).
- Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S., Does pay-for-performance improve the quality of health care?, Ann Intern Med. 2006 Aug 15;145(4):265-72.
- Rosenthal MB, Frank RG., What is the empirical basis for paying for quality in health care?, Med Care Res Rev. 2006 Apr;63(2):135-57.

Since these surveys generally did not include sufficient detail to answer the main question, it was almost always necessary to go back to the original papers. Next I searched Pubmed for more recent papers that could not have been in these surveys. This was done through a search in Pubmed limited to the years 2008 – 2010 using the keyword "pay for performance" and "P4P". This yielded three additional papers presenting original evidence on the effects of P4P programmes.

Appendix B Summary of empirical papers on the effects of P4P

This appendix presents relevant details of the 27 empirical papers on P4P. The literature is subdivided into four groups: 1. papers based on P4P-programs initiated by private health insurance firms in the US; 2. papers based on the recent P4P-demonstration projects by Medicare in association with the Premier hospital group in the US; 3. a paper based on an older Medicare P4P-demonstration project in the US; 4. Papers based on the QOF P4P-program in the UK.

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
A. Private health insurers in the US								
Chen	2010	Primary care physicians in Hawaii	Diabetes	Monitoring HBa1c and cholesterol	Improvement over previous year determines size of bonus	Yes, but non-random: control group made up of physicians who chose not to participate	1,5% - 7,5% on top of usual fee	Statistically significant improvement in all 3 indicators; significant fall in hospitalization
Chen	2010	Individual primary care physicians and specialists in Hawaii	Heart failure, cancer, diabetes	Heart failure, cancer screening (3), HBa1c testing, vaccination (3)	Improvement over previous year determines size of bonus	Yes: 1 other health plan in south of US (not Hawaii)	1,5% - 7,5% on top of usual fee	Statistically significant improvement in 2 out of 7 indicators (screening for colorectal

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
								cancer and testing HBa1c)
Rosenthal	2009	Midwives and their patients in Las Vegas	Prenatal care	Entering care and completing regular visits	Bonus per patient, paid both to midwife and patient	Yes; instrumental variables used to account for selection bias	$100 bonus for midwife and $100 bonus for patient	Statistically significant decline in neonatal intensive care unit admission and health spending in first year of life; no significant effect on low birth weight
Rodriguez	2009	Physican groups in California	Not specified	Physician communication, care coordination, access to care, office staff interaction	Not specified	No	Not specified	statistically significant improvement in all four indicators after introduction of P4P

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Pearson	2008	Physican groups in Massachusetts	Depression, asthma, cancer, chlamydia, diabetes, child health	Antidepressant medication management, asthma medication use, cancer screening (2), chlamydia screening, cholesterol screening, diabetes care (4), well-child visits	Various designs; no details presented	Yes: physician groups in Massachusetts not (yet) included in the P4P program	Maximum of $200 - $2,500 per physician per indicator (not entirely clear from paper); on averge 2,2% income is at risk in P4+I7P incentives+I 7	Highly incentivized groups did not improve more than comparison groups
Mandel	2007	Pediatric practices	Asthma	Medication control, flu shots, written self-management plans	Payment for participation, delivery of data and achievement of fixed thresholds; in combination with	No: before/after comparison	Maximum 7% bonus on top of regular fee	Percentage of population receiving 'perfect care' increased from 4% to 88%

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
					education			
Cutler	2007	1 California medical group	Chronic heart disease (heart failure), diabetes	1 proces indicator (% patients receiving LDL-test); 1 intermediate outcome measure (LDL< 130 mg/ dL)	Payment depending on reaching certain percentile score (50% and 75%)	Yes: non participating patients in the same medical group	Undisclosed due to confidentiality issues; authors calculations indicate that payments were <5% of regular fee per patient	Improvement on both indicators; statistical significance not reported

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Felt-Lisk	2007	Individual physicians, small practices and larger medical groups	Well-baby visits	Number of cases in which sufficient visits where made; details differed across health plan	Payment per baby	Yes: national and state mean	Maximum $100 - $470 dollar per baby	Positive effects in 3 out of 5 plans; largest effects in plan featuring largest bonus
Gilmore	2007	Hospitals in Hawaii	Cancer, heart failure, asthma, diabetes, immunization	Cancer screening (3); use of drugs for secondary prevention of heart attack (3); use of asthma drugs; diabetes checks (2); childhood immunization (2)	Payment depending on percentile score; no details presented	Yes: non participating physicians insured by the same health insurer	Maximum 7,5% of base fee	Statistically significant effect of seeing participating physician in regression with "recommended care" as the dependent variable
Nahra	2006	Hospitals in Michigan	AMI, Heart failure	Appropriate medication: aspitin, beta blockers, ACE inhibitors	Bonus based on reaching threshold	No	Maximum bonus payment 2 percent	Percentage improvement in indicators ranging from 4 to 14 percent; statistical significance not reported

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Shepard	2006	Substance abuse counselers	Substance abuse, primarily heroin	Completion of 5 or more aftercare sessions	Bonus per patient completing 5 aftercare sessions	Yes, randomized	US$ 100 bonus per patient completing at least 5 aftercare sessions	Statistically significant increase in no. of patients completing at least 5 aftercare sessions
Young	2006	Primary care physicians	Diabetes	4 process indicators: HBa1c check, urinalysis, LDL check, eye exam	Relative performanc e compared to other participatin g physicians (no details given)	No; pre/post comparison in rate of change in improvement	On average $1500 (authors' estimate)	Statistically significant increase in rate of improvement in only 1 of 4 indicators (eye exams)
Levin-Scherz	2006	1 integrated delivery network comprining 1100 primar care physicians and >4000 physicians	Diabetes and asthma	Diabetes: 4 process indicators: HBa1c check, nephropaty screening, LDL check, eye exam; asthma: 1 process indicator (inhalator use by children)	Bonus for achieving fixed targets; no details given	No, comparison with state mean and national mean	Maximum bonus n 2005 % $ 86 million; maximum bonus $5000 per physician; no data given on actual	Statistically significant increase in all 4 diabetes indicators; no increase in asthma indicator

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
							payments	
Rosenthal	2005	Physican groups in California	Cancer, diabetes	Cervical cancer screening, mammography, HBa1c testing	Bonus for attaining fixed target (% screened)	Yes: 1 other health plan operating in Oregon and Washington	Maximum bonus about 0,8% of overall revenu of group	Statistically significant improvement in cervical screening
Beaulieu	2005	Individual physicians	Diabetes	7 process (screening, testing) and **3 outcome indicator**s (blood pressure <130/80, HBa1c < 7,5%, LDL <100mg/dl); total weight of outcome in composite indicator 60%	Bonus upon reaching benchmark or improving by more than 50% on composite indicator	No	Actual payment between $3000 and $12000 per physician	Statistically significant improvement in 6 out of 7 process indicators and 2 out of 3 outcome indicators (HbA1c and LDL)

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Greene	2004	Internists, family practitioners, pediatricians	Acute sinusitis	Adherence to agreed standard of recommended care	Not specified	No; pre/H4post comparison in adherence to standards	Not specified	Significant improvement in adherence to standards
Roski	2003	40 primary care clinics	Smoking	Documentation of smoking status and documentation of advice to quit smoking	Payment depending on reaching fixed targets (75% and 65% of current smokers)	Yes, random assignment of clinics to treatment/control group	Maximum bonus $10,000 per clinic	Statistically significant improvements in both indicators but no statistically significant impact on smoking cessation rates
Amund-son	2003	20 physician groups in Minneapolis	Smoking	Documentation of smoking status and documentation of advice to quit smoking	Payment depending on reaching fixed targets (80% on both indicators)	No	Maximum bonus per group $43,750	Large increase in performance (25% and 50%), but this may be due to other factors

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Fairbrot-her	1999	Pediatricians and family practice physicians in New York city	Childhood immunization	Up to date immunization status	Payment for achieving 80% coverage and for improving 20% or 40% from baseline	Yes, random assignment of physicians to treatment/control group	maximum bonus $7,500	No improvement in immunization rates compared to control group
Hillman	1999	53 primary care practices in Philadelphia	Childhood immunization	% of children receiving appropriate vaccinations	Payment for top 6 performers	Yes, random assignment of physicians to treatment/control group	20% bonus on top of regular fee for 3 best performing practices; 10% bonus for next 3 practices	No improvement in immunization rates compared to control group
Hillman	1998	52 primary care practices in Philadelphia	Cancer screening	% of women 50 years and older that have received pap test, colorectal screening, mammography or breast exam	Payment for top 6 performers	Yes, random assignment of physicians to treatment/control group	Maximum bonus $ 1,260	No improvement in screening rates compared to control group
B. Medicare Premier demonstration project								

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Linden-auer	2007	Hospitals	Heart failure, AMI, pneumonia	10 indicators: % appropriate medication (8), % receiving appropriate diagnostic tests (2)	Top 2 deciles receive bonus	Non-selected hospitals	1%-2% bonus on top of usual fee	Statistically significant improvement in 7 indicators
Glickman	2007	Hospitals	AMI	Process of care (medication, diagnostics), smoking cessation counseling, **hospital mortality**	Top 2 deciles receive bonus	Yes: other hospitals participating in the same quality improvement program, but without P4P	1%-2% bonus on top of usual fee	Statistically significant improvement 3 out f 14 process indicators (including smoking cessation counceling) ; no improvement in mortality
Grossbart	2006	Hospitals	AMI, heart failure, pneumonia	AMI: 8 process of care indicators; heart failure: 4 process of care indicators; pneumonia: 5 process of care indicators	Top 2 deciles receive bonus	Yes: other hospitals that chose not to participate in a quality improvement program	1%-2% bonus on top of usual fee	Statistically significant improvement 8 out of 17 process indicators (including improvements on all 5 indicators for pneumonia)

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
C. Older Medicare demonstration project								
Kouides	1998	52 primary care practices in Rochester, New York	Elderly influenza immunization	Immunization rate	Bonus for reaching 70% or 85% immuni-zation rate	Yes, random assignment of physicians to treatment/ control group	10% or 20% bonus on top of usual fee	Statistically significant 7 percent (4 percentage points) improvement in immunization rates
D. Quality and Outcomes Framework (QOF) in UK primary care								
Campbell	2009	Family practices	asthma, diabetes, heart disease	indicators for process of care: 13 for asthma, 15 for heart disease, 21 for diabetes	payment based on number of points earned on 136 indicators	No; analysis is based on change in trend in composite indicator for each condition	up to 25% of total income	Increase in rate of improvement for asthma and diabetes (i.e. accelerating trend); no change in trend for heart disease

First author	Publi-cation year	Type of provider	Condition(s)	Performance indicator(s)	Design	Control group	Size of payment	Effect
Millett	2007	Family practices	Smoking by people with diabetes	Documentation of smoking cessation advice, prevalence of smoking	Payment based on number of points earned on 2 indicators	No; analysis is based on % change before/after	Not specified, probably small since only 8 points can be earned out of a total of 99 for all aspects diabetes care	Statistically significant increase in % of smokers receiving cessation advice between 2003 and 2005; statistically significant decrease in % of smokers between 2003 and 2005 (fall from 20% to 16.2%)