Pay-for-performance for healthcare providers

Design, performance measurement, and (unintended) effects

Frank Eijkenaar

Chapters based on articles published in peer-reviewed scientific journals, with kind permission of Springer Science + Business Media (chapters 2 and 4), SAGE Publications Ltd./Inc. (chapters 3 and 6), Elsevier Inc. (chapter 5), and Lippincott Williams & Wilkins (chapter 7).

Printing and layout: Optima Grafische Communicatie, Rotterdam, the Netherlands.

ISBN: 978-94-90420-40-6

Pay-for-Performance for Healthcare Providers

Design, performance measurement, and (unintended) effects

Prestatiebeloning voor zorgaanbieders

Vormgeving, prestatiemeting en (onbedoelde) effecten

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 14 november 2013 om 13.30 uur

door

Frank Eijkenaar

geboren te Hendrik-Ido Ambacht

PROMOTIECOMMISSIE

Promotor

Prof.dr. W.P.M.M. van de Ven

Overige leden

Prof.dr. F.T. Schut

Prof.dr. E.W. Steyerberg

Dr. J.C.C. Braspenning

Copromotor

Dr. R.C.J.A. van Vliet

CONTENTS

PUBLICATIONS

Chapters 2 through 7 are based on the following articles:

Chapter 2:
Eijkenaar, F. 2013. Key issues in the design of pay-for-performance programs. *European Journal of Health Economics* 14(1): 117-131.

Chapter 3:
Eijkenaar, F. 2012. Pay-for-performance in Health Care: An International Overview of Initiatives. *Medical Care Research and Review* 69(3): 251-276.

Chapter 4:
Emmert, M., Eijkenaar, F., Kemter, H., Esslinger, A.S., Schöffski, O. 2012. Economic evaluation of pay-for-performance in health care: a systematic review. *European Journal of Health Economics* 13(6): 755-767.

Chapter 5:
Eijkenaar, F., Emmert, M., Scheppach, M., Schöffski, O. 2013. Effects of pay-for-performance in health care: a systematic review of systematic reviews. *Health Policy* 10(2-3): 115-130.

Chapter 6:
Eijkenaar, F., van Vliet, R.C.J.A. 2013. Performance profiling in primary care: does the choice of statistical model matter? *Medical Decision Making*. DOI: 10.1177/0272989X13498825.

Chapter 7:
Eijkenaar, F., van Vliet, R.C.J.A. 2013. Profiling individual physicians using administrative data from a single insurer: variance components, reliability, and implications for performance improvement efforts. *Medical Care* 51(8): 731-739.

CHAPTER 1

Introduction

1.1 BACKGROUND

Healthcare systems around the world are characterized by a suboptimal delivery of healthcare services. For example, while healthcare expenditures continue to rise in many countries (OECD, 2012), clear deficiencies in the quality of care have been demonstrated, including limited adherence to professional medical guidelines (McGlynn et al., 2003; Steel et al., 2007), limited progress in improving patient safety (Benning et al., 2011; Shekelle et al., 2011), and avoidable complications and mortality (Langelaan et al., 2008; de Brantes et al., 2010). In addition, there appears to be considerable unwarranted variation in quality of care and utilization, both among geographic areas and among healthcare providers (Fuchs, 2004; Wennberg, 2010; Douven et al., 2012).

As a result, policymakers are exploring methods to increase the quality and efficiency of care. There has been a growing belief that many deficiencies stem from flawed provider payment systems creating perverse incentives for healthcare providers. In many countries, providers are largely being compensated on a *fee-for-service* basis (even when they are members of a group paid by capitation), which, provided that fees exceed marginal costs, entails a strong financial incentive to increase the quantity of provided services but not necessarily the quality with which services are provided. Assuming a provider's actions are partly motivated by financial considerations (i.e., his utility is partly determined by net income/profit), paying providers on a fee-for-service basis could result in *supplier-induced demand* (Evans, 1974; Ginsburg & Grossman, 2005), which exists when a provider influences a patient's demand for care against the provider's interpretation of the best interests of the patient (McGuire, 2011). As concluded by McGuire (2011), the available empirical evidence makes a convincing case that providers can influence quantity and indeed sometimes do so for their own purposes. Therefore, paying providers solely on a fee-for-service basis could result in overtreatment or inappropriate treatment, which does not contribute to an efficient delivery of care and may even negatively impact patient health (Institute of Medicine, 2001).

Against this background, there has been an increased emphasis in many countries on reforming provider payment systems. By focusing on restructuring the financial incentives at the supply-side, policymakers aim to increase efficiency by making it financially worthwhile for providers to provide high-quality, patient-centered care in a cost-conscious way. One result of this development is the emergence of the term *pay-for-performance* (P4P), which originated in the United States (US) and has been used to designate any payment scheme designed specifically to directly stimulate providers to increase the quality and efficiency of care. Although several innovative health insurers in the US already experimented with the concept in the late 1980s and early 1990s (e.g., Hanchak et al., 1996; Fairbrother et al., 1999), it was not until the early 2000s when P4P received a boost as a result of two seminal reports of the Institute of Medicine on medical errors ("To Err is Human", 1999) and overall quality of care ("Crossing the Quality Chasm", 2001). By identifying serious flaws in the

quality of care and by demonstrating how provider payment systems contribute to these flaws, the reports stimulated employers, purchasers of care, and government agencies to start working on alternative payment schemes incorporating explicit incentives for quality. By 2007, the number of P4P-programs in the US had grown to over 200, targeting a variety of providers and increasingly also focusing on costs of care. Most of these programs are sponsored by private health insurers, although state and federal government agencies are actively experimenting with P4P as well.

Over the past decade, P4P has attracted widespread interest, with programs being uncritically implemented in other high-income countries (e.g., Australia, Canada, France, Germany, the Netherlands, New Zealand, Spain, the United Kingdom) and more recently also in low- and middle-income countries (e.g., China, the Philippines, Rwanda, Taiwan, Tanzania) (Epstein, 2006; McNamara, 2005; Witter et al., 2012). To a large extent, this widespread interest in P4P appears to be a result of its intuitive appeal. Given that healthcare providers are responsive to financial incentives (Christianson & Conrad, 2011), that a considerable share of total healthcare consumption can be directly influenced by healthcare providers, and that it has become increasingly possible to objectively measure quality of care (the field of quality measurement has progressed significantly over the past 20 years, and so have the breadth and sophistication of measures to assess quality), to many it makes sense to use explicit financial incentives to stimulate healthcare providers to improve the quality and efficiency of care.

In contrast to what this widespread interest in P4P suggests, however, to date P4P does not appear to have been effective in improving the quality and efficiency of care (Petersen et al., 2006; Christianson et al., 2008; Van Herck et al., 2010). A broad evidence base is lacking, and studies with the strongest research designs have shown inconclusive results. Moreover, several studies have found evidence of unintended consequences of P4P, including risk selection (Shen, 2003; Chen et al., 2010) and neglect of unrewarded aspects of performance (Campbell et al., 2009). In part, this may have been a result of the limited knowledge about crucial aspects of the design and implementation of P4P. In addition, although interest in how payment affects the *type and amount* of services provided and the *costs* of care to purchasers is not new, examining how financial incentives may impact on *quality* of care is a fairly new area of inquiry (Christianson & Conrad, 2011). This thesis addresses these issues by exploring key conceptual and practical issues in the design and implementation of P4P, by synthesizing existing empirical literature on effects of P4P, and by addressing important empirical questions about the complex issue of performance measurement. Before formulating the specific research questions central to this thesis, however, the next two sections first provide a theoretical background regarding the use of financial incentives in health care and an illustration of how P4P may be applied in practice.

1.2 THEORETICAL BASIS OF USING FINANCIAL INCENTIVES TO STIMULATE EFFICIENT PROVIDER BEHAVIOR

1.2.1 Agency theory

A theoretical basis for using financial incentives to promote improvements in quality and efficiency can be found in agency theory, a relatively new theory in economics that is linked to the assessment of the effectiveness and efficiency of provision of information within a set of institutional agreements (Vosselman, 1996). In general, the main focus of the theory is problems that may occur within contractual relationships between two parties, the principal and the agent. These relationships are characterized by information asymmetry, conflicting interests, and outcome uncertainty. Regarding *information asymmetry*, the theory assumes that information is a commodity that can be purchased (Eisenhardt, 1989). The principal (the relatively ill-informed party delegating decision-making authority regarding specific tasks to the agent, the relatively well-informed party) often has to incur substantial costs if he wants information on the amount of effort made by the agent, which actions the agent has chosen to perform and on which information these are based, and whether or not the agent has made appropriate decisions. This is especially true because the agent typically has no incentive to reveal this information to the principal because of *conflicting interests* (MacDonald, 1984). The principal has an interest in as much effort as possible on the side of the agent, while the risk-averse agent is assumed to derive disutility from effort (Shavell, 1979; MacDonald, 1984). In addition, since agency relations are characterized by *uncertainty of outcome*, the principal is not able to draw clear conclusions about the agent's contribution to the outcome.

Agency theory is particularly relevant to many important relationships in health care. The most relevant relationships here are those between the healthcare provider and the patient and, by extension, between the provider and the purchaser of care (Dranove & White, 1987; Blomqvist 1991; Blomqvist & Léger 2005; Vermaas, 2006). The provider is generally assumed to be the well-informed agent and the patient (who has an interest in full attention and effort of his physician) and purchaser (who has an interest in cost-conscious behavior and efforts to provide high quality care on the part of the provider) are generally seen as the ill-informed principals. However, conflicts of interests are likely because the interests of the patient and the purchaser are unlikely to be the only variables determining the provider's utility; income, reputation, leisure, and workload may also be of influence (Evans, 1974; Dranove, 1988; McGuire, 2000). In addition, in most instances the purchaser and especially the patient will not be able to determine whether or not the provider has acted in their best interests because of information asymmetry and because the outcome is uncertain; in addition to the provider's actions, the outcome also depends on the actions of the patient (e.g., lifestyle, compliance with treatment) and unforeseeable external factors.

Several problems may arise in agency relationships, which mainly are a result of conflicting interests giving the agent an incentive to exploit his information surplus. The most relevant problem here is the agency problem, which may occur before (adverse selection) or after (moral hazard) the contract is concluded (Eisenhardt, 1989). *Adverse selection* occurs when the principal concludes a contract based on false or incomplete information provided by the agent; the principal would never have agreed if he were equally informed. *Moral hazard* may arise when the principal is unable to monitor the agent's actions, but has information on the outcome of the activity. The principal has no guarantees the agent will pursue the best possible outcome. After all, the agent knows the principal is unable to discern between his contribution to the outcome and the influence of other factors (Vosselman, 1996). A well-known example of moral hazard in the context of health care is the issue of supplier-induced demand, in which the provider exploits his information surplus to induce demand against his interpretation of what is in the best interests of the patient (full attention of his doctor, the right type and level of care) and the purchaser (efficient provision of appropriate care).

Agency theory offers several strategies the principal could apply to deal with such problems. These strategies may be directed at aligning the agent's interests with his own interests or at reducing his arrears in information. In health care, the focus is often on the means available to purchasers (acting as prudent buyers of care on behalf of their members) to prevent providers' from exploiting their information surplus (Schut & van Doorslaer, 1999). Vermaas (2006) discerns three strategies: selecting the provider, controlling the provider, and monitoring the provider (including profiling activities such as informing providers about their performance relative to each other and/or to a norm). Financial incentives fall under the second strategy and will typically be applied to align interests. The choice for a particular incentive scheme depends on both parties' attitudes towards risk and on the information possessed by the purchaser on the outcome and on the provider's efforts (Shavell, 1979; Vosselman, 1996; Vermaas, 2006). In health care, some dimensions of the quality of care are not observable and hence not contractible. This is the *problem of multitasking* (Holmstrom & Milgrom, 1991), which refers to "the challenge of designing incentives to motivate appropriate effort across multiple tasks when the desired outcomes for some tasks are more difficult to measure than others" (Eggleston, 2005: 211). In practice, therefore, different payment schemes incorporating diverging incentives are often combined to achieve a balanced incentives structure (see below). In addition, the literature on financial incentives in health care suggests that desired outcomes are not likely to be achieved by financial incentives alone and that monitoring activities are required as well (Christianson & Conrad, 2011). Yet monitoring requires implementation of an often expensive information system, and providers will insist on being compensated appropriately when confronted with (financial) risk. The greater the cost of monitoring activities, the less likely the use of financial incentives will be cost-effective in changing provider behavior (Christianson & Conrad, 2011).

1.2.2 Methods for paying healthcare providers

Because of the multitasking problem and the large degree of outcome uncertainty (Eggleston, 2005), payments for healthcare providers will always at least consist of a component that is unrelated to performance. This "base payment" will typically comprise the majority of a provider's revenues. There are various base payment methods that purchasers of care could apply, which can be discerned according to the *providers* and the *care* covered by the payment. Regarding the former, roughly three options are possible: one separate payment for care provided by one type of provider (e.g., a physician or a hospital), one combined payment to a "main contractor" for care provided by at least two types of providers who are relevant for a part of the relevant basic benefits package, and one integral payment to a main contractor for care provided by all relevant providers. Under the latter two options, the main contractor receiving the payment could in turn use various payment methods to compensate associated "subcontractors". Regarding the *care* covered by the payment, a distinction can be made in payment per visit or procedure and payment per "bundle" of different types of care. Bundled payments can be made per patient per admission, per patient per period (e.g., a disease-episode or a year), or per insured per period (e.g., a month or a year). Table 1.1 shows how these options can be combined into various base payment methods.

Each of the base payment methods listed in Table 1.1 has advantages and disadvantages. A disadvantage of *payment per visit/procedure* is that it provides an incentive to provide unnecessary care, implying excessive use of resources and possibly also detrimental effects on patients' health. In addition, it provides an incentive for "upcoding", that is, incorrectly classifying patients in treatment categories with higher fees or tariffs. Furthermore, payment per visit/procedure contains no intrinsic incentive to provide care of high "clinical" quality because the additional care required as a result of complications is reimbursed. On the other hand, there is an incentive to achieve a good patient satisfaction/experience as patients can then more easily be seen often. Another advantage is that there is no incentive for *stinting* on quality. After all, because the costs of providing necessary (expensive) treatments are fully reimbursed, there is no incentive not to provide these treatments (provided the fees exceed the marginal costs to provide the care). Finally, a disadvantage of separate payments for one type of provider (which by definition is the case under payment per visit/procedure) is that there is no incentive for efficient care coordination and collaboration among providers. To a lesser extent, this also holds for payments for care provided by at least two types of providers.

In case of *bundled payment per patient*, there may be an incentive to still provide care when this may not be warranted (the "grey area"). The extent to which this incentive exists depends on how the payment system is designed. For example, under the Diagnosis-Related Group (DRG) system in Medicare in the US there is an incentive to admit a patient to the hospital even when this patient could have been treated in an outpatient setting. This incen-

TABLE 1.1 Base payment methods and some examples

	Separate payment for the care provided by one type of provider	One payment for the care provided by at least two types of providers that are relevant for a part of the basic package	One integral payment for the care provided by all providers that are relevant for the basic package
Payment per visit/procedure	• Fee per office visit • Fee per ultrasound		
Payment per patient per admission	• Payment to a hospital (excl. specialist) per Diagnosis-related group in Medicare in the United States		
Payment per patient per period	• Payment to a medical specialist per referral from a general practitioner (as for formerly publicly insured patients in the Netherlands)	• Payment to a hospital (incl. specialist) per diagnosis treatment combination (DBC) in the Netherlands • Payment to a main contractor for care related to a chronic condition (e.g., "chain-DBCs" in Netherlands)	
Payment per insured per period	• Capitation payment to a general practitioner (as for formerly publicly insured patients in the Netherlands)	• Budget for a general practitioner that also covers a specified percentage of follow-up care (e.g., GP fundholding in the UK)	• Health maintenance organizations (HMOs) in the United States • Global budgets for physician groups (e.g., 'total purchasing' in the UK)

tive does not exist under the broader Dutch system of Diagnosis Treatment Combinations (DBCs), which instead provides an incentive to the hospital for efficient substitution of care that is part of the DBC. For both systems, the opportunities for upcoding increase (but the incentives for risk selection decrease) as the system contains more categories. Incentives for quality and prevention depend on the extent to which future complications and health problems are part of the bundled payment. For example, when the costs of readmissions due to complications are part of the payment, incentives for quality are large because high-quality care decreases the risk of complications.

Finally, *bundled payment per insured* provides an incentive to reduce costs for care part of the bundle as well as incentives for risk selection and cost shifting. The incentive to select patients/insured with low expected expenses increases as the bundle covers more (follow-up) care, and can be mitigated by adjusting the payment for relevant risk characteristics. Possibilities for cost shifting decrease as the bundle covers more care, and are minimized under integral payments. A classic example of possibilities for cost shifting is the capitation payment to Dutch general practitioners (GPs) for formerly publicly insured persons (before 2006). The care that GPs had to provide was not explicitly defined in the contract, so GPs could shift costs simply by referring patients to a medical specialist. Incentives for

quality and prevention increase (and incentives for undertreatment decrease) as a larger share of the costs of future complications and other health problems fall under the bundled payment, and are maximized under integral payments. With integral payments, the main contractor has an incentive for prevention and "health maintenance" because this prevents future expenses. Because of this, the main contractor also has an interest in passing these incentives along to subcontractors.

1.2.3 Pay-for-performance as a supplement to base payments

An option to mitigate the disadvantages of base payment methods is to combine different base payment methods (Robinson, 2001; McGuire, 2011; Christianson & Conrad, 2011), as is currently the case in Dutch general practice where GPs are being compensated through a mixture of payments per visit/procedure, bundled payments per patient ("chain-DBCs" for chronic conditions), and payments per insured per period (capitation). However, it is impossible to remove all disadvantages by combining base payment methods. Moreover, other problems may arise, like paying twice for the same care when combining payments per visit/procedure with bundled payments. Another option to mitigate the disadvantages of base payment methods is to supplement base payments with *explicit financial incentives for quality and cost containment* via P4P. Regarding (clinical) quality, such payments could depend on scores on measures of structure (e.g., working with an electronic medical record), process (e.g., administering beta-blockers after a heart attack), and/or outcome (e.g., mortality within 30 days among patients who had bypass surgery). Although improving health outcomes will be the ultimate goal of P4P, a major problem with using outcome measures is that compared to process measures they are much more sensitive to random chance and other factors that are difficult to influence by providers, such as patients' adherence to treatment and other patient characteristics (casemix). The payments may also depend on expenses. For example, they could depend on the difference between a normative (expected) level of expenses and actual expenses. As with clinical outcomes, however, random chance and casemix may influence providers' scores to a large extent, necessitating large patient numbers as well as risk adjustment to mitigate incentives for risk selection and to obtain relevant and meaningful results.

A major problem with P4P is that precise metrics for provider actions that promote quality are difficult to quantify and that some dimensions of quality will never be contractible (Eggleston, 2005). Consequently, explicitly incentivizing specific aspects of performance can distort resource allocation to the measured aspects and away from unmeasured aspects (Newhouse, 2002). This issue provides a strong case for combining prospective fixed payments with retrospective volume-based payments, in addition to P4P. Not only can this potentially reduce incentives for risk selection and quality stinting while maintaining incentives for cost containment (e.g., Ellis & McGuire, 1990; Ma, 1994; Newhouse, 1996, 2002; Ma & McGuire, 1997; Pauly, 2000), it also enables balancing incentives for quality across contractible *and* non-contractible dimensions (Eggleston, 2005).

1.3 TWO PROTOTYPICAL PAY-FOR-PERFORMANCE PROGRAMS

As noted, P4P has widely been adopted as a performance improvement strategy in health care. There are several programs that stand out, both in terms of scope (e.g., the amount of money at stake, the number of participants, the number of performance measures) and in terms of duration. Two of these are briefly discussed below: the Hospital Quality Incentive Demonstration in the US and the Quality and Outcomes Framework in the UK. Both have served as examples for the design and implementation of many other P4P-programs throughout the world.

1.3.1 The Hospital Quality Incentive Demonstration

The Hospital Quality Incentive Demonstration (HQID) was a P4P-program managed by the Centers for Medicare and Medicaid Services (CMS) and Premier Inc., a coalition of 2,500 hospitals. The HQID ran from October 2003 until September 2009 and was designed to acknowledge and reward hospitals providing high-quality care. Of the 421 hospitals invited to participate, 266 chose to do so (Lindenauer et al., 2007). Although the payments were applied only for Medicare beneficiaries, performance was measured for *all* patients admitted for the following "clinical areas": heart failure, acute myocardial infarction (AMI), pneumonia, coronary artery bypass graft (CABG) surgery, and knee- and hip-replacement. Eligible hospitals could participate in each of these areas, provided they had at least 30 patients with the relevant condition at the end of the reporting year. Incentive payments were paid as add-on to the relevant DRG payment per admission, and were based on an overall composite score per clinical area (Ryan et al., 2012a). Most areas consisted of both process measures (e.g., aspirin at arrival for AMI patients) and outcome measures (e.g., inpatient mortality for CABG patients), although the emphasis was on process quality (26 of the 33 measures were process measures). Most measures were adjusted for demographic and clinical risk characteristics of patients, such as age, sex, and preexisting (chronic) conditions. Performance data were self-reported by participating hospitals and extensively checked and validated by CMS and Premier. Hospitals were allowed to exclude certain patients from counting toward their quality performance, provided there were appropriate reasons for doing so (Ryan, 2010).

The demonstration consisted of two phases (Ryan et al., 2012a, 2012b). In phase 1 (2003-2006) CMS paid a 2 percent add-on to the relevant annual DRG payment per Medicare beneficiary to hospitals scoring in the top 10 percent of all participating hospitals. Hospitals scoring in the second highest decile received a 1 percent add-on. Thus, receipt of payments not only depended on hospitals' own performance, but also on the performance of other hospitals. From year 3, financial penalties of maximally 2 percent were imposed on hospitals scoring below the 20th percentile of hospitals' scores in year t-2. In phase 2 (2006-2009), several alternative designs were tested. For each clinical area, hospitals could earn payments

for attainment (scoring above the median of year t-2), top performance (being in the top 20 percent in the current year), and improvement (scoring above the median of year t-2 and being in the top 20 percent with the largest percent improvement). Of the annual budget for incentive payments, 60 percent was reserved for top performance and improvement, and 40 percent for attainment. Payments to hospitals averaged to $12 million per year in phase 2, against $8.2 million in phase 1 (Ryan, 2009a).

The HQID served as a pilot for a much larger national P4P-program. Mandated by the Affordable Care Act, CMS launched the "hospital value-based purchasing program" in 2012, which is mandatory for all acute care hospitals. The US Department of Health and Human Services described the goals of this program as follows: "to transform Medicare from a passive payer of claims to an active purchaser of quality health care for its beneficiaries" and "to transform how Medicare pays for care and to encourage hospitals to continually improve the quality of care they provide" (Health and Human Services, 2011: 26490, 26543). The program differs from the HQID in several respects. In addition to patient safety and clinical quality, performance scores are now also based on patient experience and spending. In addition, payments are calculated differently: for each performance measure, two scores are determined, one for absolute performance and one for improvement, both relative to a minimum performance level and a benchmark. The score a hospital gets is the higher of these two scores. Explicit performance targets have largely been abolished and replaced by a linear point system to translate scores into payments. Because payments are being financed by a generic 1 percent cut of DRG payments amounting to $850 million annually (which will gradually be increased to 2 percent in 2017), the program is largely budget-neutral (Health and Human Services, 2011).

1.3.2 The Quality and Outcomes Framework

One of the largest P4P-programs in the world is the Quality and Outcomes Framework (QOF) in the UK, which was implemented in 2004. Under the QOF, GP practices receive substantial financial rewards for scoring well on a large number of performance measures (Roland, 2004; Doran et al., 2006). The 2011/2012 QOF consists of 142 measures divided over four domains: clinical (87 measures), organizational (45 measures), patient experience (one measure), and additional services such as child health surveillance and maternity services (nine measures). The measure set and the weights attached to individual measures are based on negotiations between the Government and the British Medical Association. More recently, the National Institute of Clinical Excellence (NICE) has been involved in selecting, defining, and updating the measures. Within each domain, measures are divided over different areas (31 in total). For example, there are twenty clinical areas, including diabetes, heart failure, and hypertension. Similar to the HQID, practices are scored mainly on process measures, although the QOF also contains some outcomes. Performance scores are not adjusted for casemix, but GPs are allowed to exclude certain patients (e.g., those

who are noncompliant with treatment) from the performance measurements. Audits and penalties for fraud are to prevent inappropriate use of this system of "exception reporting" (Doran et al., 2008a).

For each practice, performance data are extracted automatically from a uniform system of electronic medical records and collated in a central database. This health IT system, for which GP practices were largely compensated (Doran & Roland, 2010), provides practices with ongoing insight in their performance as well as automated prompts and reminders. Practice-specific performance is translated into payments via a point system. Each measure has a lower target, an upper target, and a maximum number of points that can be earned. Between the two targets, performance is measured on a continuous scale and practices earn more points for reaching higher levels (Doran et al., 2008b). In 2011/2012, each point is worth £125, and 1,000 points can be earned in total. In effect, bonus payments can add up to as much of 30 percent of practices' revenues (Doran & Roland, 2010). Although the program is voluntary, participation is virtually 100 percent (about 8,600 practices).

1.4 RESEARCH QUESTIONS AND RELEVANCE

In contrast to what the popularity of P4P in practice suggests, its effectiveness has not been convincingly confirmed. As argued by several commentators, this lack of evidence may have partly been a result of flaws in the design of current P4P-programs (Rosenthal & Frank, 2006; Petersen et al., 2006; Rosenthal & Dudley, 2007; McDonald & Roland, 2009; Jha, 2013). Despite over a decade of experimenting with P4P, we still know very little about which specific design features contribute to (un)desired effects (Roland, 2012). Given that the interest in P4P is more likely to increase than decrease in the coming years in view of the continued problems in healthcare delivery and the absence of a single ideal payment system, knowledge of crucial design features is therefore urgently required. In this respect, insight in how P4P is being designed in practice and the extent to which this is adequate is also important. This insight is largely lacking, as there has been no comparative investigation of the design of major P4P-programs around the world. Two important research questions are therefore:

Q1: What are crucial design features of a successful P4P-program?
Q2: How is P4P currently being designed in practice and to what extent is this design adequate?

Regarding question 1, a P4P-program can viewed as "successful" if it is effective in attaining its goal (e.g., substantially increasing providers' adherence to professional medical guidelines) without unintended consequences (e.g., deterioration of other important aspects of the care that are not rewarded, such as continuity of care and patient satisfaction).

Along with the interest in P4P, the literature on the effects of P4P has expanded rapidly over the past decade. Although this is a desirable development, the evidence has become fragmented. Several reviews have attempted to synthesize the evidence, but they all had different foci (e.g., only including experimental studies, only focusing on prevention, etc.) and hence different conclusions. Consequently, it is challenging to comprehend this evidence and to extract success factors and pitfalls when it comes to designing and implementing P4P. In addition, literature reviews have typically overlooked a crucial aspect of P4P performance: cost-effectiveness. Although high-quality care is clearly an important goal, resources are scarce and ideally allocated to improvement efforts yielding most value for money. In addition to effectiveness, therefore, it is important to assess the cost-effectiveness of improvement efforts. The complexity of P4P-program design and the fact that running a P4P-program (e.g., engaging providers, collecting and validating performance data, calculating incentive payments) likely involves significant transaction costs, cast doubt on whether P4P can be a cost-effective improvement strategy. Two additional research questions are therefore:

Q3: What is the current state of evidence on the cost-effectiveness of P4P?
Q4: What is the current state of evidence on effects of P4P?

One of the most crucial aspects of the design of P4P is performance measurement. As noted, the performance of a healthcare provider has many dimensions, and the measurable aspects can be measured in various ways. Especially in health care, performance measures may be particularly sensitive to specific patient characteristics (e.g., age, sex, socioeconomic status, preexisting conditions, severity of disease) and random chance. To account for that, an appropriate statistical model is essential. Various models are available for analyzing performance differences among providers. In practice, purchasers would prefer to use a model that is easy to implement, maintain, and explain to providers. However, performance data in health care typically have specific features rendering simple models potentially unsuitable for modeling these data. It is unclear, however, to what extent the choice of statistical model really affects the results of provider performance comparisons (rankings). A fifth research question is therefore:

Q5: To what extent does the choice of statistical model used for risk adjustment affect the results of comparative provider performance assessments?

In addition to risk adjustment to prevent systematic misclassification of providers due to differences in casemix (Ash et al. 2012; Iezzoni 2003), performance measurements require adequate reliability to prevent random misclassification of providers due to chance (Adams et al. 2010a; Safran et al. 2006). When performance measurements have low reliability, they

are driven by random chance instead of true performance, and P4P incentives based on them may arbitrarily and unfairly penalize or reward providers. For performance measurements and comparisons to be reliable, a sufficient number of patients per provider must be sampled. In addition, variation between providers must be sufficiently large relative to variation within providers. Previous research examining performance variation and reliability has mainly focused on groups of physicians and/or used data from large public purchasers or pooled, often cross-sectional data from multiple purchasers. Yet in practice performance comparisons are still predominantly being applied by individual (private) purchasers. In addition, since many physicians still work in solo or small-group practices and individual physicians make important decisions that affect performance, assessing individual physicians' performance continues to be the predominate approach to provider performance measurement and comparison. However, when single-purchaser data are used to profile individual physicians' performance, adequate reliability is uncertain due to small sample sizes. A sixth research question is therefore:

Q6: To what extent can individual physicians be reliably compared with respect to their performance on measures derived from the administrative data of a single private care purchaser?

1.5 GOAL AND STRUCTURE OF THIS THESIS

The goal of this thesis is to provide answers to the six research questions. In doing so, it aims to provide insight in key conceptual and practical issues in the design and implementation of P4P for healthcare providers, as well as to provide recommendations regarding these issues. Accordingly, it aims to provide an important contribution in the process towards maximizing the value of using P4P to increase the quality and efficiency of care.

The remainder of this thesis is structured a follows. As shown in Table 1.2, the thesis consists of three main parts, each containing two chapters. Regarding *part 1* (design and implementation of P4P), chapter 2 examines research question 1 by identifying and synthesizing relevant theoretical and empirical literature as well as findings from previous work on P4P-program design into a single comprehensive overview of key design features. Then, using this overview, research question 2 is addressed in chapter 3 by systematically describing and critically reviewing major P4P-programs that have been implemented throughout the world. By providing lessons from experiences with P4P in practice and facilitating comparison of typical program design in different settings, the knowledge obtained in this chapter will be of particular interest to policymakers and purchasers intending to apply P4P. Regarding *part 2* (effects of P4P), chapter 4 focuses on research question 3 by systematically reviewing the empirical literature on the cost-effectiveness of P4P as assessed by economic evaluations. Research question 4 is examined in chapter 5; by conducting a systematic

TABLE 1.2 Structure of the main body of this thesis

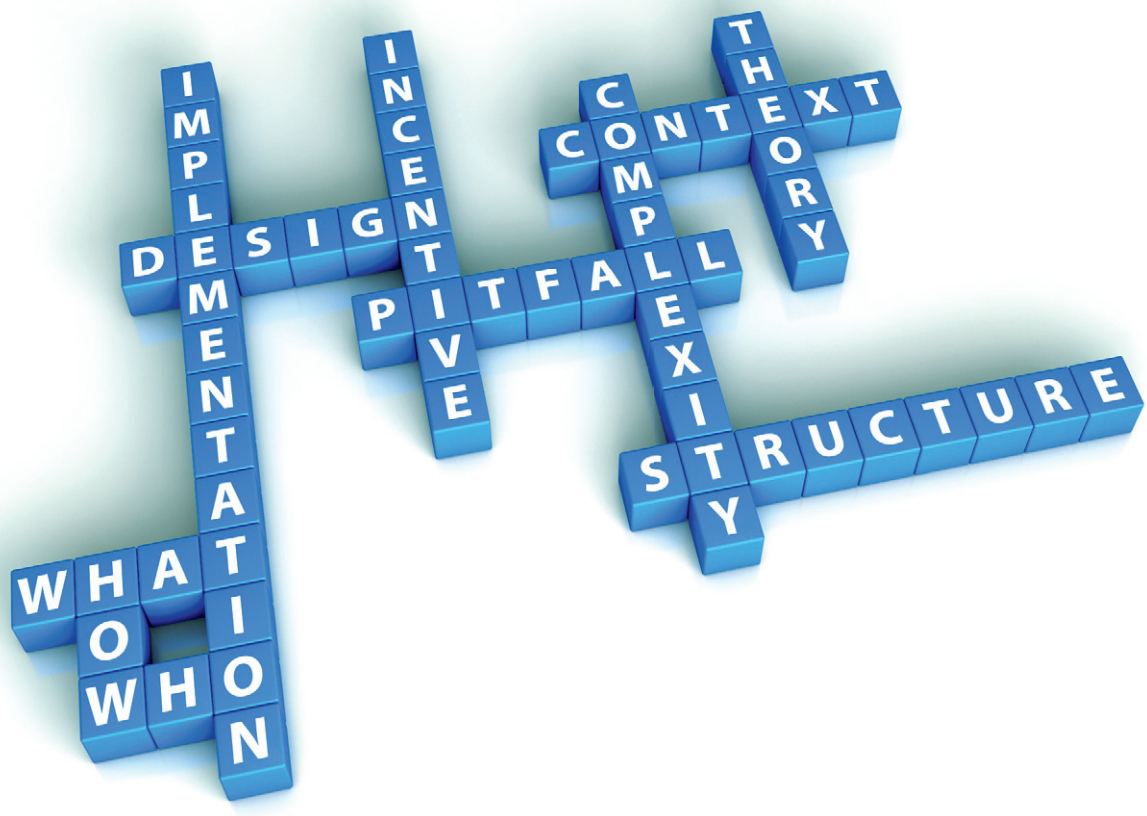
	Research question	Addressed in chapter
Part 1. Design and implementation of P4P		
• Design and implementation of P4P in theory	1	2
• Design and implementation of P4P in practice	2	3
Part 2. (Unintended) effects of P4P		
• Cost-effectiveness of P4P	3	4
• Effects of P4P in a broad sense	4	5
Part 3. Statistical issues in performance measurement		
• Impact of the choice of statistical model on provider performance rankings	5	6
• Reliability of individual physician performance comparisons based on single-purchaser data	6	7

review of published systematic reviews, the chapter provides a structured overview of the existing evidence on P4P effects. Regarding *part 3* (statistical issues in performance measurement), chapters 6 and 7 address research questions 5 and 6 using patient-level administrative claims data from a large Dutch health insurer. Chapter 6 empirically examines the extent to which a variety of statistical models produce different rankings of Dutch GPs and health centers regarding their performance on several measures of quality and resource use. Chapter 7 assesses the reliability of performance measurements for GPs by analyzing GP-level variation in performance across multiple years. Finally, chapter 8 summarizes the main findings by answering the research questions, provides some reflections on the results and discusses their relevance for the Dutch healthcare system, and offers some suggestions for further research.

CHAPTER 2

Key issues in the design of pay-for-performance programs

European Journal of Health Economics, 2013, 14(1): 117-131.



ABSTRACT

Pay-for-performance (P4P) is increasingly being used to stimulate healthcare providers to improve their performance. However, the evidence on P4P effectiveness shows inconclusive results. Flaws in P4P-program design may have contributed to this limited success. Based on a synthesis of relevant theoretical and empirical literature, this paper discusses key issues in P4P-program design and implementation. The analysis reveals that designing a fair and effective program is a complex undertaking. The following tentative conclusions are made: (1) performance is ideally defined broadly, provided that the set of measures remains comprehensible, (2) concerns that P4P may encourage "risk selection" and "teaching to the test" should not be dismissed, (3) sophisticated risk adjustment is important, especially for outcome and resource use measures, (4) involving providers in program design is vital, (5) group-level incentives are preferred over individual-level incentives, (6) whether to use rewards or penalties is context-dependent, (7) payouts should be frequent and low-powered, (8) absolute performance targets are generally preferred over relative performance targets, (9) multiple performance targets are preferred over single performance targets, and (10) P4P should be a permanent component of providers' compensation and is ideally "decoupled" from base payments. However, the design of P4P-programs should be tailored to the specific setting of implementation, and empirical research is needed to confirm the conclusions.

2.1 INTRODUCTION

In many countries, healthcare delivery is suboptimal. For example, McGlynn et al. (2003) have shown that in the United States (US) adherence to recommended care processes is near 50 percent. In the Netherlands, this is somewhat higher, but there is large variation among providers and among specific guidelines (e.g., Grol, 2001). Similar deficits were found in the United Kingdom (UK), Australia, and New Zealand (seddon et al., 2001). As a response, a multitude of strategies has been developed to spur improvements in performance. Pay-for-performance (P4P) is one of these strategies. In P4P, healthcare providers receive explicit financial incentives for reaching targets on predefined performance measures. The premise of P4P is that providers are responsive to financial incentives (Hillman et al., 1989; Donaldson & Gerard, 1989; Dudley et al., 1998; Gosden et al., 2000, 2001; Town et al., 2004) and that each of the commonest payment methods (i.e., fee-for-service, capitation, and salary) is not designed to stimulate good performance and separately creates incentives for undesired behavior. Given that performance measurements have become more accurate over the past two decades, it therefore seems appropriate to use financial incentives explicitly to stimulate improvements in performance. The main goal of P4P is to improve patient health outcomes while mitigating unintended consequences (such as increasing disparities). By contributing to better prevention and disease management, as well as by including expenses measures, if effective P4P could also mitigate cost growth.

P4P is now widely being applied in the United States (US) and the United Kingdom (UK) (Rosenthal et al., 2006; Baker & Delbanco, 2007; Roland, 2004) and is increasingly being implemented in many other countries (Rochon et al., 2006; Duckett et al., 2008; Gross et al., 2008; Buetow, 2008; Benavent et al., 2009; Lee et al., 2010). However, in contrast to what its popularity in practice suggests, P4P effectiveness has not been convincingly confirmed. A broad evidence base is lacking, and existing studies show mixed or inconclusive results (Petersen et al., 2006; Rosenthal & Frank, 2006; Christianson et al., 2008). Moreover, unintended and undesired effects of P4P have been demonstrated (Shen, 2003; Karve et al., 2008; Werner et al., 2008; McDonald & Roland, 2009; Campbell et al., 2009; Friedberg et al., 2010; Chen et al., 2010). Nonetheless, in general, the potential of P4P to improve performance remains undisputed. There is consensus that the way in which P4P is designed and implemented has important consequences for the incentives that physicians experience and how they might respond to them (Mehrotra et al., 2010a). As argued by several authors, the fact that P4P has not been very successful has partly been a consequence of flaws in program design (Rosenthal & Frank, 2006; Petersen et al., 2006; Rosenthal & Dudley, 2007; McDonald & Roland, 2009). Although the idea underlying P4P is simple, designing a fair and effective program is a complex undertaking involving many different aspects to consider.

The goal of this paper is to provide an overview of key issues in the design and implementation of P4P. Other authors have already provided important contributions in this area (Town et al., 2004; Young et al., 2005; Rosenthal & Dudley, 2007; Damberg et al., 2007; Conrad & Perry, 2009; Mehrotra et al., 2010a). However, this work typically addresses a selection of design elements, without discussing other potentially relevant aspects in detail. This paper synthesizes relevant theoretical and empirical literature as well as findings from the previous work into a single comprehensive overview. The first section discusses issues regarding the definition of performance and important prerequisites for preventing undesired behavior ("what to incentivize"). The next section deals with the question whether P4P should focus on individual physicians or groups of physicians ("whom to incentivize"). Finally, section three discusses consecutively whether programs should use penalties or rewards, the size of the incentive and the role of the base payment system, whether payment should be made for absolute performance or for relative performance, the frequency of payments, and the duration of the P4P incentives ("how to incentivize"). Throughout the paper, issues regarding incentive salience and provider participation are also discussed. The salience of the financial incentives incorporated in a P4P-program is an important predictor of the program's effect on behavior. If providers are aware of the program and the targets to be attained, and actually experience the incentives in daily practice, behavioral response is likely. Likewise, the willingness of providers to participate and their possibilities of "exit" determine the success of the program to a great extent.

2.2 WHAT TO INCENTIVIZE: HOW IS PERFORMANCE DEFINED?

2.2.1 Dimensions and measurement of performance

Depending on the goals of the stakeholders involved, programs will vary in how "good performance" is defined (Ittner & Larcker, 2002). Cost and utilization control were the main focus of early P4P-programs in the US (e.g., Moore et al., 1980), mainly because of the context in which they were implemented (pay-for-volume was the status quo), but also because measurement is relatively straightforward and the means by which savings were achieved (e.g., more prevention, less overtreatment) was also expected to be beneficial for the quality of care. More recently, however, purchasers have increasingly been using P4P to spur improvements in the quality of care. Quality is a multidimensional concept embodied in structures (e.g., having an up-to-date patient registry), processes (e.g., regularly checking the blood sugar levels of diabetes patients), and (intermediate) outcomes (e.g., optimal blood sugar levels for diabetes patients) (Donabedian, 1988). Although structures and processes are imperfect surrogates for outcomes, they are used frequently in P4P-programs because of the difficulty of measuring and risk-adjusting outcomes (Dudley et al., 1998). A related performance aspect is patient satisfaction or patient centeredness, which, although

clearly associated with quality of care, is not necessarily positively correlated with desired clinical processes and outcomes (Weyer et al., 2008).

The number and characteristics of included performance measures are likely to affect the eventual effect of the program on overall performance (Town et al., 2004). If a program only includes a few measures regarding one specific performance aspect (e.g., diabetes care), this could result in a disproportionate focus on a specific behavior (i.e., improving diabetes care). If, on the other hand, many different measures pertaining to many performance dimensions and aspects are included, the program may be too complex and providers may have difficulties in processing the incentives. Consequently, providers may not exhibit the desired behavior the purchaser wishes to stimulate (Town et al., 2004). Thus, a balance is needed between "narrow and shallow" and "broad and deep." It also seems important to combine objective (clinical) measures (e.g., adherence to clinical guidelines) with subjective measures such as patient satisfaction and continuity of care (Gibbons, 1998). Ultimately, the exact definition of "good performance" depends on the context in which the program is implemented.

In practice, measure sets are typically quite narrow, which mainly is a result of strict inclusion criteria such as consistency with other quality improvement activities, a firm evidence base, good psychometric properties, and availability of data at acceptable cost (Rosenthal et al., 2007b; Damberg et al., 2007; Sorbero et al., 2007; Baker & Delbanco, 2007). To minimize the burden and cost of data collection, many programs largely rely on claims data, which are easy and inexpensive to collect. However, claims data are not intended and often not suitable for generating performance information. To supplement claims data, purchasers may require providers to provide additional performance information based on extractions of medical records and by administering patient surveys. However, extracting data from medical records is often time consuming and expensive. Also, it imposes substantially higher burdens on smaller practices than on larger ones, and increased reimbursement to support record reviews may be necessary (Landon & Normand, 2008). Information technology (IT) such as electronic medical records (EMR) may considerably reduce the cost and burden of data collection. Under the Quality and Outcomes Framework (QOF), a large national P4P-program in the UK, primary care practices receive substantial financial rewards for scoring well on a large number of performance measures (Roland, 2004). For each practice, performance data are extracted automatically via a uniform EMR and collated in a national database. This has several advantages, including complete and accurate data and improved possibilities for performing checks on self-reported data. In addition, because practices have ongoing insight into their performance and receive relative performance feedback, the system contributes to incentive strength. However, such a comprehensive IT infrastructure involves substantial investments. In the UK, primary care practices were largely compensated for health IT (Doran & Roland, 2010), but in other settings, this may not always be feasible and providers may have to share in the costs. An

option is to make the financial incentives conditional on IT adoption, which is increasingly being done in many P4P-programs. In the US, EMRs are increasingly used for the purpose of data collection, although still on a relatively small scale (Sorbero et al., 2006; Damberg et al., 2007; Landon & Norman, 2008).

2.2.2 Risk adjustment

Patients are not randomly distributed across providers, and there is no level playing field regarding the attainability of performance targets. Consequently, providers who perform above average may be classified as average or even below average, whereas providers who perform below average may be classified as average or even above average, purely as a result of differences in casemix. This provides a strong incentive for providers to select healthy and compliant patients and to avoid severely ill and noncompliant patients. Adequate risk adjustment reduces this perverse incentive (in this paper, "risk" refers to patient characteristics that directly or indirectly affect providers' performance but cannot be influenced by providers, including sociodemographic characteristics and severity of disease). In general, outcome measures require more sophisticated risk adjustment than process measures because the latter are more within providers' control. It is therefore not surprising that structural and process measures are used much more often in current P4P-programs than outcome measures. Indeed, in addition to a lack of routinely available clinical data, the limited use of outcome measures in practice stems from concerns among purchasers about the adequacy of risk-adjustment models (Damberg et al., 2007). Over the years, risk adjustment has become more sophisticated. As a result, it is increasingly being applied in P4P-programs, and its importance is widely underscored (Sorbero et al., 2006; Rosenthal et al., 2007b; Damberg et al., 2007).

Because risk adjustment contributes to a fair allocation of performance payments, it may increase provider support and participation. However, as noted by Christianson et al. (2007:19), "application of risk-adjustment techniques is often controversial. They can be difficult to explain and require sophisticated statistical methods to implement, which can cause [providers] to view them as arbitrary 'black boxes' and to be suspicious of their validity." Although transparent application and communication can mitigate these problems, even sophisticated risk-adjustment models may be insufficient to effectively remove incentives for selection (Hofer et al., 1999). In addition, because of the complexity of patient care, providers are likely to have better information about their patients than the most detailed database and may therefore still be able to improve their performance through selection (Dranove et al., 2003). Moreover, even if information on outcome quality can be routinely collected and risk adjustment would be adequate, these measures will often not be useful for P4P purposes because of low reliability as a result of small sample size (Krein et al., 2002; Nyweide et al., 2009). In addition to clinical outcomes, this will often also hold for measures of utilization and resource use (Hofer et al., 1999; Krein et al., 2002; Nyweide et

al., 2009; Mehrotra et al., 2010b). Therefore, one should be cautious with including outcome and resource use measures in P4P-programs. They should only be considered for inclusion if risk adjustment is sophisticated and if sample size is large enough to yield sufficient reliability. Yet, other strategies may still be necessary to minimize incentives for selection. In the UK, for example, performance measures (including outcomes) in the QOF are not risk-adjusted. Instead, for each measure, practices are allowed to exclude patients (e.g., those who are noncompliant) from the measurements. While this provides practices with a tool to increase income by excluding "difficult" patients or patients for whom targets had been missed rather than because of an appropriate reason, there is little evidence of inappropriate use of "exception reporting" (Doran et al., 2008a; Gravelle et al., 2008), although more research is needed to confirm this. Extensive inspections and penalties for fraud may have contributed to preventing this behavior.

Risk selection is not just a theoretical concept. Hofer et al. (1999) showed empirically that the easiest way for physicians being profiled on the blood sugar levels of their diabetes patients to have a substantial improvement in performance would be to deselect from their panel those patients with high blood sugar levels in the previous year. They demonstrate that if physicians with the worst performance in the prior year manage to deselect the one to three patients with the highest blood sugar levels, they would in most cases achieve substantially improved performance than average in the current year. In their analysis, about half of this improvement was due to patient selection. Shen (2003) investigated whether a performance-based contracting system for nonprofit providers of substance abuse treatment resulted in providers selecting less severely ill clients in their treatment program in order to improve their performance. After implementation of the system, the proportion most severe patients increased in the control group whereas in the intervention group this proportion decreased, providing a clear indication that providers engaged in selection. Another study showed that public reporting of hospital- and surgeon-specific risk-adjusted mortality of coronary artery bypass grafting (CABG) patients led to substantial selection by providers (Dranove et al., 2003): relative to patients in states without such public reporting, a significant decline in the severity of illness of CABG patients was observed in the two intervention states. McDonald and Roland (2009), comparing unintended consequences of large P4P-programs in California and England, found that the inability of Californian physicians to exclude individual patients from performance calculations caused frustration and led some physicians to deter noncompliant patients. Finally, in Taiwan, a national P4P-program for diabetes includes two unadjusted outcome measures. Because providers are free to choose which patients to enroll in the program, they have an incentive and a clear tool for selection. Indeed, older patients and patients with greater disease severity or comorbidity were more likely to be excluded from the program than younger patients and patients with less disease severity or comorbidity (Chen et al., 2011).

2.2.3 Teaching to the test

As a result of explicitly targeting specific aspects of care, P4P incentives may cause providers to focus disproportionately on those aspects of care that are measured and incentivized, possibly to the detriment of other, often more indeterminate aspects that are not (easily) measured (Holmstrom & Milgrom, 1991; Gibbons, 1998). In the literature, this is known as "teaching to the test", which may occur especially in multitasking environments (such as medical care). However, it is also possible that rewarding specific behaviors will lead to positive spillover effects on unincentivized aspects of performance. As noted by Mullen et al. (2010:66), "which response dominates will depend on the technology of quality improvement in medical practices, about which little is known. For example, screening and follow-up measures, such as mammography and hemoglobin A1c (blood sugar) testing for diabetics, may both be increased by a general improvement in information technology, such as a computerized reminder program, despite differences in administration technique and patient populations." In an empirical analysis of performance data of physician medical groups contracting with a large network health maintenance organization, Mullen et al. (2010) did not find evidence of positive or negative spillovers on unincentivized aspects of care, although some rewarded performance measures improved. Another US study (Glickman et al., 2007) found that among hospitals participating in a quality-improvement program, P4P had limited incremental impact on quality of care for acute myocardial infarction (AMI). In addition, no evidence was found that P4P had an adverse impact on improvement in processes of care for which there were no financial incentives. Two studies have addressed teaching to the test with respect to the QOF in the UK, with more than 130 measures in about 30 different areas the most comprehensive P4P-program in the world. Steel et al. (2007) found neither improvement nor deterioration in unincentivized conditions. However, Campbell et al. (2009) found a positive spillover effect on unincentivized aspects of an included condition, a deterioration of unincentivized aspects of two other included conditions (while incentivized aspects continued to improve), and a reduction in the continuity of care immediately after the QOF was implemented. Most current P4P-programs include less performance domains and much smaller sets of measures per domain than the QOF. In the US, while purchasers underscore the importance of a broad set of measures, sets are typically quite narrow (Rosenthal et al., 2007b; Baker & Delbanco, 2007). However, the somewhat stronger evidence of teaching to the test in the UK may also have been a result of the magnitude of rewards, which can be up to 30 percent of practice revenues. Rewards of this size may have "crowded out" GPs' intrinsic motivation, leading to negative spillover effects on unrewarded performance aspects (see below).

Although evidence of teaching to the test is limited, theory and practice suggests that the risk cannot be ignored and that unincentivized aspects should be monitored. As Mullen et al. (2010:86) argue, "even though we fail to find conclusive evidence of negative spillovers (…), the concern that P4P encourages 'teaching to the test' should not be dismissed. Given

the complex and largely unobservable nature of healthcare quality, we can only study some potential unintended consequences but we cannot confirm or reject the existence of all such effects (…). The negative incentives of P4P-programs still exist and should be taken seriously given evidence that providers do indeed respond to incentives." Negative spillovers can be mitigated by adopting a varied set of performance measures. This also contributes to incentive salience because the fraction of providers' patients to which the incentive applies is large. The set should at least incorporate "high-impact" measures, that is, measures pertaining to conditions with a high prevalence and/or disease burden. However, especially regarding clinical quality, lack of data often hampers inclusion of important measures. Therefore, if P4P is to contribute to improved patient outcomes, efforts should continue to focus on creating reliable and easy to apply methods for extraction and validation of patient-level data, and the merits of IT for these purposes should be explored further. As noted, however, one should be cautious that the program does not become too complex because individuals often have difficulties in processing complex decisions tied to financial incentives (Mehrotra et al., 2010a). Yet, in P4P it is particularly important to carefully monitor the more indeterminate aspects such as continuity of care and patient centeredness (both core features of good patient care) as these aspects will be among the first aspects that may be neglected when the extrinsic motivation of providers is emphasized (Marshall & Harrison, 2005). However, adequate measurement of these aspects is often more difficult and more expensive than measurement of clinical processes or resource use. Consequently, even monitoring may be not feasible. It is therefore important that providers are actively involved in measure selection and program design.

2.2.4 Providers' intrinsic motivation

Financial incentives based on productivity and financial results may have a negative impact on physician satisfaction whereas incentives based on quality and patient satisfaction may positively affect physician satisfaction (Grumbach et al., 1998). A possible reason may be that the former goals are less aligned with physicians' professional norms and values and are therefore less acceptable to them (Dudley et al. 2004). Such dissatisfaction mitigates the likelihood of a desired response and increases the likelihood of undesired behavior because the incentives may "crowd out" providers' intrinsic motivation to provide high-quality care. Research has shown that extrinsic incentives may indeed result in outcrowding (Deci et al., 1999). Although this literature primarily pertains to educational settings, the idea seems to apply particularly well to physicians who are believed to be driven for a large part by professionalism and have been socialized to put the interest of their patients above anything else (Freidson, 2001). The introduction of P4P could then play a trivializing role regarding the nonfinancial motivation (Berwick, 1995; Christianson et al., 2008). However, this is also true for the base payment system. Moreover, outcrowding will be more significant as a result of base payments than of P4P because it involves larger sums of money. P4P aims to correct

perverse incentives emanating from base payments and in order to make sure that these are not exacerbated, insight into how outcrowding occurs is required.

According to Marshall and Harrison (2005:5), outcrowding may occur in two ways: "firstly, external incentives may impair self-determination, resulting in a shift in the locus of control and the resulting loss of professional autonomy. Secondly, external drivers may damage self-esteem, resulting in the perception that professionalism is no longer valued." In addition, when extrinsic incentives are provided for performing a particular task, individuals tend to view that task as irksome or hard to perform (Freedman et al., 1992). Outcrowding is more likely to occur in creative tasks, in overly bureaucratic schemes, and in the more indeterminate aspects of professional practice (Marshall & Harrison, 2005). To prevent outcrowding, purchasers should make sure that the incentives are viewed as legitimating and reinforcing of internal motivators (Frey, 1997; Conrad & Christianson, 2004). If the incentives are aligned with providers' internal value framework, the likelihood that the program will be successful increases (Marshall & Harrison, 2005). Alignment may be achieved by focusing on the more technical aspects of performance and by closely involving providers in program design and in developing, selecting, and validating the performance measures for which they will be held accountable (Young et al., 2005). All else equal, P4P may then compensate the loss in intrinsic motivation that occurs as a result of base payments. Outcrowding can also be mitigated by making participation voluntary. Even when providers are actively involved in the development process, imposed participation may be perceived as a loss of autonomy, which in turn may lead to undesired behavior. However, if participation is selective, performance differences among providers may be created, sustained, and/or enlarged, which may lead to and/or increase inequalities in access to high-quality care. Clearly communicating to providers the program's characteristics and potential merits and actively involving providers in program development mitigates this problem. But even if a high participation rate can be attained, reaching consensus will often be a long and difficult process and inevitably involves making compromises, which may result in diverging definitions of performance. It is therefore important that the program is designed such that it stimulates desired behavior and that agents (i.e., the healthcare providers) are incentivized to act in the interests of the principal (i.e., the purchaser).

2.2.5 Summary

In sum, performance is ideally defined broadly, provided that the set of performance measures remains comprehensible for providers. The set should at least incorporate "high-impact" measures of different performance dimensions, and the more indeterminate aspects should be monitored. However, measures should conform to strict criteria before they can be used in P4P-programs, including good psychometric properties and availability of complete and accurate data. Outcome and resource use measures should only be included if risk adjustment is sophisticated and if sample size is large enough. However, even then providers

may have incentives for selection, necessitating other risk-mitigating measures. To prevent undesired behavior, it is vital that providers are actively involved in program design, though monitoring for undesired consequences and structured feedback to providers about such consequences occurring will likely remain necessary.

2.3 WHOM TO INCENTIVIZE: INDIVIDUALS OR GROUPS?

For performance issues that can be improved most efficiently through group effort (e.g., those that require collective action), incentives should be directed toward the group level. For the extent to which issues are under individual physicians' control, incentives may be most effective when targeted at individuals (Rosenthal & Dudley, 2007; Gaynor et al., 2004; Town et al., 2004). However, health care is increasingly provided in settings in which professionals from various medical disciplines cooperate in the treatment of patients. Consequently, it is becoming increasingly difficult to ascribe a "good performance" to an individual practitioner. Therefore, it would often be logical to target P4P at groups of physicians rather than individual physicians. (In this paper, we follow Town et al.'s (2004:99) definition of a medical group, that is, an actor in which two or more physicians operate as a partnership, have a common profit center, pool income, pay expenses, and distribute profits to group members. Another possibility is an arrangement in which physicians retain their own income and contribute to common office expenses). In group incentives, in which the financial risk is shared among the physicians in the group, performance is affected through an effect on group culture, selection and socialization of new members, sharing of information, peer pressure, and collaboration (Town et al., 2004). They may be more effective than individual incentives because inefficiencies in health care are often viewed to be a result of a failure of systems (Institute of Medicine, 2001; Enthoven & Tollen, 2005) and because of enabling factors like assistance of other professional and support staff (Young et al., 2005), collaboration, peer review, and available infrastructure. However, it is important to assess whether and how incentives are passed along to group members (Frølich et al., 2007). When such mechanisms are not (effectively) in place, the effect of the program may be mitigated because the incentive to improve performance experienced by individual group members is weak (Alchian & Demsetz, 1972; Gaynor & Gertler, 1995). Free-riding on the efforts of peers may then be difficult to detect and penalize. As noted by Town et al. (2004), problems of free-riding will increase as group size increases because it is more difficult for social influence and monitoring to operate through peer relationships. The problem will be most pronounced in large groups where significant interdependencies among group members are absent. Peer pressure may then not be sufficient to offset the dilution of incentives that naturally occurs in group settings (Gaynor et al., 2004). In addition to diluted incentives, a potential disadvantage of directing P4P at groups from a purchaser perspective is that

groups generally have more bargaining power than individuals and are more effective in defying or negotiating the terms of external incentive programs (Oliver, 1980; Town et al., 2004). Based on interviews with sponsors of hospital P4P-programs in the US, Damberg et al. (2007) noted that in negotiating the terms of their P4P contracts, sponsors experience greater bargaining power of hospitals compared to individual physicians. Finally, behavior may be hard to change in groups because of a shared culture. However, group culture may also present an advantage in that achieved performance improvements are likely to be sustained as a result of peer pressure and socialization of new members.

Individual and small-group incentives have an important practical disadvantage. The success of a P4P-program depends on the reliability of the performance measures used, which requires sufficiently large panels of patients (Landon et al., 2003). Especially when variation in performance attributable to the physicians is small, which tends to be the case particularly for outcome and resource use measures, large numbers of patients per measure are needed to generate reliable measurements and comparisons (Krein et al., 2002). Patient panels of individual physicians and small groups are typically too small to measure performance reliably (Hofer et al., 1999; Krein et al., 2002; Huang et al., 2005a; Scholle et al., 2008; Nyweide et al., 2009; Adams et al., 2010a). Thus, if P4P targets individual physicians or small groups, measured performance differences may reflect to a significant degree random variation (Christianson et al., 2008; Mehrotra et al., 2010b), possibly resulting in misclassification of providers and incorrect allocation of incentive payments (Nyweide et al., 2009; Adams et al., 2010a). Constructing composite scores could enhance low reliability due to small sample size per measure (Caldis, 2007) and has the additional advantage that it hampers gaming behavior. However, it requires rich data and complex calculations (e.g., for determining the relative weights of individual measures) and considerations (Reeves et al., 2007). Also, composites provide less actionable information on quality than individual measures and do not guarantee reliability levels sufficient to enable inclusion of large shares of providers (Scholle et al., 2008; Smith et al., 2013). Aggregating data across purchasers may also be an option (Higgins et al., 2011). However, for several reasons (e.g., possible violations of anti-trust regulation, technical difficulties, patient privacy), this does not occur on a large and systematic scale yet (Rodriguez et al., 2012).

On balance, group incentives seem preferred over individual incentives, mainly because performance profiles are more likely to be reliable (Huang et al., 2005a). However, this may not always be the case because there may be less variation among groups than among individuals (Smith et al., 2013). In addition, when performance is compared across groups, it is important that there are sufficient numbers of physicians in each comparison group to detect meaningful differences. Lack of adjustment for clustering at the physician level (in addition to adjustment for patient characteristics) could lead to overestimation of the statistical significance of differences between groups (Greenfield et al., 2002). In addition, groups differ considerably in size and composition, and it is unclear how to treat the many provid-

ers working in small practices with small numbers of patients for many measures (Mehrotra et al., 2010b; Landon & Normand, 2008). Although health care is increasingly provided in group settings, small practice settings will likely remain important, necessitating strategies to facilitate inclusion of small practices (Landon & Normand, 2008). As methods for data aggregation and constructing composite scores continue to evolve (Higgins et al., 2011), it will be increasingly possible to include measures with small sample size and to target P4P at small groups. Of note, purchasers should be cautious in applying hybrid structures (e.g., using both group and individual incentives for a team with high interdependence among team members) because they have shown to perform worse than pure structures (Town et al., 2004), perhaps because they are less transparent and hence less salient.

2.4 HOW TO INCENTIVIZE: HOW IS THE PROGRAM STRUCTURED?

2.4.1 Rewards versus penalties

Because individuals generally weigh losses more heavily than gains, a larger behavioral response can be expected if individuals perceive the incentive as a (possible) loss as opposed to a (possible) gain (Kahneman & Tversky, 1979). This implies that withholds will be more effective in improving performance than positive bonuses. For example, withholding €1,000 from base payments with the possibility of releasing this amount if performance targets are met will elicit a stronger behavioral response than offering providers a €1,000 bonus for good performance (Damberg et al. 2007). However, research has shown that incentive schemes incorporating losses tend to be perceived as unfair and may result in negative reactions among those incentivized (Kahneman et al., 1986). Consequently, the program may not be acceptable to providers, and they may choose not to participate. This may especially be a problem if the bargaining power of the purchaser (e.g., an insurer) is relatively low and if providers can choose from among multiple insurers to contract with (Arrow, 1986). But even if providers can be convinced or enforced to participate, the behavioral response to financial penalties may not be the desired response. The prospect of a loss may cause physicians to behave opportunistically, and incentives for gaming and other undesired behavior may be large. (Not receiving a bonus from a pool of money available for performance improvement may also be perceived by providers as a financial penalty because their relative income position deteriorates, but negative reactions will likely be stronger under absolute financial penalties).

A possible way to still take advantage of the expected strong provider response to penalties while mitigating the likelihood of negative reactions is to combine rewards and penalties. For example, providers could be offered a choice between a €1,000 bonus for meeting targets and entering a deposit of €500 with the prospect of a €2,000 bonus (Mehrotra et

TABLE 2.1 Features of schemes adopting penalties and/or rewards

Scheme	Income increase or decrease possible?	Incentive strength	Likelihood of negative reactions
1. Penalties for poor performance only	Decrease only	High	High
2. Rewards for good performance only	Increase only	Moderate	Low
3. Penalties for poor performance, (larger) rewards for good performance	Both	High	Moderately high
4. Choice between 2 and 3 provided that the potential increase in income is larger in 3	Depends on choice	Moderately high	Moderately low

al., 2010a). In case the provider chooses the second option and fails to reach the target, he loses the deposit. Thus, providers are offered a choice between a possible increase in income without the possibility of a loss in income and a larger possible increase in income with the possibility of a loss in income. Such a scheme also provides insight into differences among providers in their expectations about their potential for performance improvement. Furthermore, it will likely be received positively by providers and increases the likelihood of high participation rates. Table 2.1 displays the features of four possible schemes.

Despite the advantages of using rewards, purchasers may still opt for using "old" money (e.g., redistributing money to high performers based on a generic reduction of base payments). They could argue that programs using rewards may not be sustainable and object to investing additional resources in settings with substantial inefficiencies (Christianson et al., 2008). It may be an option to use efficiency savings to finance the program. However, performance improvement will, at least in the short term, often be accompanied by cost increases because a substantial share of quality problems is related to undertreatment. Another option is to make use of inflation. Providers could be given the prospect they will at least receive their current absolute income in the next period and, if they reach certain performance targets, they will also receive a mark-up based on the general increase in price levels. In that case, the perceived decrease in income for low performers is relatively small. However, negative reactions cannot be ruled out. Thus, if positive incentives are not possible, the extent to which P4P will improve overall performance depends on whether providers can be convinced or enforced to participate and whether provider behavior can be effectively monitored and, if necessary, countered.

In practice, the use of negative incentives in P4P has declined rapidly. In the US, although withholds are still applied in 10-20 percent of current programs, more than 60 percent only use bonuses, mainly because of anticipated negative reactions and the importance being attached to a collaborative rather than a combative tone (Sorbero et al., 2006; Damberg et al., 2007; Baker & Delbanco, 2007). Also in other countries, P4P-programs typically only provide positive incentives.

2.4.2 Incentive size

All else equal, the higher the revenue potential for providers, the larger their response and the impact on performance, up to a certain point. Large incentives are salient and increase the likelihood that the costs of performance improvement, including the opportunity costs of not doing something else, are covered (Grossman & Hart, 1983; Young & Conrad, 2007; Conrad & Perry, 2009). These costs will vary by the base payment system and the set of performance measures, so the payment level required to achieve improvements is not a static figure (Christianson et al., 2007). In general, the relationship between incentive size and performance will be positive with diminishing marginal increases in performance above a certain payment level. This is because the marginal utility of income generally diminishes and because every unit of performance improvement will be harder to attain than the previous unit. Also, there is evidence that the reference-/target-income hypothesis is applicable to physicians (Rizzo & Blumenthal, 1996; Rizzo & Zeckhauser, 2003), suggesting that when physicians reach a certain income level, additional payment will not lead to further significant improvements. Large payments, therefore, need not necessarily be more effective than smaller payments. Although large payments may still be necessary to persuade providers to participate, compared to small payments they are more likely to impair providers' intrinsic motivation (Frey, 1997; Deci et al., 1999). Consequently, the likelihood of undesired behavior increases since positive net gains of this behavior are more likely. Monitoring for this behavior may be difficult and costly, so in determining incentive size purchasers will often be confronted with a trade-off between an increased (but at some point diminishing) impact on performance and reduced intrinsic motivation. Yet, if payment levels are set high enough, the positive effect on incentivized performance may be greater than would be obtained through intrinsic motivation alone (Damberg et al., 2007). This is illustrated by Gneezy and Rustichini (2000), who show empirically that in financial incentive schemes one should "pay enough or not pay at all." However, increasing incentive size to surpass the loss in intrinsic motivation is of course an imperfect solution that may not be sustainable and could lead to problems like teaching to the test (Holmstrom & Milgrom, 1991; Prendergast, 1999). Therefore, relatively low-powered payments seem to be preferred, provided that they are based on measures that are aligned with providers' professional norms and values.

Empirical research on the influence of incentive size is scarce. Hillman et al. (1998, 1999) suggest that the limited success of the programs they evaluated may have been due to the small bonus size, as well as short program duration (less than two years) and lack of physician awareness. Conversely, Mullen et al. (2010) found that a dramatic increase in payment size triggered behavioral response. They investigated whether movement in selected quality measures changed when in addition to PacifiCare (a large network health maintenance organization [HMO]in California that had been running its own P4P-program called QIP), five other health plans in the Integrated Healthcare Association (IHA) coalition adopted P4P using a common measure set. Implementation of the IHA program considerably in-

creased the size of potential bonuses for medical groups compared to what they could have potentially earned under QIP. The authors found that while the QIP alone had not been able to generate improvements in quality, some quality measures did improve after the other plans adopted P4P. Thus, the authors concluded that payment size matters (Mullen et al., 2010). Finally, in the UK QOF, which has been successful in improving performance in primary care, performance payments can be up to 30 percent of practice income (Doran & Roland, 2010). However, it is unclear to what extent observed improvements can be attributed to these generous payments. In addition, as shown by McDonald and Roland (2009: 123), the large financial incentives have likely changed the nature of the office visit: "The requirement to enter data into the electronic medical record to respond to the large number of targets was described as reducing eye contact, increasing time spent on data collection, and potentially crowding out the patient's agenda."

The opportunity costs of complying with P4P incentives (i.e., the gains forgone of doing the next best alternative) are determined by the base payment system (Frølich et al., 2007). Especially under fee-for-service payments, these costs can be substantial because time and effort put in improving performance cannot be used to treat patients. Opportunity costs can be mitigated by replacing base payments by performance-related payments. However, multitasking predicts that important performance dimensions will likely never be contractible and that mixed payment is appointed (Eggleston, 2005). Even if performance would be entirely contractible, even on outcomes, the optimal compensation scheme would often still have a component of income that is guaranteed because practice in health care is inherently uncertain and physicians tend to be risk-averse (Town et al., 2004). Performance-related payments will therefore be supplemental to base payments. In addition, it seems warranted to "decouple" P4P payments from base payment as much as possible (Mehrotra et al., 2010a). Augmenting base payment from €1,000 to €1,100 will generally elicit a smaller behavioral response than providing a separate €100 bonus because individuals perceive the difference between €0 and €100 as larger than the difference between €1,000 and €1,100. Without decoupling, the P4P payment may be perceived as negligible compared to the base payment and the behavioral response may be small (Thaler, 1985). However, decoupling adds to administrative complexity (Mehrotra et al., 2010a).

2.4.3 Absolute versus relative performance

Performance-related payments can be based on absolute performance (e.g., performing a foot exam for at least 90 percent of eligible diabetics), relative performance (e.g., belonging to the 10 percent of physicians with the highest rates of performed foot exams), and improvement in performance (e.g., large payments for large improvements with improvement weighted more heavily at higher performance levels than at lower levels). Absolute targets are transparent and will be more acceptable to providers than relative targets because they involve less uncertainty. However, in a system in which the same P4P-program is applied

uniformly to a large group of providers, absolute targets may not be very efficient because a substantial portion of bonus payments may be awarded to providers already at or above the targets. Furthermore, for improvement beyond targets and improvement not reaching targets, providers receive zero incremental payment (Rosenthal & Dudley, 2007). The goal gradient hypothesis predicts that a goal should be perceived attainable by providers; otherwise, little response can be expected (Heath et al., 1999). Similarly, little effort can be expected after the goal has been achieved. These difficulties can be solved by differentiating required performance targets across groups, depending on groups' baseline performance (for individual-level incentives, such an arrangement will probably not be feasible because of high transaction costs). For groups with low baseline performance, target and payment could be set relatively low, whereas for high-performing groups, target and payment could be set relatively high.

Relative schemes stimulate continual improvement. However, because they encourage competition, they may reduce collaboration and dissemination of best practices and may sustain performance gaps across providers (Rosenthal & Dudley, 2007). Furthermore, the behavior of competing providers is to a large extent beyond the individual provider's control but does influence that provider's ranking. The strength of the incentive may be limited because "type I errors (false positive rewards based on relatively poor performance of others) and type II errors (false negative penalties or foregone rewards because of relatively good performance of others)" are likely (Conrad & Perry, 2009:361). Moreover, compared with absolute targets, relative targets involve more uncertainty for providers regarding their possibilities and/or the efforts needed to become eligible for payment. Because individuals tend to be risk-averse, P4P-programs accompanying little uncertainty will be more appealing to providers and will therefore lead to higher participation rates than programs accompanying much uncertainty. Conversely, an advantage of a relative scheme over an absolute scheme is that the total amount of incentive payments is known ex ante (Rosenthal & Dudley, 2007), which gives providers the prospect of certain payment if targets are reached. In an absolute scheme, if more providers than expected reach the threshold(s), either new money has to be generated or payment per eligible provider has to be decreased. The former is exactly what happened in the QOF in the UK. By 2006-2007 (the third year), primary care practices on average scored more than 95 percent of the available points, which exceeded the predictions of the Department of Health, which had anticipated 75 percent attainment (Doran & Roland, 2010). While generating new money will be difficult, reducing payments will probably lead to negative reactions among providers and a reduced effect of the program in the future (Conrad & Perry, 1999). If there is not much flexibility in increasing the pool of incentive payments, the pool could be set to a maximum about which participating providers should be informed in advance.

Both relative and absolute schemes using single targets risk being resisted by providers because they explicitly create "winners" and "losers." Because providers may perceive

losing as a penalty, a single target scheme may provoke undesired behavior. As noted, this difficulty can be resolved by varying required (absolute) performance targets across providers, conditional on baseline performance. Another option is to confront all participating providers with a series of (absolute) targets with large payments for reaching high targets and low payments for reaching low targets. Such a scheme also rewards improvement. The downside of this approach is that the program may be viewed as unfair and demotivating by high performers. An option could be to choose a particular target as a starting point (e.g., 50 percent) and to increase payments as higher targets are reached. Providers with scores below 50 percent then get nothing or could be given a financial penalty. Another option is to eliminate targets altogether and to use a continuous gradient (Mehrotra et al., 2010a). Yet, a scheme using targets may be a stronger stimulus than a continuous scale because providers have clear goals to work toward.

Again, the QOF provides some (weak) empirical evidence. In the QOF, each performance measure has a lower (e.g., 40 percent) and an upper target (e.g., 90 percent). Between these targets, performance is measured on a continuous scale and practices earn more points for reaching higher performance levels. Improvements in the quality of care were most pronounced for GPs with the lowest scores, narrowing inequalities in quality of care, especially for chronic conditions (Doran et al., 2008b). This may well have been a result of the use of the continuous scale because even for the worst performers, the lower targets were often attainable and for them, improvements would entail large increases in income.

Alternatively, purchasers could opt for a system that rewards high-value care, provided by anyone (Rosenthal & Dudley, 2007:743). This can be achieved by "paying all providers an additional fee for each appropriately managed patient or for each recommended service [so that] every provider has an incentive to deliver the best care to each patient seen." Drawbacks of this approach (e.g., actuarial uncertainty for the purchaser) have to be traded-off against its advantages (e.g., simplicity and certainty for providers, as well as less incentives for risk selection compared to explicit targets). A recent study by showed that within a health plan that implemented a "piece-rate" P4P-program (i.e., providers received a payment for each patient meeting a performance benchmark), childhood immunization rates increased significantly more than among health plans that did not (Chien et al., 2010). Also, the program did not exacerbate disparities nor have a negative effect on children with chronic conditions.

In sum, differentiating required absolute performance targets across providers and/or applying a series of tiered absolute targets, possibly combined with additional fees for each appropriately managed patient, are preferred over a uniform, single threshold system and schemes using relative targets. Advantages of combining different approaches within a single program should be weighed against increased complexity and reduced incentive salience.

2.4.4 Frequency of payments

Providing a monthly €100 bonus with an additional payment of €500 based on overall improvement will generally be a more effective lever of improvement than a single €1,700 bonus at the end of the year. This is because people tend to discount future gains by a certain rate, which increases with the length of the delay (Frederick et al., 2002). In addition, people generally discount losses at lower rates than gains and large outcomes more than small outcomes (Thaler, 1981; Frederick et al., 2002). Thus, minimizing the time lag between care delivery and payment is warranted, especially when large payments are used, also because the costs of improving performance are typically incurred without much delay. A high frequency becomes even more important when providers experience uncertainty regarding the net gains of improvement efforts (as with relative targets) because, compared to schemes involving little uncertainty, possible gains will be discounted at higher rates. A second reason why a high payment frequency is important is that in risk-averse people, each additional unit of income leads to a smaller increase in utility than the previous unit. A large lumpsum payment will likely be less effective than a series of smaller, more frequent payments because each payment is judged as a new gain rather than an addition to the previous gain (Thaler, 1985; Damberg et al., 2007). Finally, a high payment frequency increases incentive salience.

In practice, data collection and validation may considerably delay payments, and long performance periods may be necessary to yield sufficient reliability. In a randomized experiment, Chung et al. (2009) investigated whether the impact of P4P increases when payments are provided quarterly instead of annually. They found no difference between the two trial arms in average quality score or in total bonus amount earned. However, physicians also received quarterly performance feedback, and the authors were unable to disentangle the effects of P4P and feedback. Also, regardless of the payment frequency, the size of the incentives may have been too small to elicit a noticeable impact on performance (bonuses were potentially 2.5 percent of the average physician's annual income), although this was not specially examined.

Clearly, for performance on outcomes that occur in the long term, a high payment frequency is not possible. In that case, P4P-programs will have to resort to structural and process measures, as well as to general measures like patient experience, which can be measured on a more regular basis. At least in theory, for these measures, a high payment frequency contributes to incentive strength. This does not imply that P4P can only be used for short-term objectives. For example, in long-term contracts with hospitals, payment could be linked to five-year mortality for different conditions. However, for specific types of care (e.g., rehabilitation and preventive care) P4P will not often be linked to clinical outcomes because they occur too far in the future. Instead, other types of outcomes may be included such as patient-reported outcomes or, regarding rehabilitation, patients' general abilities to independently perform activities of daily living.

2.4.5 Program duration

As noted by Town and colleagues, expectations about the future stability of new incentive schemes may influence whether providers will be responsive to these schemes. The decision to invest in quality improvement (e.g., employing an expensive IT infrastructure) requires making projections about future payment rates and expectations about return on investment (Town et al., 2004). Thus, the duration of the program as well as providers' expectations thereof seem important predictors of its effectiveness. Programs that are perceived as a stable systemic change will probably be more effective than programs that are perceived as a temporary effort. In addition, the effects of external rewards tend to last only through the period of incentive delivery; as soon as the scheme is abolished, performance may revert to the baseline level (Deci et al., 1999; Conrad & Perry, 2009). P4P aims to counterbalance perverse incentives in the base payment system (e.g., the incentive to do more than necessary in a fee-for-service system), so abolishing P4P incentives would mean that providers are confronted again only with the base payment incentives. Thus, once implemented, P4P payments ideally remain a component of providers' compensation. However, it is unclear whether programs using solely new money (generated through efficiency savings or otherwise) are sustainable in the long run.

The frequency of turnover of performance measures, that is, the duration of incentivizing specific aspects of performance within the program, is also of relevance (Young et al., 2005). A high frequency can be demoralizing for providers, especially when measures in which substantial effort has been put are replaced as soon as targets are reached. Yet, periodic reevaluation of measures will be essential, also from an efficiency viewpoint; it may not make sense to continue using measures for which performance has reached a plateau. In that case, replacing and/or updating measures are warranted, also because variation in performance may have become too small to measure performance reliably and to discriminate across providers (Krein et al., 2002; Scholle et al., 2008).

2.5 DISCUSSION

This paper provides an overview of key issues in the design and implementation of P4P-programs by synthesizing theoretical and empirical literature. The design of P4P-programs is important since it determines the way in which providers' behavior is influenced. To prevent undesired behavior, careful consideration of how the incentives are framed is vital, especially in multitasking environments (Holmstrom & Milgrom, 1991). Although the idea underlying P4P is simple, this paper has shown that designing a fair and effective P4P-program is a complex undertaking that requires consideration of many interrelated aspects and potential pitfalls. Nonetheless, several tentative conclusions can be drawn, which are summarized in Table 2.2.

TABLE 2.2 Conclusions with respect to P4P-program design and implementation

What to incentivize
- Performance is ideally defined broadly, provided that the set of measures remains comprehensible
- Concerns that P4P may encourage "risk selection" and "teaching to the test" should not be dismissed
- Outcome and resource use measures should only be used with adequate risk adjustment and sufficient sample size
- In addition to risk adjustment, other strategies to mitigate incentives for risk selection may still be necessary
- Measure sets should at least incorporate "high-impact" measures. The less technical / more indeterminate aspects of care such as patient satisfaction and continuity of care are ideally also included or at least regularly monitored
- P4P incentives should be aligned with professional norms and values; it is therefore vital that providers are actively involved in program design and in developing, selecting, and maintaining the performance measures
- Monitoring, feedback, and information technology are important in preventing undesired provider behavior

Whom to incentivize
- Group-level incentives will typically be preferred over individual-level incentives mainly because performance profiles are more likely to be statistically reliable as a result of larger numbers of patients
- Individual-level or small-group incentives as well as using measures with small available samples sizes will become increasingly feasible as methods for constructing composite performance scores continue to evolve
- One should be cautious with applying schemes incorporating both individual- and group-level incentives
- Participation is ideally voluntary provided that broad participation among eligible providers can be realized

How to incentivize
- Whether rewards or penalties should be used is context-dependent. Offering providers a choice among schemes also including penalties may be a viable option
- Although increasing the size of the incentive increases its strength (up to a certain point), relatively low-powered incentives are preferred, provided that providers' opportunity costs of improving performance are covered
- Provider-specific absolute performance targets and/or a uniform series of absolute targets, possibly combined with piece-rates for each appropriately managed patient, are preferred over single targets and relative targets
- The time lag between care delivery and payment should be minimized
- P4P should be a permanent component of provider compensation, and is ideally decoupled from base payments
- Performance measures should be reevaluated periodically and regularly be replaced or updated (as necessary)

However, conclusions on the appropriateness of design are inherently context-dependent; judgment about whether a particular P4P-program is designed appropriately will vary according to the setting in which it was implemented. For example, when providers are capitated, payment can be relatively small because, all else equal, the opportunity costs of improving performance are low compared to when providers are paid through fee-for-service. In addition to the base payment system, other relevant contextual factors are the characteristics of the practice environment (e.g., the level of IT); whether P4P is implemented in a single-purchaser healthcare system or in a system with multiple (competing) purchasers and, in case of the latter, the extent to which there is overlap in provider networks (much overlap may result in conflicting incentives for individual providers and increased complexity in provider decision-making); whether P4P is implemented in a system in which financing and delivery of care are integrated (such as HMO-like entities in the US, Israel, and Switzerland) or in a system with a purchaser-provider split (in an integrated system, P4P would be enacted by the organization's management, which likely has more possibilities to directly influence providers' behavior and align providers' incentives than purchasers operating more or less independently from providers); whether providers have fixed patient panels (if not, computerized algorithms are necessary to attribute care to pro-

viders and it will be more difficult to generate reliable performance profiles); whether there are concurrent improvement efforts (e.g., public reporting) targeting the same or different performance aspects; and the legal environment (e.g., data aggregation across competing purchasers may a violation of antitrust regulation). Recently, research has begun to address the influence of context (e.g., McDonald et al., 2009; Van Herck et al., 2010; Sutton et al., 2012). As shown in this work, this influence is likely to be substantial.

Several difficulties mitigate the strength of our conclusions. First, given a particular context, appropriate design choices may conflict. For example, group incentives and a broad measure set with outcome measures will often be preferred over individual incentives and measure sets not incorporating outcomes. However, as this paper has shown it is important to minimize provider uncertainty. For the individual provider, uncertainty regarding the net gains of improvement efforts increases when the incentive is targeted at the group level and when perceived possibilities for performance improvement decrease as a result of adding outcome measures to the measure set. Similarly, this paper has argued that using a tiered series of absolute targets is preferred over using a single target. However, such a scheme also adds to complexity, which may dilute incentive strength since individuals typically have difficulties in processing complex decisions tied to financial incentives (Mehrotra et al., 2010a). Second, practical difficulties may impede appropriate design. For example, where individual incentives are preferred, small sample sizes may necessitate targeting groups or aggregating scores. Similarly, although minimizing the time lag between care delivery and receipt of payments is warranted, data collection and validation are often time consuming and could result in payment coming long after the period of care delivery. Third, empirical evidence regarding the influence of specific design choices in practice is scarce. As a result, the weight of different design choices in terms of incentive strength is largely unknown. In particular, several authors have called for more research investigating specifically the "dose-response" relationship in P4P (Petersen et al., 2006; Frølich et al., 2007; Christianson et al., 2008; Mehrotra et al., 2009). Until further empirical research on these specific topics becomes available, lessons will have to be drawn from applications of P4P in practice. However, although evaluation studies may provide valuable information, without explicitly examining design issues it will be difficult to isolate the influence of specific design choices on P4P performance. In addition, as noted by Petersen et al. (2006) and Frølich et al. (2007), details on program design are generally not well documented, which mitigates the relevance of such studies for these purposes even more. Finally, there are important limitations in the interpretation of the theories applied in this paper for predicting provider behavior. For example, the theories mainly describe the behavior of individuals, not groups of individuals or organizations (like hospitals). The impact of P4P-program design on provider behavior may be different when groups or organizations are regarded (Damberg et al., 2007).

2.6 CONCLUSION

Designing a fair and effective P4P-program is a complex undertaking. This complexity and the limited effectiveness thus far cast serious doubt on whether P4P can be cost-effective. In addition to the performance payments themselves, data collection and validation as well as payment calculation likely involve significant transaction costs. Therefore, adequate evaluations of P4P-programs would not only assess the impact on quality but also include comprehensive cost analyses. However, a recent review identified only nine economic evaluations of P4P-programs and concluded that current evidence is insufficient to support P4P cost-effectiveness (see chapter 4). Nonetheless, P4P may be able to mitigate cost growth through better prevention and disease management and through inclusion of expenses measures. Recently, purchasers have begun to incorporate such measures in their P4P-programs (Institute of Medicine, 2007; Robinson et al., 2009). Yet, empirical research investigating the influence of specific design choices and contextual factors is needed to enable fine tuning of P4P-programs tailored to the setting of implementation. In the meantime, it would be sensible if purchasers would (continue to) consider other (non-financial) improvement strategies in their efforts to achieve more value for money.

CHAPTER 3

Pay-for-performance in health care: an international overview of initiatives

Medical Care Research and Review, 2012, 69(3): 251-276.

ABSTRACT

Pay-for-performance (P4P) has become a popular approach to performance improvement in health care. Most of the literature on P4P has focused on the United States and there is limited insight in the characteristics of major programs initiated in other countries. This paper systematically describes and reviews P4P-programs initiated outside the United States. The literature search identified thirteen programs initiated in nine different countries. Although the programs share many similarities, they differ in several important respects, also when compared with the typical P4P-program in the United States. In addition, there are clearly possibilities to increase incentive strength and minimize incentives for undesired behavior. In part, observed heterogeneity will be a consequence of contextual differences, but design choices often also seem to be made arbitrarily. In designing their programs, purchasers are hampered by limited knowledge about the influence of specific design choices and effective strategies to mitigate undesired behavior.

3.1 INTRODUCTION

Rising healthcare expenditures and deficiencies in the quality of care emphasize the need to increase efficiency in health care. Pay-for-performance (P4P), a payment approach in which healthcare providers receive explicit financial incentives for reaching targets on predefined performance measures, is considered a promising strategy to spur necessary improvements. The premise of P4P is that providers are responsive to financial incentives and that each of the commonest payment methods (i.e., fee-for-service, capitation, and salary) separately creates incentives for undesired behavior while none of them are designed to stimulate good performance. Given that performance measurements have become more accurate over the past two decades, it seems appropriate to use financial incentives explicitly to stimulate improvement. In recent years, interest in applying P4P in health care has greatly increased. In the United States, P4P is widely applied by public and private purchasers. In addition, P4P is increasingly being applied elsewhere, including in Canada, Australia, and several European countries. To date, most of the P4P literature has focused on the United States (US). There is now insight in how US programs are typically designed and how they have developed. In contrast, there is limited insight in the features of major P4P-programs initiated in other countries. In this paper, we aim to describe P4P-programs that have been initiated outside the US in terms of key design elements. Careful consideration of the design of P4P is important since inadequately designed programs may result in undesired provider behavior. Several authors have argued that the limited effectiveness of P4P (Christianson et al., 2008; Rosenthal & Frank, 2006) has partly been a result of flaws in design (Rosenthal & Frank, 2006; Petersen et al., 2006; Institute of Medicine, 2007; Rosenthal & Dudley, 2007; Christianson et al., 2008; McDonald et al., 2009). A second objective of this paper is therefore to assess the extent to which programs are designed appropriately.

3.1.1 New contribution

If P4P is to achieve desired results, careful consideration of program design is vital. The inconclusive evidence on the effectiveness of P4P suggests that design has not been optimal. In this respect, insight into the design of current P4P-programs would be useful in improving P4P performance. With the exception of the United States, there has been no comparative investigation of the characteristics of major programs initiated across the world. In this paper, we aim to fill this gap by systematically describing and critically reviewing non-US P4P-programs using a theory-based organizing framework. For purchasers and policy makers this information will be of interest as it provides lessons from experiences with P4P in practice and enables comparison of typical design in different settings. The paper proceeds as follows. The next section provides an overview of key elements of appropriate P4P-program design. Subsequently, after explaining the methods and search strategy, identified programs are described and assessed on the appropriateness of their design. The final section discusses the results.

3.2 KEY ELEMENTS OF P4P-PROGRAM DESIGN

The structure of a P4P-program has important consequences for how providers experience the incentives and how they might respond to them (Mehrotra et al., 2010a). Key elements of program design can be divided into three categories: what to incentivize, whom to incentivize, and how to incentivize. Table 3.1 summarizes the most important conclusions, based on chapter 2. It must be emphasized that these conclusions are largely based on theory and there is limited empirical evidence on the influence of specific design choices in practice.

TABLE 3.1 Conclusions with respect to P4P-program design and implementation

What to incentivize
• Concerns that P4P encourages "risk selection" and "teaching to the test" should not be dismissed
• Performance is ideally defined broadly. The more indeterminate aspects are ideally included or monitored
• Measures should be included only if risk adjustment is sophisticated and sample size is sufficient. Other risk-mitigating measures may still be necessary to prevent selection
• It is vital that providers are actively involved in program design and the selection of performance measures
Whom to incentivize
• On balance, group incentives are preferred over individual incentives, mainly because of larger sample size
• Inclusion of solo and small-group practices should be facilitated
• Participation is ideally voluntary provided that broad participation among eligible providers can be realized
How to incentivize
• Whether positive or negative incentives should be used is context-dependent
• Combining positive and negative incentives may be a viable option
• Payments should at least cover the (opportunity) costs of improving performance
• Frequent, low-powered payments are preferred over large lumpsum payments
• Absolute targets and piece-rates are preferred over relative targets
• Multiple targets and differentiated targets across providers are preferred over uniform single targets

3.2.1 What to incentivize: definition of performance

The set of performance measures determines a program's eventual effect on performance (Town et al., 2004). If the set includes only a few measures, providers may focus disproportionately on incentivized performance ("teaching to the test") and unincentivized aspects may be neglected, especially the more indeterminate aspects such as continuity of care and patient satisfaction. Therefore, a broad and varied set seems important. However, measures should conform to strict criteria if they are to be used for the purpose of P4P. In particular, especially outcome and resource use measures should be adjusted for relevant patient characteristics. However, sophisticated risk adjustment is complex and not necessarily sufficient to prevent risk selection (Hofer et al., 1999; Dranove et al., 2003). Therefore, other risk-mitigating measures may be necessary. In addition, to prevent P4P from crowding out providers' intrinsic motivation (Berwick, 1995; Marshall & Harrison, 2005), which may lead to undesired behavior, it is important that the design of the program and included measures are aligned with professional norms and values (Conrad & Christianson, 2004; Marshall & Harrison, 2005). Hence, providers should be closely involved in program design.

3.2.2 Whom to incentivize: individuals or groups

In P4P, it is crucial that performance profiles are sufficiently reliable. Patient panels of individual physicians are typically too small to measure individual performance reliably (Hofer et al., 1999; Krein et al., 2002; Nyweide et al., 2009). Consequently, targeting individuals will often result in misclassification of providers and incorrect allocation of payments (Adams et al., 2010a; Nyweide et al., 2009). In addition, enabling factors such as essential infrastructure and peer review are often available in group settings (Young et al., 2005). However, in groups P4P payments may not be effectively distributed to group members, and it may be tempting for members to free-ride on the efforts of peers, especially in large groups (Gaynor & Gertler, 1995; Gaynor et al., 2004; Town et al., 2004). Nonetheless, on balance group incentives seem to be preferred over individual incentives. Yet small practices will remain important and their inclusion should be facilitated (Landon & Normand, 2008). Finally, participation is ideally voluntary. Imposed participation may impair providers' intrinsic motivation and may lead to negative provider reactions. However, efforts may be needed to yield high participation rates and to prevent creating, maintaining, or widening disparities.

3.2.3 How to incentivize: structure of the incentive scheme

Individuals tend to respond more strongly to negative incentives than to positive incentives of equivalent size (Kahneman & Tversky, 1979). However, negative incentives are likely to be perceived as unfair and may result in negative reactions (Kahneman et al., 1986; Town et al., 2004). On the other hand, programs using rewards may not be sustainable. If rewards are not possible, the extent to which P4P will be successful depends on whether providers can be convinced to participate and whether undesired behavior can be prevented. Another aspect is the size of the incentives. All else equal, the larger the revenue potential for providers, the larger their response (e.g., Mullen et al., 2010). Payments must be large enough to offset the cost of improving performance (Young & Conrad, 2007; Conrad & Perry, 2009). Yet large payments are not necessarily more effective than smaller payments because physicians often have a target income (Rizzo & Blumenthal, 1996; Rizzo & Zeckhauser, 2003) and because of diminishing marginal utility of income. In addition, large payments are likely to impair providers' intrinsic motivation more than smaller ones (Deci et al., 1999; Frey, 1997), which may lead to undesired effects (McDonald & Roland, 2009). In addition to size, payment frequency also seems relevant. Individuals value immediate outcomes more than future outcomes of the same size (Frederick et al., 2002). Thus, paying €100 monthly may be more effective than paying €1,200 annually, also because it enhances incentive salience (Damberg et al., 2007). Finally, absolute targets may be more effective than relative targets because they are transparent and create less uncertainty regarding the efforts required to become eligible for payment (Conrad & Perry, 2009). In addition, relative targets encourage competition, which may reduce collaboration and dissemination of best practices (Rosenthal & Dudley, 2007). But when a program is applied uniformly to a large number of providers, absolute

targets may not be efficient because payments are made for performance already being delivered. Also, the goal-gradient hypothesis predicts that little response can be expected if the target is perceived unattainable or if the target is already attained (Heath et al., 1999). Adopting a tiered series of targets or differentiating targets based on baseline performance (with payment size conditional on level of attainment) can resolve this issue. Alternatively, purchasers may opt for paying providers a "piece-rate" for each appropriately managed patient or each recommended service (Rosenthal & Dudley, 2007; Chien et al., 2010).

3.3 METHODS

3.3.1 Search strategy and selection procedure

To ensure a comprehensive inclusion of available literature on existing P4P-programs, we consulted several sources, including Medline, the Internet, experts in the field of P4P, and reference lists of retrieved documents. During our search, we used information obtained from a particular source as input for consulting other sources. For example, if we found an article describing the features of a seemingly relevant program, we searched the different sources using program-specific keywords to obtain additional information.

We started by searching Medline through PubMed using the following keywords: pay-for-performance, P4P, pay for quality, bonus, malus, reward, penalty, withhold, financial/monetary/economic incentive, quality-/performance-/efficiency-based incentive/pay*/funding/remuneration/reimbursement, and value-based purchasing. These keywords were combined with the following terms: physician, doctor, practitioner, clinician, specialist, hospital, facility, clinic, nursing home, provider, HMO, MCO, IPA, POS, PPO, primary care, general practice, long-term care, elderly care, preventive care, and rehabilitation. This yielded 21 relevant articles. However, many P4P-programs will not have been described in the scientific literature as indexed in Medline. In addition, P4P-programs often have specific names that do not include common search terms like "pay-for-performance" and "financial incentives". For such programs, documents will often be available on websites, especially in case of a publicly administered program. Using the same keywords, we searched the Internet via Google, yielding another 25 documents and websites. Consultation of country-specific experts resulted in four additional documents. Four additional documents were identified by screening the reference lists of retrieved documents. While reviewing the literature on non-US P4P-programs, we simultaneously looked for documents describing typical P4P-program design in the US. We identified two review articles via Medline and four additional documents via the Internet.

To be included, documents had to be written in English, Dutch, or German and had to contain a clear description of at least one of the following program features: number and type of included performance measures, adopted risk-mitigating measures (including risk

adjustment and methods of data aggregation), monitoring and feedback mechanisms to detect and counter undesired behavior, information on sample size, strategies to engage providers, information on providers, strategies to facilitate inclusion of providers with small patient panels, participation (voluntary/mandatory, rates), size and type of financial incentives, number and type of performance targets, and payment frequency. We initially excluded documents published before 2005 to ensure sufficiently up-to-date program descriptions. However, this sometimes resulted in incomplete data, so for these programs we extended our search by searching for documents published since 2000. For the US programs, we only included documents providing an overview/review of typical program design in the US (i.e., descriptions of single programs were excluded).

3.3.2 Data extraction and abstraction

In extracting and summarizing the data, three steps were followed. For each of these steps, separate abstraction forms organized according to the three main categories of P4P-program design (Table 3.1) were used. The first step entailed reducing and categorizing the information in the documents to a table with only information about program design and relevant contextual factors (e.g., health system features and concurrent improvement efforts). Everything related to the features listed above was written down in detail to ensure important information was not missed. In step 2, using a new form, we compressed the table constructed in step 1, considerably reducing its length. For the purpose of presentation, this table was subsequently split in three parts, one for each of the three main categories. These tables can be found in Appendices 3.1 to 3.3. In the final step, program characteristics were summarized in a single table, incorporating only the key findings. Since the literature search and the data extraction and coding were performed by one person, it was not possible to compare results and resolve differences among multiple reviewers. Instead, to ensure reliability both the searches and the data extraction and coding were performed at two points in time, with four months in between.

3.3.3 Critical review of programs' design

An important objective of this paper is to critically review the design of identified programs. Because the literature reviewed in section 3.2 does not provide much insight into the weights of the various design elements, we assessed the appropriateness for each element separately. This was done for each program by awarding a "+" (appropriate design) or a "–" (inappropriate design). The programs were assessed on the following aspects: the set of measures is sufficiently broad and varied; risk adjustment is applied for outcome and resource use measures; efforts are made to mitigate providers' risk; providers are monitored for undesired behavior and receive adequate feedback if necessary; providers are actively involved in the design; performance is only measured on the level of the individual physician if sample size is sufficient; strategies are used to facilitate inclusion of small providers; participation is

voluntary and the majority of eligible providers participate; negative incentives are used appropriately; incentive size and frequency appear appropriate; absolute performance targets or piece rates are used; multiple targets are used (with larger payments for reaching higher targets) or targets are based on baseline performance.

3.4 DESCRIPTION AND CRITICAL REVIEW OF IDENTIFIED P4P-PROGRAMS

The systematic search identified 54 documents describing the features of thirteen non-US P4P-programs initiated in nine different countries. Table 3.2 contains general characteristics, Table 3.3 shows descriptive information on the programs' design, and Table 3.4 provides insight in the extent to which programs have been designed appropriately. Seven programs are regional while six have been implemented on a national level (one regional program, Advancing Quality in England, was later absorbed into a new program that applied to the whole of England; here we discuss the regional version). Eight programs were initiated by a public purchaser (typically in a single-purchaser healthcare system) and five by private insurers responsible for managing the care for their enrollees. P4P is often combined with nonfinancial incentives; providers regularly receive performance feedback and scores are publicly reported in at least five cases. (One program, NHI-P4P in Taiwan, differs from the other programs in the sense that it consists of five separate disease-specific programs. Yet in this paper it is treated it as a single program).

3.4.1 What is being incentivized?

Regarding the performance dimensions that are targeted, all programs incentivize clinical quality, which in five programs is the sole focus. Other dimensions include patient experience/satisfaction, financial performance, access, and capacity (i.e., structural measures referring to organizational and administrative aspects of performance such as record keeping and providing/receiving education). In the programs targeting multiple dimensions, clinical quality mostly gets most weight (54 percent on average) and contains most measures. There is variation in targeted areas within dimensions. For example, the clinical dimension contains twenty areas in the Quality and Outcomes Framework (QOF; 86 measures), three in the Practice Incentive Program (PIP; 22 measures), and eight in the Performance Management Program (PMP; 8 measures). Clinical aspects typically pertain to chronic care and/or prevention, although acute care (Advancing Quality [AQ], Clinical Practice Improvement Payment [CPIP], National Health Insurance [NHI]-P4P, QOF) and mental health (Clinical Practice Improvement Payment [CPIP], Physician Integrated Network [PIN], Program of Quality Improvement [PQI], QOF) are also common. Four programs adopted a measure set containing at least 30 measures pertaining to clinical quality and patient experience/

TABLE 3.2 General characteristics of identified P4P-programs

Program	Country/ Setting	Year	Public Reporting?	Performance Feedback?	References
Advancing Quality	United Kingdom, regional	2008	Yes	Yes, quarterly	West (2008), NHS North West (2008, n.d.), Premier Inc. (2010)
Clalit P4P	Israel	1998	Unknown	Yes, relative	Gross et al. (2008), Balicer et al. (2011)
Clinical Practice Improvement Payment	Australia, Queensland	2008	Unknown	Unknown	Ward et al. (2007), Duckett et al. (2008), Clinical Practice Improvement Centre (2008, 2010), Queensland Health (2010)
Ergebnis Orientierte Vergütung	Germany, regional	2001	Yes, quarterly	Yes, continuous	Gerdes et al. (2008, 2009), Ludwig Boltzmann Institut für Health Technology Assessment (2009), Walle (2009), Schlingensiepen (2009), Fachklinik Herzogenaurach (2010), Hochrhein-Institut (n.d.)
Maccabi P4P	Israel	2001	Yes, monthly	Yes	Friedman et al. (2003), Gross et al. (2008)
National Health Insurance P4P	Taiwan	2004	Unknown	Unknown	Chang (2004), Cheng (2006), Lee et al. (2010), Tsai et al. (2010), Li et al. (2010a), Chen et al. (2011), Kuo et al. (2011)
Performance Management Program	New Zealand	2006	Yes, semi-annually	Yes, quarterly	Gross et al. (2008), Buetow (2008), New Zealand Ministry of Health (2010), District Health Boards New Zealand (2010)
Physician Integrated Network	Canada, Manitoba	2004	Unknown	Yes, quarterly	Frohlich et al. (2006), Manitoba Health (2007, 2010, n.d.), Katz et al. (2010)
Practice Incentive Program	Australia	1998	Unknown	Yes	Scott (2007), Australian National Audit Office (2010), Medicare Australia (2011)
Primary Care P4P	Netherlands, regional	2005	No	Yes	Kirschner et al. (2008, 2009)
Primary Care Renewal Models	Canada, Ontario	2007	Unknown	Unknown	Price Waterhouse Coopers (2001), Wilson (2006), Anderson et al. (2006), Kantarevic et al. (2010), Li et al. (2010b)
Program of Quality Improvement	Argentina, Buenos Aires	2005	Unknown	Yes	Rubinstein et al. (2009), Wikipedia (n.d.)
Quality and Outcomes Framework	United Kingdom	2004	Yes, annually	Yes, continuous	Roland (2004, 2006), Doran et al. (2006), Doran et al. (2008a, 2008b), Guthrie et al. (2006), McDonald et al. (2009), Health and Social Care Information Centre (2009), Doran and Roland (2010)

satisfaction or access. In Table 3.4, these programs are awarded a "+". Maccabi also scores a "+" because a measure unknown to providers is defined retrospectively and included with 10 percent weight. Especially CPIP, the Primary Care Renewal Models (PCRM), PMP, and Ergebnis Orientierte Vergütung (ERGOV) include few domains and small measure sets.

TABLE 3.3 Design features of identified P4P-programs and the typical P4P-program in the United States

Design element	AQ	Clalit	CPIP	ERGOV	Maccabi	NHI-P4P [a]	PC-P4P	PCRM	PIN	PIP	PMP	PQI	QOF	US P4P [b]
What is being incentivized?														
Dimensions [c]	C, QL, PE	C, PS	C	C (ADL)	C, F, PS	C	C, Ca, PE	C	C	C, Ca, A	C, Ca, F	C, Ca	C, Ca, PE	C, F, IT
Number of measures	>30	>21	19	20	23	>23	80	11	38-40	38	14	31	134	10-25
Measures type [d]	P, FO	P, IO, FO	S, P	FO	P, IO	S, P, IO, FO	S, P	P	P	S, P	S, P	S, P, IO	S, P, IO	S, P, IO
Provider involvement	yes	?	yes	yes	yes	limited but increasing	yes	?	yes	yes	no	yes	yes	yes
Risk-mitigating measures [e]	RA, CS	RA (age)	?	RA, ER, MS	RA of target	MS; providers enroll patients	CS	Low target if patients refuse	MS	Indirect RA (age/sex)	RA	?	ER	RA, CS, MS
Who is being incentivized?														
Individual/group	group	group	group	group	group	both	group	individual	group	both	group	group	group	group
Provider type [f]	H	H, PC	H	Rehab.	PC	H, SC, PC	PC	PC	PC	PC	PC	PC	PC	PC
Participation [g]	V	C	C	V	C	V	V	V	V	V	V	C	V	V
How is being incentivized?														
Incentive type [h]	R, (P)	R	R	R, P	R	R	R	R	R	R	R	R	R	R
Maximum payment size [i]	4% of tariff, £260-702K	?	5-10% of DRG-fee	?	?	varies, e.g., 7% for BC	<5%	C$31K, C$6,86 per patient	?	A$58K per practice	?	8%	30%	P: 7%, H: 2.5%
Type of targets [j]	R	A	A	R	R	mostly A	R	A	A	A	A	A	A	A, R
Number of targets [k]	2 (achievement); 1 (improvement)	1	?	cont.	1	FFS for processes, 1 for outcomes	6	clinical:3-5; utilization:1	5	1 or FFS; ≥2 in some areas	cont., 1 target	3	cont., 2 targets	?
Payment frequency [l]	AN	AN	S-AN	Q	AN	M to AN	AN	AN	AN	Q or AN	S-AN	AN	AN	AN

Note: AQ = Advancing Quality; CPIP = Clinical Practice Improvement Payment; ERGOV = Ergebnis Orientierte Vergütung; NHI = National Health Insurance; PMP = Performance Management Program; PIN = Physician Integrated Network; PIP = Practice Incentive Program; PC = Primary Care; PCRM = Primary Care Renewal Model; PQI = Program of Quality Improvement; QOF = Quality and Outcomes Framework.

a. Data pertain to all five disease-specific programs. The number of measures includes structural measures that are mainly used as eligibility criteria for participation.
b. Sources: Rosenthal et al. (2006, 2007), Sorbero et al. (2006), Damberg et al. (2007), Baker and Delbanco (2007), Med-Vantage (2009).
c. A = access; ADL = activities of daily living; C = clinical; Ca = capacity; F = financial/efficiency/resource use; IT = information technology; PE = patient experience; PS = patient satisfaction; QL = quality of life.
d. S = structure; P = process; IO = intermediate outcome; FO = final outcome.
e. RA = risk adjustment; CS = composite scores; ER = exception reporting; MS = minimum size (i.e., providers can only participate if they have sufficient patients).
f. H = hospital; PC = primary care provider; SC = specialist care provider.
g. V = voluntary; C = compulsory.
h. R = rewards; P = penalties.
i. BC = breast cancer; P = physicians; H = hospitals.
j. A = absolute; R = relative.
k. FFS = enhanced fee for service; cont. = continuous.
l. AN = annual; S-AN = semi-annual; Q = quarterly; M = monthly.

In view of the relatively large payments in CPIP and uncertain financial consequences in ERGOV, concerns about teaching to the test are particularly large in these programs. The most comprehensive program is QOF, containing more than 130 measures in about 30 areas. Despite this, there is mixed evidence of teaching to the test in the QOF. One study showed neither improvement nor deterioration in unrewarded conditions (Steel et al., 2007). Another study showed a positive effect on unrewarded aspects of an included condition, a deterioration of unrewarded aspects of two other included conditions, and a reduction of the continuity of care (Campbell et al., 2009). These latter two findings may have been a result of the large bonuses, which may have crowded out general practitioners' (GP) intrinsic motivation. McDonald and Roland (2009: 123) found that "the requirement to enter data into the electronic medical record to respond to the large number of targets was described as reducing eye contact, increasing time spent on data collection, and crowding out the patient's agenda." But regarding incentivized performance, average attainment exceeds the maximum target for almost all measures (Doran & Roland, 2010), suggesting that money has not been the sole motivation for reaching high performance (Campbell et al., 2008).

The programs differ in the use of risk-mitigating measures. Risk adjustment is used in AQ, ERGOV, and PMP for outcome and financial measures. Especially in AQ and ERGOV, risk adjustment seems fairly sophisticated, controlling for sociodemographic and morbidity-based risk factors. Not all programs that include outcomes (seven in total) apply risk adjustment. NHI-P4P includes unadjusted outcomes for diabetes, breast cancer, and tuberculosis (TB), which may have resulted in the finding that for diabetes, older patients and patients with greater disease severity or comorbidity were more likely to be excluded from the program than younger patients and patients with less disease severity (Chen et al., 2011). QOF also includes unadjusted outcomes, but practices are allowed to exclude certain patients from the measurements. While this provides them with a tool to increase income by excluding "difficult" patients or patients for whom targets had been missed rather than because of an appropriate reason, there is little evidence of inappropriate use (Doran et al., 2008a; Gravelle et al., 2008). Audits and penalties for fraud may have contributed to preventing this. Across the programs, various other risk-mitigating methods are used. For example, in Maccabi performance targets are differentiated according to how current performance is affected by population features and casemix. Yet in general, although the documents provide limited information on the use of risk-mitigating measures, the results raise doubts about whether differences in (patient) risk are sufficiently equalized, especially in NHI-P4P, Clalit, Maccabi, Primary Care (PC)-P4P, and PQI.

In most programs providers are actively involved in program design. Provider support is considered a critical success factor and is being realized in various ways, including consensus meetings (AQ, PIN, PC-P4P, PQI), delegating measure development to providers (CPIP, Maccabi, PIP, QOF), and adjusting measures based on provider feedback (CPIP, PIN, PC-P4P, PQI). A notable exception is PMP, in which measures seem to be imposed top-down.

According to Buetow (2008: 42), it is "unclear that the program appropriately reflects the values and goals of (…) providers of primary health care (…)." Appendix 3.1 provides more details on "what is being incentivized".

3.4.2 Who is being incentivized?

Across the programs, payments are mostly provided at the group level. Targeted "groups" vary in size and structure, ranging from hospitals (AQ, CPIP) to large multispecialty organizations (PMP) to primary care practices (QOF, PC-P4P, PIP). In five programs, payments are provided only to individual physicians or to both individuals and groups. Targeting individuals is appropriate for measures under physicians' direct control. In PIP, payment is provided to the primary care practice for measures for which this does not seem to hold. For example, GPs receive an enhanced fee for each Pap smear, but the practice receives a fixed amount per patient if a specified percentage of patients are screened. In NHI-P4P, payment is provided to hospitals for diabetes and cancer, but directly to physicians for asthma and TB. In PCRM, payments are mostly made to GPs. However, for many measures included in these programs, sample size may well be insufficient to generate reliable profiles, especially for outcomes and resource use (Hofer et al., 1999; Krein et al., 2002; Scholle et al., 2008; Mehrotra et al., 2010b). This also seems relevant for PIP, PC-P4P, and QOF, as many GPs still work in solo or small group-practices. For several programs, documents state that measures are only included if they are sufficiently under providers' control and/or if sample size is sufficient. However, it is unclear when this is the case. Some programs (e.g., AQ, PC-P4P) aggregate individual measures into composites, which could enhance reliability (Kaplan et al., 2009; Holmboe et al., 2010). In PC-P4P this resulted in fair reliability, despite that many practices were solo or duo practices. Although it is hard to draw conclusions, there are concerns about whether providers can be adequately discriminated from each other, and thus whether payment allocation occurs adequately.

In seven programs (AQ, Clalit, CPIP, Maccabi, PMP, PQI, QOF) participation rates are virtually 100 percent. In PIP, PCRM, and NHI-P4P participation exceeds 50 percent. Low participation may be problematic in ERGOV, PCRM, and NHI-P4P. In ERGOV, participating clinics are recognized as preferred providers. Especially if receiving care from non-preferred providers involves large out-of-pocket payments, this may be a strong incentive for clinics to participate. But participation may still not be attractive as it involves a considerable administrative burden while financial consequences are highly uncertain. To achieve meaningful differences, participation needs to increase (Gerdes et al., 2009). Also in NHI-P4P for breast cancer, low participation seems to be a result of the additional financial risk that participation involves and the fact that hospitals experience survival rates, which determine whether or not they receive a bonus, to be largely beyond their control (Kuo et al., 2011). Appendix 3.2 provides more details on "who is being incentivized".

TABLE 3.4 Appropriateness of the design of identified P4P-programs

Design element	AQ	Clalit	CPIP	ERGOV	Maccabi	NHI-P4P	PC-P4P	PCRM	PIN	PIP	PMP	PQI	QOF
What to incentivize?													
Broad and varied set of measures	+	+/-	-	-	+	+/-	+	-	+/-	+	-	+/-	+
Risk adjustment for outcomes and resource use	+	+/-	n.a.	+	+/-	-	n.a.	n.a.	n.a.	n.a.	+	-	-
Risk-mitigating measures (excl. group incentives)	?	?	?	-	?	+	-	+/-	+/-	?	?	?	+
Monitoring for undesired behavior / feedback	+	?	?	+	?	+	+/-	?	+/-	?	+/-	?	+
Active provider involvement in design	+	?	+	+	+	+/-	+	?	+	+	-	+	+
Whom to incentivize?													
Individual incentives only if sample size sufficient	n.a.	n.a.	n.a.	n.a.	n.a.	?	+/-	?	n.a.	?	n.a.	n.a.	?
Strategies to facilitate inclusion of small providers	+	?	?	-	?	+/-	+	+	-	+	?	?	+
Voluntary participation, high participation rates	+/+	-/+	-/+	+/-	-/+	+/-	+/-	+/-	+/-	+/+	+/+	-/+	+/+
How to incentivize?													
Negative incentives used appropriately	+	n.a.	n.a.	-	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Size of payments appropriate	+	?	+	?	?	+/-	+	-	?	+	-	+	+/-
Frequency of payments appropriate	-	-	+/-	+	-	+/-	-	-	-	+	+/-	-	-
Absolute targets or piece-rates	-	+	+	-	-	+	-	+	+	+	+	+	+
Multiple targets or baseline-based targets	+	-	?	+/-	-	-	+	+	+	+/-	+	+	+

Note: AQ = Advancing Quality; CPIP = Clinical Practice Improvement Payment; ERGOV = Ergebnis Orientierte Vergütung; NHI = National Health Insurance; PMP = Performance Management Program; PIN = Physician Integrated Network; PIP = Practice Incentive Program; PC = Primary Care; PCRM = Primary Care Renewal Model; PQI = Program of Quality Improvement; QOF = Quality and Outcomes Framework. "+" denotes appropriate design; "–" denotes inappropriate design; "n.a." = not applicable; and "?" denotes insufficient information.

3.4.3 How is being incentivized?

Two programs have adopted financial penalties. In ERGOV, bonuses for high performers are financed by maluses for low performers. Although this contributes to financial sustainability, incentives for gaming may be large (Gerdes et al., 2009). To prevent this, clinics are required to supply data via an online tool that enables auditing and checks. In AQ, there were no penalties for low performers, but hospitals that failed to meet targets for data accuracy and completeness received a penalty or were excluded from the program. (The current version of AQ involves withholding of payments rather than bonuses.) Compared with ERGOV, the financial risk of participation was smaller. Also, there was less uncertainty as payments were fixed. Regarding payment size, there is much variation across programs. In AQ, in addition to payments for patient-reported outcomes, hospitals could earn a 4 percent add-on to the national tariff for the associated activity. In CPIP, bonus potential is 5 to 10 percent of the average DRG price. In NHI-P4P, payments per patient per year are often maximized. For cervical cancer, fees may be increased by up to 50 percent. For breast cancer patients, eligible hospitals receive a bundled payment, which is higher than the regular payments. Hospitals also meeting targets for disease-free survival are eligible for a bonus of up to 7 percent of the bundled payment. The payments in PMP are "small in relation to total PHO incomes" (Buetow, 2008:40). In PCRM, "the monetary values of the P4P incentives are a relatively very small proportion of the total income of GPs" (Li et al., 2010b:15). In contrast, in QOF payments can be up to 30 percent of practices' revenues, which seems to have contributed to improvements (Doran & Roland, 2010). Yet it is likely that smaller payments would have generated similar results.

Eight programs only incentivize absolute performance. In AQ, hospitals in the top two quartiles were eligible for payment, and in ERGOV, clinics are judged on their performance relative to the mean. In Maccabi, only the three clinics in each of the five clinic size categories that best achieve their own target receive a bonus proportional to the degree of target attainment. Six programs use three or more targets or a sliding scale. PIN typically uses five targets per measure with a large difference between tiers. A similar approach is used in PCRM, which may well have contributed to the finding that improvements in incentivized measures were typically largest among GPs with low or medium baseline performance (Li et al., 2010b). PMP and QOF use a sliding scale. Providers in PMP earn more for a larger percentage improvement from baseline to the target. In QOF, each measure has lower and upper targets delineating the scale. Improvements in performance were most pronounced for GPs with low scores at baseline (Doran et al., 2008b), which could have been a result of the sliding scale on which practices are scored. NHIP4P provides piece rates for process quality. For example, for breast cancer, hospitals are rewarded for each patient completing recommended treatment. This may well have contributed to observed improvements in process and outcome quality (Kuo et al., 2011). While the use of multiple targets is important especially in relative schemes, the four programs using relative targets typically use only one

or two targets. An exception is PC-P4P; practices above the 25th percentile are eligible for a bonus and divided into six groups. Appendix 3.3 provides more details on "how is being incentivized".

3.4.4 Comparison with P4P-programs in the United States

In the US, 256 P4P-programs were in place in 2007. The last column of Table 3.3 shows the features of a typical US P4P-program (Rosenthal et al., 2006; Sorbero et al., 2006; Damberg et al., 2007; Baker & Delbanco, 2007; Rosenthal et al., 2007b). Similar to non-US programs, clinical quality is the most commonly incentivized dimension, followed by resource use/efficiency (40-50 percent of programs), information technology (IT) adoption (30-40 percent), and patient experience or satisfaction. Especially efficiency and IT adoption are more common in US programs. Measure sets are mostly quite small in the US, typically ranging from 10-25 measures. Nine non-US programs fall within or are close to this range. These relatively small numbers are mainly a result of strict inclusion criteria such as consistency with other improvement strategies and endorsement of professional organizations. To our knowledge no study has found evidence of teaching to the test as a result of P4P in the US (e.g., Glickman et al., 2007; Mullen et al., 2010). Although outcome measures are increasingly used, process and structural measures are much more common. Similar to non-US programs, if outcomes are included, they usually pertain to intermediate effects in physician P4P-programs and complication and mortality rates in hospital programs. The limited use of outcomes seems to be a result of lack of data and concerns about the adequacy of risk adjustment, although risk adjustment is increasingly being applied, especially in P4P for hospitals. Finally, many purchasers underscore the importance of active provider engagement. Mechanisms are often in place to obtain ongoing input, and providers usually have the option to raise concerns about scores and data.

In the US, the vast majority of P4P-programs target groups. In hospital programs, purchasers use a variety of strategies to avoid small numbers, including using multi-purchaser data, constructing composite scores, and only using measures that apply to large numbers of patients. In physician programs, most purchasers require that a minimum number of patients can be attributed to a provider. This is also the case in some of the non-US programs, although the retrieved documents were generally not very specific on this topic. Similar to non-US programs, most US programs are in primary care, although programs for hospitals and specialists have been increasing in number. In 2007, there were more than 40 hospital programs and of all physician programs in 2006, 41 percent only target primary care physicians, none target specialists only, and 59 percent target both.

More than 60 percent of US P4P-programs only use bonuses, but withholds are still relatively common (10-20 percent of programs). The average payment size is about 7 percent for physicians and 2.5 percent for hospitals (Med-Vantage, 2009). Payment frequency is mostly annual, although ongoing payments are becoming increasingly common. Programs

using absolute targets have increased in number and most physician programs now reward absolute performance. Yet relative targets are still used in about 50 percent of all programs. This compares well with the non-US programs in that relative targets are mainly used by competing purchasers.

3.5 DISCUSSION

This study provides an international overview of P4P initiatives in health care. The thirteen identified programs have similar design in several respects. They all incentivize clinical quality and most of them only use positive incentives, actively involve providers in design, target primary care providers, and pay on an annual basis. However, there is also considerable heterogeneity regarding the breadth of measure sets, use of risk-mitigating measures, payment size, and number and type of targets. In most programs there seems to be ample room to increase incentives for desired behavior and to mitigate incentives for undesired behavior. In particular, shortcomings pertain to number and type of included performance measures, risk adjustment of outcomes and resource use, reliability of measurements, payment frequency, and number of targets. Modification seems relevant mainly for Clalit, ERGOV, Maccabi, and NHI-P4P, but also for other programs there is room for improvement, notably regarding measure sets, risk-mitigating methods, and payment size. For some aspects design seems adequate in most programs. These include provider involvement in design (nine programs), voluntary participation (nine programs), and type of targets (absolute targets or piece-rates in nine programs). AQ and QOF seem to have been designed particularly well. The effectiveness of QOF has been evaluated in several studies, and the positive results (Doran & Roland, 2010) seem to correspond with this finding. AQ had not been formally evaluated at the time of writing.

Despite that the design of NHI-P4P seems to be lacking in several respects, several studies have found positive effects of this program (e.g., Kuo et al., 2011; Lee et al., 2010; Li et al., 2010a). This may seem surprising, but the shortcomings in NHI-P4P's design pertain mainly to aspects that mitigate undesired behavior, including a relatively narrow definition of performance (concerns about teaching to the test), no risk adjustment for outcomes (incentives for selection), and limited provider involvement in design (provider support unlikely). The evaluations typically lack an assessment of these types of undesired consequences, and the one study we know of (Chen et al., 2011) found evidence of selection. Thus, although incentivized aspects appear to have improved, this may have come at the cost of worse performance on unincentivized aspects.

The programs share several design features with the typical US P4P-program: clinical quality is most common and generally gets most weight (50 percent or more); measure sets are usually relatively small; outcomes are not often included and when they are, they

pertain to similar aspects; provider engagement is considered a vital; most programs target physician groups in primary care; and payments are mostly made on an annual basis. There are notable differences as well. Negative incentives and ongoing payments are more often applied in the US. Furthermore, although for physician programs payment size appears to be similar (perhaps somewhat higher in the non-US programs), for hospitals generally more generous payments are used in the non-US programs. Finally, relative targets are more often used in the US. This may be explained by the competitive nature of the US care system. Providers (as well as purchasers) are used to competitive forces, so the use of relative targets may be more acceptable to them. This is backed by the finding that among the non-US programs, relative targets are only used in programs initiated by competing purchasers (AQ is an exception). Competition may also be an explanation for the finding that US programs rely more on efficiency measures; of the three non-US programs including financial performance or efficiency, two were initiated by competing HMO-like entities. Among other "competitive programs", ERGOV uses a budget neutral approach and PC-P4P considerably reduced payments for budgetary reasons. US programs also more often include measures of IT adoption. Besides the fact that IT applications themselves may benefit performance, the problems associated with using claims data to generate performance information and the high costs of manually extracting data from medical records are probable explanations for the (increasing) use of such measures. As the diffusion of P4P continues and the adequacy of performance measurement becomes more relevant, it can be expected that such measures will increasingly be used also in non-US programs.

3.5.1 Implications

Our findings have several implications for the future of P4P as a performance improvement effort. First, inadequately designed programs may stimulate undesired provider behavior, and more insight is required in how such behavior can be prevented. Several studies have shown that risk selection is not just a theoretical concept (Shen, 2003; Dranove et al., 2003; McDonald & Roland, 2009; Chen et al., 2011) and although evidence on teaching to the test is both limited and mixed (Steel et al., 2007; Glickman et al., 2007; Campbell et al., 2009), Mullen et al. (2010:86) rightly argue that:

> The concern that P4P encourages teaching to the test should not be dismissed. Given the complex and largely unobservable nature of health care quality, we can only study some potential unintended consequences but we cannot confirm or reject the existence of all such effects (…). The negative incentives of P4P (…) should be taken seriously given evidence that providers do indeed respond to incentives.

Many current P4P-programs have shortcomings with respect to design elements that relate to preventing undesired behavior (specifically teaching to the test and risk selection), and

there is large variation in the use of risk-mitigating measures. This suggests that purchasers, though clearly concerned about them, are uncertain about how to effectively prevent undesired effects. Thus, because such effects can potentially undermine the entire program, more insight into how they can be prevented is required. For example, research should continue to focus on developing adequate risk adjustment than can be applied transparently in practice and on the merits and drawbacks of potentially viable alternatives or supplements such as exception reporting. Second, if P4P is to contribute to improving patient outcomes, payment allocation must be based on timely, reliable, and accurate performance data. Many shortcomings in the design of current programs, including low payment frequencies, small measure sets, limited use of outcomes, and lack of risk adjustment, can be traced back to a lack of data. Efforts should continue to focus on creating methods for registering, extracting, and processing patient-level data, and the merits of IT for these purposes should be explored further. Third, breakthrough improvements require coordination across disciplines and alignment of incentives for all providers in the continuum of care. Current programs focus too much on a specific sector and type of provider (physician groups in primary care). Aligned incentives require strategies to facilitate inclusion of small practices (e.g., developing methods for aggregating performance data) as well as incorporating incentives that encourage coordination. Customized IT and forms of prospective payment like bundled case rates will prove vital in attaining these goals. If structured around patients rather than providers, prospective payment with performance-based elements can both reduce the problem of overuse of low-value services and reward providers for effectively coordinating care (Rosenthal, 2007a). Fourth, it is crucial that programs are evaluated using convincing control groups. Of the identified programs, only seven have been evaluated, and often only partially. There is a paucity of empirical research on the influence of specific design choices, and we still know very little about which designs are most effective in a given context. Therefore, studies should not only assess effectiveness but also include assessments of undesired effects and the influence of specific design elements. This not only provides insight in which parts require modification, but also important lessons on program design. Finally, provider input is essential in developing risk-mitigating measures, preventing unintended consequences, generating reliable performance information, designing payments models that encourage coordination and improvement across the continuum of care, and conducting sound program evaluations.

3.5.2 Limitations

This study has several limitations. First, conclusions about (appropriate) design were not always possible because the documents often lack specific information. For example, if risk adjustment is being applied, the program scored a "+" for this aspect, despite limited information about the adequacy of the risk-adjustment model. Thus, the analysis only provides indications of the adequacy of design. Second, conclusions about the adequacy of design

are largely based on theory and have not been confirmed in practice. Many of the programs identified in this study have not been (extensively) evaluated and regarding the programs that have been, the extent to which specific design features have contributed to changes in performance is unclear. Third, it is likely that relevant programs were not identified as a result of the search strategy, specifically the language restriction. Finally, this study suffers from publication bias. We know, for example, that there are other P4P-programs in effect in Canada (Canadian Medical Association, 2010), Germany, Italy (Fiorentini et al., 2011), Spain (Gené-Badia et al., 2007; Pedrós et al., 2009), and The Philippines (Peabody et al., 2011), but we did not find sufficient information on these programs to be included in this review.

3.6 CONCLUSION

P4P is now widely being applied in many healthcare systems and there are no indications that this will change in the near future. However, current evidence suggests that designing an effective P4P-program is a highly complex undertaking. Given the limited knowledge about "what works" in P4P, it may then not be very surprising that the design of current programs seems to be lacking in several respects and that purchasers struggle with developing effective programs. To get the most out of P4P, well-conducted evaluations are critical for generating the information needed for fine-tuning P4P to the specific settings of implementation. In particular, empirical research investigating the influence of specific design choices in specific settings is needed, as well as insight in the perverse incentives of P4P and how these can be prevented. In parallel, if P4P is to contribute to improved patient outcomes, efforts should continue to focus on creating reliable and easy-to-use methods for generating comprehensive patient-level (performance) data.

APPENDICES

Appendix 3.1 Characteristics of identified programs: what is being incentivized?

Program	Performance dimensions (weight)	Performance measures	Measure selection; provider involvement; data collection	Methods employed to mitigate providers' risk
AQ At least 30 measures	• Clinical (60%) • Patient-reported outcome measures (PROMs) (20%) • Patient experience (PE) (20%)	• Clinical: 28 processes, 2 final outcomes; divided over 5 acute care areas • PROMs: quality of life before and after treatment • PE: no information	• Measures developed in context of CMS/Premier Hospital Quality Incentive Demonstration in the US • Endorsed by clinicians, NICE, and royal colleges • Data self-collected; targets for completeness/accuracy; centralized support	• Risk-adjustment: survival index for acute myocardial infarction, PROMs • Composite score for each clinical domain
Clalit At least 21 measures	• Clinical (30%) • Service (25%) • Management evaluation (25%) • Marketing (10%) • Quality of routine work / plans (10%)	• Clinical: 10 processes, 8 intermediate outcomes • Service: patient satisfaction, percentage of patients who plan to transfer, percentages of patients with complaints	• Selection based on experience from other countries, clinical relevance, and data collection possible via electronic health record • No information on provider involvement, Medical Association may be involved	• Measures age-standardized • Measures explicitly selected based on sufficient sample size per clinic
CPIP 19 measures	• Clinical (100%) • Intention to add outcome and access measures, functional status, and physician communication	• 12 structures and 7 processes in 10 areas • Acute, chronic, and metal health care • Area can be disease- or care-specific (e.g., stroke vs. intensive care)	• Selection based on disease burden, consensus, and available evidence • Developed by clinical networks, endorsed as measures of quality. • Data collected by clinical units	• Measures only included if they are sufficiently within providers' control
ERGOV 20-item tool	• Outcome quality of rehabilitation care for stroke patients • Patients' ability to perform activities of daily living	• Self-care (7 items), communicative skills (4), mobility (4), cognitive activity (5) • 6 types of assistance per item	• Quality assessment tool combining items from widely used measurement instruments with good psychometric properties • Endorsed by clinics	• Risk-adj. ($R2=0.84$) • Data self-reported online; checks using own assessment • Patient exclusion • ≥100 patients
Maccabi 23 measures	• Clinical (45%), added in 2004 • Financial (35%) • Service (20%) • "X-measure" is defined afterwards and included with 10 percent weight in total score	• Clinical: 12 processes, 5 outcomes (largely prevention/chronic) • Financial: deviation from budget and use of hospital services and drug formulary • Service: patient satisfaction, retention, complaints	• Set based on own system, HEDIS, CAPHS, and UK NHS • Selection based on relation with outcomes, evidence • Physicians from national and regional levels actively involved; public involved in defining patient satisfaction • Data collection: electronic administrative/medical files	• Measures included if sufficiently within providers' control • Risk-adjustment: different targets per region based on how baseline performance is affected by casemix

Program	Performance dimensions (weight)	Performance measures	Measure selection; provider involvement; data collection	Methods employed to mitigate providers' risk
NHI-P4P At least 23 measures	• Clinical (100%) • diabetes mellitus (DM), asthma, Breast/cervical cancer (BC/CC), tuberculosis (TB) • Intention to add clinical outcomes, hypertension, hepatitis B/C, schizophrenia	• DM: 2 structures, several processes, 2 intermediate outcomes • Asthma: 2 structures, several processes • BC: 4 structures, processes, 2 outcomes • CC: 2 processes • TB: 4 structures, several processes, 1 final outcome	• Measures (also) selected based on disease burden • Data self-reported by providers and entered automatically in a database • Intention to increase provider participation in program development and measure selection	• Sample size requirement • Providers decide which patients to enroll. Government increases the no. of patients physicians have to enroll (in 2009 for DM: 30% of population, ≥ 50 patients)
PC-P4P 80 measures	• Clinical (50%) • Practice features (PF, 25%) • Patient experience (PE, 25%) • Intention to add drug safety and mental health	• Clinical: 31 processes divided over 7 areas, chronic/prevention • PF: 22 structures, 4 areas • PE: 27 measures, 2 areas, survey after visit	• Inspired by UK QOF • Selection: evidence, reliability, validity, professional guidelines • Development/selection in cooperation with GPs • Data self-reported (e.g., via own information system, patient sample, surveys)	• Reliability mediocre for PF, good for PE. • Reliable composites • Clinical: generally 100 patients per condition and sufficient practices to draw reliable conclusions
PCRM 11 measures	• Clinical (100%) • Utilization (e.g., obstetrical, palliative care)	• Clinical: 5 processes, 5 areas; preventive care • Utilization: 6 processes, 6 areas	• Measures developed and selected by government • Data collection via billed codes and self-reported documentation	• Targets adjusted downwards for e.g., patients declining care for religious or medical reasons
PIN 38-40 measures	• Clinical (100%) • Intention to add access, ongoing care, coordination, and mental health	• 24 processes, 6 areas, chronic care; 14 processes, prevention • Sites in a depression trial also judged on 2 additional processes	• Source: Canadian Institute of Health Information and Centre for Health Policy • Selection: consensus meetings, expert opinion • Data system populated via clinics' electronic records	• Measures included if "specific" to clinics and data are reliable and valid • Measures adjusted based on feedback • Checks with registry
PIP 38 measures	• Clinical (3 areas) • "Capacity" / general involvement in improvement efforts (4 areas) • Access (6 areas)	• Structure and process • Clinical: chronic care (annual cycles, 2 areas with 12 + 7 processes), prevention (3 process) • Capacity (4, e.g., practice nurses) • Access (12, e.g., specialist care rural)	• Measures nationally developed/negotiated; physicians are consulted • Payment formula developed with negotiating body with government and physicians • Practices self-report data and maintain separate patient records	• "Patient load" per practice; sum of the fractions of care provided to a patient weighted for age and sex; determines 75 percent of the payments
PMP 14 measures	• Clinical (60%) • Capacity (10%) • Financial (30%) • District Health Boards (DHB) and health organizations (HO) may add extra measures	• Clinical: 8 processes (mainly prevention) • Capacity: 2 measures • Access: 2 measures • Financial: GP-referred lab and medication expenditure	• Measures and targets developed by HO advisory group and DHB / Ministry team including experts • Providers have little influence on design • Data collated in national database	• Financial: target includes prior resource use, adjustment for policy changes and for low historic utilization by high-need groups

Program	Performance dimensions (weight)	Performance measures	Measure selection; provider involvement; data collection	Methods employed to mitigate providers' risk
PQI 31 measures	• Clinical (64%) • Capacity (33%) • Referral to specialists for high prevalence conditions (3%)	• Clinical: 3 outcomes, 16 processes (chronic, prevention, mental) • Capacity: 7 structures (medical education activities and quality of documentation)	• Inspired by UK QOF • Measures selected by purchaser with physician input; meetings to discuss measures/ weights, targets • Data extracted from hospital data system (also for physicians) or from groups' electronic records	• Switch made from individual incentives to group incentives, also to increase reliability
QOF 134 measures	• Clinical (69.7%) • Organizational (16.75%) • Patient experience (9.15%) • Additional services (4.4%)	• Clinical: 86 measures, 85 percent process, 20 areas; acute, chronic, metal, prevention • Organizational: 36 measures, 5 areas • Patient experience: 3 measures, 2 areas • Additional services: 9 measures, 4 areas	• Measures developed by professional organizations • Selection/weights based on negotiations between the British Medical Association and the government • Data collection: uniform electronic medical record managed by GPs, extracted to national database	• Exclusion of certain patients by GPs; exception reporting • Annual inspections by Primary care trusts, large penalties for fraud

Appendix 3.2 Characteristics of identified programs: who is being incentivized?

Program	Individual or group?	Characteristics of targeted providers	Participation
AQ	• Group • Payments allocated to clinical teams, to be invested in patient care	• All 24 hospitals in the Northwest region of England that provided emergency care • Hospitals can be public or private	• Voluntary
Clalit	• Group • First introduced at clinic level, later expanded to districts and hospitals • Bonuses can be used freely	• All 1,500 primary care clinics • Clinics are part of 1 of 8 districts • Clinics receive a budget from district management to manage all patient care	• Compulsory for clinics since 2001
CPIP	• Group • Earned funds allocated to hospitals (stimulated to allocate ≥80% to units)	• 16 districts in Queensland • Program only applicable to the 22 public hospitals	• Compulsory
ERGOV	• Group	• 13 rehabilitation clinics (pilot). • project financially supported by clinics • Clinics assess patients at admission and at discharge	• Voluntary
Maccabi	• Group • Bonuses used at units' discretion	• Districts and clinics ("units") • Units receive budget to manage all care	• Compulsory
NHI-P4P	• Asthma: individual • Diabetes mellitus (DM), breast/ cervical cancer (BC/CC): group (oriented to physicians, but payment given to hospitals) • Tuberculosis (TB): both	• Asthma: pediatricians, internists, and GPs • DM: hospitals. Physician participation 47%; 17% of patients eligible • BC/CC: hospitals (BC: 44% of patients in 2004) • TB: hospitals (43% participated in 2006), physicians, case managers. 70% of patients	• Voluntary

Program	Individual or group?	Characteristics of targeted providers	Participation
PC-P4P	• Both: 23% solo-practice, 30% duo-practice	• 72 GP practices (1.7% of total). 60 participated in 2 measurements and 12 in 3 measurements • Mean number of patients per practice was 5,226 • Practices participated in improvement trajectory of Dutch GP Association's Accreditation Program • Gatekeeping and patient enrolment required	• Voluntary
PCRM	• Preventive care bonus payments: Group: FHN; Individual: FHG, FHO, and CCM • Utilization special payments: always paid to individual GP, but mostly not possible for FHG, CCM	• FHN: 4% of GPs, ≥3 GPs, capitation for enrolled patients (80%), fee-for-service (FFS) for non-enrolled patients, list size requirement of 2400 patients per GP • FHO: 25%, ≥3 GPs, capitation, patient enrolment • FHG: 30%, ≥3 GPs, FFS, patient enrolment • CCM: 3%, solo, FFS, patient enrolment	• Participation in different primary care renewal models is voluntary
PIN	• Group, but payment often divided over participating physicians • Primary care groups receive funds on behalf of member clinics • Free allocation of funds by clinics	• Phase 1: typically 15 to 30 physicians (mostly GPs, but also specialized physicians and other practitioners) • Currently (phase 2): 13 primary care groups • Criteria to participate: electronic medical record, ≥5 GPs, 6,500 patients, access for other GPs	• Voluntary
PIP	• Often group; solo possible • In some areas, part of the payments are paid to individual GPs • Spurred to use funds to upgrade practice/offer GPs more income	• GP practices throughout Australia • In 2009/10, 4,881 practices (21,000 GPs) participated (67% of total; 82% of primary care) • Participation if accredited (criteria set by the Royal Australian College of GPs)	• Voluntary
PMP	• Group • PHOs may distribute bonuses to practices; district health boards may not approve how primary health organizations (PHO) proposed to use the payments • PHOs stimulated to use bonuses to improve services and infrastructure	• All 81 non-governmental, not-for-profit PHOs (doctors, nurses, and other professionals) • PHOs vary in size and structure and provide care themselves or through provider contracting • 95% of population enrolled with a PHO, almost all primary care practices are part of a PHO	• Voluntary for PHOs; GPs obliged to supply data if PHO participates
PQI	• Group • Bonuses likely distributed to physicians based on individual performance	• All 5 primary care groups of organization • Groups consist of 10-15 physicians (GPs, internists, pediatricians) and are responsible for the care of 10,000-15,000 patients	• Compulsory
QOF	• Mostly group, but solo-practices possible (6% of total in 2008)	• 8,600 primary care practices (almost 100%) • On average 5,500 patients, 3.5 physicians • Gatekeeping and patient enrolment required	• Voluntary

Appendix 3.3 Characteristics of identified programs: how is being incentivized?

Program	Positive or negative incentives	Incentive size	Base payments	Performance targets	Payment calculation	Payment frequency
AQ	• Positive (2010) • Penalties for data inaccuracy/ incompleteness	• Clinical: 2-4% add-on to tariff (bonuses totaling £3.2M in 1st year and £1.6M in 2nd year) • Patient-reported outcomes and experience: both £1M/year • Max. between £260K-702K/year depending on hospital size	• National tariffs for clinical conditions	• 1st year, clinical: relative, +4% or +2% for reaching top or 2nd quartile of achievement • 2nd year, clinical: relative; attainment (1 target), most improved (1 target), top performance (2 targets)	• Hospitals ranked on composite score per clinical area, measures weighted equally	• Annually, 2 or 3 month lag
Clalit	• Positive (unclear if also negative)	• Depends on the savings on the budget	• Budget to manage all patient care	• Absolute target per measure based on HEDIS/best district's performance last year	• Budget savings x overall %-grade • Grade ≥75% to be eligible for pay • High quality, no savings: small pay	• Annually
CPIP	• Positive	• Max. 5-10% of average diagnosis-related group (DRG) price • 2008: actual payment A$50-100 per patient or 1-3% increase in marginal revenue per patient	• Mostly case-mix adjusted prospective DRG pay	• Thus far pay for reporting; move to P4P where clinical networks set absolute targets	• % add-on based on improvement • 6 domains: set amount per patient per measure (within limits) • 4 domains: fixed amount per measure per provider	• Semi-annually, 3 month lag
ERGOV	• Both • Thus far payments have been virtual	• Depends on how clinic performs relative to other clinics and on size of multiplier (can be adjusted so that clinics do not go bankrupt) • Budget-neutral: maluses for low-performers used to finance bonuses for high-performers	• Typically per diem or flat-rated payment per patient	• Quality tournament in which bonuses and maluses are determined by relative differences (deviation from mean)	• 100 points per patient (5 per item) • Value at discharge subtracted from predicted value, calculated using admission scores from all patients. Residuals per clinic averaged and multiplied by no. of patients to calculate bonus	• Quarterly
Maccabi	• Positive (probably)	• Not reported	• Budget to manage all patient care	• Each unit receives own target based on baseline and top performance in the region; all units expected to show same percentage improvement	• Maximum total bonus amount • For each of 5 clinic sizes, 3 clinics best achieving their targets (in %) receive bonus proportional to degree of attainment per measure	• Annually

Program	Positive or negative incentives	Incentive size	Base payments	Performance targets	Payment calculation	Payment frequency
NHI-P4P	• Positive • Financing not from global budgets	• Diabetes (DM): NT$1,8K/patient/ year for process measures • Asthma: NT$1,2K/patient/year • Breast cancer (BC): 2, 3, 4, 6 and 7% add-on to bundled payment for 1, 2, 3, 4, and 5 year disease-free survival; on average NT$126K • Cervical cancer (CC): 10-50% add-on to current fees • Tuberculosis (TB): hospitals NT$13K/case; physicians NT$1,5K/ case; case manager NT$6K (first 6 months $3K, then $500/month)	• Fee-for-service under global budgets. 53 procedures reimbursed through fixed case payments	• Positive scores on structures • For DM, Asthma, BC, TB: enlarged fees for processes • DM: relative target for outcome measures • BC: absolute target for disease-free survival • CC: number of and monthly growth in Pap smears • TB: cure rates	• DM: outcomes: top 25% on composite (mean rank) • BC: stage-specific targets for survival (4 pathology stages + overall), descending targets and ascending payment over the five survival-years • CC: 10-50% add-on to current fees based on size of improvement • TB: payment varies by 4 treatment stages and is larger if cured earlier	• Monthly to annually
PC-P4P	• Positive	• Clinical: €3,45/patient insured with either one of the 2 insurers • Practice features and patient experience both €1,72 • 1st 'round': average pay/practice max. €15K (€7,5K on average, about 5-10% of practice income) • 3rd 'round': payment lowered from €3,20 to €1,50/insured	• Risk-adjusted capitation and fee-for-service	• Relative targets • Ascending bonus-levels from 25th percentile (6 tiers) • 3rd round: improvement explicitly incentivized (7 equally sized tiers)	• Fixed sum per point per patient • Attainment (60% weight): if > 25th percentile, then score of 0-100%. Scores averaged to subject and then to theme: 0 points if ≤25th percentile, then 6 groups • Improvement (20% weight): practices divided in 7 groups	• Annually • Improvement: practice feature scores available after 3 years
PCRM	• Positive	• "Very small % of GPs' income" • Prevention (enrolled patients): C$6.86/patient if documentation provided to government, up to C$11K for attainment • Utilization: C$18-20.5K depending on rurality of area	• Fee-for-service, risk-adjusted capitation, or combination	• Preventive care: 3-5 absolute targets per measure. Low target ranges from 15-85% based on provincial averages; upper target from 50-95% based on population targets • Utilization: absolute target	• Preventive care: fixed payment per patient per measure; fixed extra payment for attainment • Utilization: fixed payment if required level of services provision to required number of patients is reached (can also be a cost-level)	• Annually

Program	Positive or negative incentives	Incentive size	Base payments	Performance targets	Payment calculation	Payment frequency
PIN	• Positive	• Bonus potential unknown • Phase 1: data management: C$5K/ clinic, max. C$5K/GP, C$350K in total. Group: C$40K/ clinic, max. C$5K/GP, C$370K in total	• Fee-for-service	• Phase 1: typically 5 absolute targets/measure, large range (e.g., 20, 40, 60, 80, and 90%) • Phase 1: only 2 areas of performance measures	• Phase 1: increasing payment per measure if higher target is reached (e.g., 50, 65, 80, 95, 100% of maximal payment per measure)	• After demonstration period (last quarter of 2008 for phase 1)
PIP	• Positive	• Depends heavily on practice size/number of patients seen • 2008/09: average payment per practice was A$61.6K or A$19.7K/GP; 5% reached average A$426K (A$36K/GP) • 2009/10: A$282M paid to practices (A$57.8K/practice)	• Fee-for-service, PIP-payments separately paid	• Processes: mostly 1 absolute target (% of patients or number of services) and/or fixed payment per service provided. Some areas: ≥2 tiers • Targets may increase with practice size	• Varying schemes for the 13 areas • One-off or periodical (fixed amount or per standardized patient), payments may have a maximum per practice • Rural loading: 0-50% add-on based on geographical size of region and remoteness of practice	• Quarterly/ annually; semi-annually in one area • Intention: only quarterly
PMP	• Positive	• "Small % of income" (<5%) • If eligible, PHOs receive setup sum of NZ$20K + NZ$0,60/ member • Payments: up to NZ$6/member • PHOs and district health boards (DHB) may increase payments	• PHOs: risk-adjusted capitation (GPs mostly fee-for-service)	• Target per measure (annually set using national outline) • PHOs/DHBs may add targets • Target attainment for 4 measures also assessed only against high-need population	• % attainment of target from baseline • Targets assessed independently for a preset fraction the $6 • Payment weighted toward progress in high-needs areas	• Semi-annually for 11 measures; annually for 2 measures (5 month lag)
PQI	• Positive	• Maximally 8% of physicians' annual income • In 2007, for the physician with the highest score payment would represent 4% of average annual income for direct patient care	• Capitation	• Typically 3 absolute targets/ measure based on baselines • About half of the measures scored as % of physicians performing required number of services (e.g., Pap smears)	• Scores translated in points and then summed (max. 1,000), more points if higher target reached • Unclear how points are translated into payment	• Annually
QOF	• Positive	• Up to 30% of practice income	• Risk-adjusted capitation	• For each measure: sliding scale within absolute targets (typically 40% and 90%)	• Scores translated into points and then summed (max. 1,000 points) • Fixed amount per point (£125), fixed no. of points per measure	• Annually

CHAPTER 4

Economic evaluation of pay-for-performance in health care: a systematic review

With Martin Emmert, Heike Kemter, Susanne Esslinger, and Oliver Schöffski

European Journal of Health Economics, 2012, 13(6): 755-767.

ABSTRACT

Background: Pay-for-performance (P4P) aims to stimulate both a more effective and a more efficient delivery of health care. To date, evidence on whether P4P itself is an efficient improvement strategy has not been systematically analyzed.
Objective: To identify and synthesize the existing literature on the cost-effectiveness of P4P.
Data sources: English, German, Spanish, and Turkish language literature published between January 2000 and April 2010 was searched in six databases: Business Source Complete, the Cochrane Library, Econlit, ISI web of knowledge, Medline, and PsycInfo.
Study selection: Articles reporting economic evaluations of P4P initiatives published in peer-reviewed scientific journals were eligible for inclusion, with full economic evaluations simultaneously considering costs and effects being the prime focus. Comparative partial economic evaluations were included if (impacts on) costs were described and the study allows for assessment of effects. Both experimental and observational studies were considered.
Results: Nine studies met our inclusion criteria. Three studies were classified as full evaluations and six studies as partial evaluations. Based on the full evaluations, P4P cost-effectiveness could not be demonstrated. Partial evaluations showed mixed results. Overall, ranges in assessed costs and effects were narrow, and evaluated P4P-programs differed considerably in design. Methodological quality assessment of included studies showed scores between 32 and 65 percent.
Conclusion: The evidence on the cost-effectiveness of P4P is scarce and inconclusive. The small number and heterogeneity of studies hampers drawing strong conclusions. Sound economic evaluations of P4P-programs are needed.

4.1 INTRODUCTION

Healthcare systems around the world are characterized by inefficiencies in the delivery of care (Kizer, 2001; Institute of Medicine, 2001; Casalino, 2003; McGlynn et al., 2003). One possible reason for that lies in provider payment systems; incentives that foster efficient delivery of high-quality care are usually lacking (Endsley et al., 2004; Sorian, 2006; Wilson, 2007; Rowe, 2006; Rosenthal, 2008). A promising strategy to improve healthcare delivery is pay-for-performance (P4P), which has become increasingly popular worldwide (Rosenthal et al., 2007b; Christianson et al., 2008). In P4P, explicit financial incentives are provided to healthcare providers based on their scores on predefined performance measures. The method assumes that physicians' behavior can be influenced by how they are compensated. Indeed, the health economics literature provides ample evidence that financial incentives can change the way in which physicians practice (e.g., Hillman et al., 1989; Hellinger, 1996; Gosden et al., 2000; Town et al., 2004). In addition, P4P assumes that increasing adherence to evidence-based guidelines, more emphasis on prevention, and carrying out early diagnosis not only improves the quality of care, but may also curb costs growth (Wheeler et al., 2007). In practice, P4P is often applied in concert with other improvement strategies, such as public reporting of performance information (Lindenauer et al., 2007; Rhoads et al., 2009).

Although good quality care is an important goal of healthcare systems, resources are scarce, inevitably leading to trade-offs and priority setting. Therefore, it is important to address the cost-effectiveness of improvement efforts (Kahn et al., 2010). While some commentators assume P4P cost-effectiveness (Scott, 2007; Smith, 2007; Greene et al., 2008; Greene & Nash, 2009; SBEG, 2009), others are skeptical (O'Kane, 2007; Schatz, 2008; Bailit Health Purchasing, 2008; Peiro & Garcia-Altes, 2008; Hahn, 2009). To date, convincing evidence on the effectiveness of P4P is largely lacking. Although a positive impact was found in some studies, improvements have generally been small (Petersen et al., 2006; Scott, 2007). Other studies have come to heterogeneous conclusions (Armour et al., 2001; Roski et al., 2003; Amundson et al., 2003; Rosenthal et al., 2007b; Schatz et al., 2007; Campbell et al., 2009; Kahn et al., 2010), and still others found unintended effects (Schatz et al., 2007; Nelson, 2007; Ryan, 2009b; McDonald & Roland, 2009). Thus, although theoretically P4P has the potential to improve quality, this has not been convincingly confirmed in practice. Together with the complexity of P4P-program design and implementation (see chapter 2), these results cast doubt on whether P4P can be a cost-effective improvement strategy. To date, there has been no systematic analysis of the literature on the cost-effectiveness of P4P. In this paper, we review the literature on P4P cost-effectiveness as assessed through economic evaluations. P4P can be considered cost-effective when improved quality of care is achieved with equal or lower costs, or when the same quality is achieved with fewer

resources. Yet even if P4P leads to cost increases it may still be regarded cost-effective if quality increases are large enough.

4.1.1 Prior reviews

There are already some published reviews investigating both the effectiveness and, to some extent, the cost-effectiveness of P4P (Christianson et al., 2008; Greene & Nash, 2009; Mehrotra et al., 2009). Christianson et al. reviewed the P4P literature to derive lessons from evaluations of purchaser P4P-programs. However, their main focus was on effectiveness. In addition, no critical appraisal of included studies was conducted. Greene and Nash provided an extensive but more general overview of the P4P literature. They analyzed 100 studies by clustering them into categories. One category was labeled "cost analysis", to which three studies were assigned. However, they did not conduct a critical assessment of the studies. In addition, it is unlikely that their search procedure identified all relevant studies on P4P cost-effectiveness. Finally, Mehrotra et al. assessed the state of the evidence on P4P in the hospital setting. They identified one cost-effectiveness study, but potentially relevant studies on the cost-effectiveness of physician P4P-programs were excluded. Overall, prior reviews did not focus specifically on the cost-effectiveness of P4P; effectiveness has been the prime aim. In addition, investigations sometimes omitted studies conducted in the physician (group) setting. Accordingly, adopted search strategies do not ensure a comprehensive collection of relevant studies. Furthermore, where studies were identified, no detailed description or critical assessment of the studies was provided. Finally, the most recent date range was until 2008. It is likely that new results have been published since then, and while P4P has traditionally focused almost exclusively on improving quality, recent developments have moved purchasers to incorporate efficiency measures in their programs (Institute of Medicine, 2007; Robinson et al., 2009).

4.2 METHODS

4.2.1 Search strategy and study selection criteria

For this review, we adhered to guidelines and recommendations from the Cochrane Collaboration (Higgins & Green, 2008), the Institute for Quality and Efficiency in Health Care (2008), the Hannoveraner Konsensus (Graf von der Schulenburg et al., 2007), and the NHS Economic Evaluation Database (Graig & Rice, 2007). We started with electronic searches in the following six databases: Business Source Complete, the Cochrane Library (i.e., Central, DARE, NHS-EED.), Econlit, ISI web of knowledge, Medline (via PubMed), and PsycInfo. Search limits included studies written in English, Spanish, Turkish, or German published between January 2000 and April 2010. We conducted our searches in the different databases using the same terms (see Appendix 4.1 for the complete search history). Our search strat-

egy was segmented into two components. The first component referred to P4P. Search terms from recent studies were used (e.g., SBEG, 2007; Mehrotra et al., 2009; Eldridge & Palmer, 2009) and expanded. The second component, referring to cost-effectiveness, is consistent with the first one. Again, terms from recent studies were used (e.g., Leatherman et al., 2003; Kilpatrick et al., 2005; Parke, 2007; Rosenthal et al., 2007a; Reiter et al., 2007; Schöffski, 2008) and expanded. MeSH terms were used to broaden our search. We also used Google to obtain knowledge of unpublished studies and new P4P initiatives. In addition, websites of government and/or scientific agencies concerned with P4P (e.g., the Leapfrog Group and the Agency for Healthcare Research and Quality in the US) were consulted. Furthermore, comments, editorials, guidelines, interviews, letters, lectures, news items, and conference presentations were scanned to identify additional studies. Also, authors and other experts were contacted to provide potentially relevant ongoing or finished studies. Finally, reference lists of identified articles and previous reviews were screened.

Two authors independently reviewed titles and abstracts and checked them against several criteria (Table 4.1). When a study seemed to meet our criteria or when a final decision could not yet be made, the full text paper was obtained. In case consensus could not be reached, a third author was consulted. Published peer-reviewed articles were the prime focus, but we did not exclude unpublished studies beforehand. Both experimental and observational studies were eligible for inclusion provided that a quantitative assessment of (the impact on) costs was performed. For determining which types of evaluation to include, we used Drummond et al.'s (2005) categorization. In this framework, evaluations are classified based on whether costs, effects, or both were analyzed, and whether comparison is made between alternatives. We only included comparative evaluations. Our main focus was on full economic evaluations that simultaneously consider costs and effects, such as cost-effectiveness analyses (CEA) and cost-utility analyses (CUA). We label these evaluations

TABLE 4.1 Inclusion and exclusion criteria

	Inclusion criteria	Exclusion criteria
Language	English, German, Spanish, and Turkish	Other languages
Publication type	Articles published in peer-reviewed journals; unpublished studies not excluded beforehand	Comments, editorials, guidelines, interviews, letters, news, and conference presentations
Study type	Experimental and observational studies including a quantitative assessment of (the impact on) costs	Qualitative studies and studies only including an assessment of effects (e.g., the impact on quality)
Evaluation type	Comparative evaluations: full economic evaluations and partial economic evaluations	Non-comparative evaluations
Quality aspects	When effects are regarded: assessment of (or information on the development of) at least one process or outcome measure of quality	When effects were regarded: evaluation of only structural measures of quality
Targeted entity	Healthcare providers	Only patients

"Type I". Certain partial evaluations do, under certain conditions, also allow for making inferences on cost-effectiveness. First, there are studies describing the impact in terms of both costs and effects without making a connection between the two. This is the case, for example, when it is shown that quality improved and costs decreased, but a link between the effects is not provided. Thus, there are two separate analyses: a cost analysis and an effectiveness analysis. Such evaluations are included and labeled "Type II". Second, some studies only assessed the monetary impact without analyzing of the impact on quality. These studies can be characterized as cost comparisons and in principle do not allow for inferences on cost-effectiveness. Nevertheless, we included these studies as they may still provide relevant insights. We divided these studies into two groups: "Type III" (studies in which the authors also provide some information on how quality has developed or provide arguments for why an equal quality level can be assumed) and "Type IV" (studies only assessing the monetary impact, without providing a description of changes in quality). To be eligible for inclusion, the evaluated program had to contain an explicit financial incentive for healthcare providers related to their performance on predefined performance measures. Articles only describing a P4P-program and studies evaluating P4P-programs only focusing on improving structure quality measures (e.g., implementation of an electronic medical record) were excluded. Finally, we excluded studies focusing only the impact of financial incentives for patients (Volpp et al., 2009).

4.2.2 Study scoring procedure

To determine the methodological quality of studies, we applied a checklist containing 35 items grouped under ten categories (Drummond & Jefferson, 1996): study design, selection of alternatives, form of evaluation, effectiveness data, benefit measurement and valuation, costing, modeling, adjustments for timing of costs and benefits, allowance for uncertainty, and presentation of results. We added three items: "cost range" (category 6), "comparison with prior studies" (category 10), and "discussion on generalizability of results" (category 10). Score possibilities are "yes" (score 1), "no" (score 0), and "not appropriate" (N/A, seventeen items). For each study, the methodological quality was determined by calculating an unweighted average score. If the score was "N/A", the item was omitted from the calculation (Gonzalez-Perez, 2002). Two authors independently carried out the scoring. In case consensus could not be reached, a third author was consulted.

4.2.3 Relevant study characteristics

Studies are analyzed according to the context in which they were conducted, the design of the P4P-program, and the design of the study. The context is important since it provides information on the generalizability of the findings to other settings. In addition, contextual factors likely have important consequences for the success of the P4P-program (Nicholson et al., 2008; McDonald & Roland, 2009; Sutton et al., 2012). Relevant aspects include the

country and sector in which the program was implemented (e.g., inpatient, outpatient, primary care; public, private), the type of care to which it pertains (prevention, acute care, chronic care), and providers' base payment system. Whether or not the program functioned in concert with other interventions is also of relevance (SBEG, 2007).

In P4P, the way in which the incentives are structured likely have a significant impact on their (cost-)effectiveness (Rosenthal & Dudley, 2007; Damberg et al., 2007; Mehrotra et al., 2010a). A first design issue is the targeted performance dimensions, which may include clinical structures, processes, and outcomes (Donabedian, 1988), and patient satisfaction. Measuring and risk-adjusting outcomes are difficult (Dudley et al., 1998), so programs often use structure and process measures as proxies. Second, the targeted entity is of relevance. Targeting groups of physicians may be more effective in changing behavior than targeting individual physicians because of enabling factors such as essential infrastructure and peer review, and because larger sample sizes increase the likelihood of reliable performance measurement (Huang et al., 2005a; Mehrotra et al., 2010b; Adams et al., 2010a). On the other hand, overall response may be attenuated as a result of individuals free-riding on the efforts of peers. Third, the size of the incentive is important. Although increasing incentive size will increase provider response (up to a certain point), large payments are not necessarily more effective than smaller payments as they may "crowd out" providers' intrinsic motivation to provide high-quality care (Frey, 1997; Deci et al., 1999; Marshall & Harrison, 2005). Undesired behavior, for example neglect of unincentivized performance aspects (Holmstrom & Milgrom, 1991), may be the result. Fourth, penalties may be more effective and efficient than rewards (even if the same amount of money is at stake) because individuals tend to be risk- and loss-averse (Kahneman & Tversky, 1979). However, penalties may lead to negative reactions among providers (Damberg et al., 2007). Rewards may thus be preferred, although they require "new" money to finance the program and purchasers may be hesitant to invest in settings with substantial inefficiencies (Christianson et al., 2008). Fifth, whether payment is based on absolute performance, relative performance, and/or improvement in performance will also be important. Relative schemes encourage competition, but response may be attenuated since cooperation and dissemination of best practices is discouraged because of uncertainty regarding the efforts required to receive rewards. Absolute schemes have the disadvantage that the budget reserved for incentive payments becomes too small if more providers than expected reach the targets (Rosenthal & Dudley, 2007). Because an individual's response to an incentive depends on an assessment of the distance of the current performance from the required level (Heath et al., 1999), using multiple targets with larger payments for reaching higher targets may be most efficient (Damberg et al., 2007). Finally, the effect may be influenced by the degree of certainty for providers, frequency of payments, and duration of the program: certainty regarding the efforts required to reach targets and the attainability of these targets contributes to incentive strength (Conrad & Perry, 2009); monthly payments may yield a greater response than annual payments because people

discount future gains and because frequent payments increase incentive salience (Conrad & Perry, 2009); and durable incentives will likely elicit a stronger response as improving performance may be worth the effort while in short-term incentives it may not.

Finally, experimental studies yield more valid results than observational studies since they use more convincing control groups. Yet results of randomized controlled trials (RCT) may be difficult to generalize and thus less relevant for daily practice (Schatz et al., 2007). Also, characteristics of the patient population and the issuer and recipients of the incentive are of relevance. Sample sizes are important to report as they provide information on the reliability of measurements. In evaluating P4P, medical record data are preferred over claims data, although the latter are still often used since they are readily available. Finally, the type of evaluation and the adopted perspective are of interest, as well as information on measurement instruments and (statistical) methods used.

4.3 RESULTS

4.3.1 Search results and classification of included studies

The search strategy identified 1,644 articles. After title/abstract review and eliminating duplicates, 63 studies remained for detailed reflection. Screening of reference lists, expert consultation, and Internet searches yielded eleven additional articles. Thus, 74 full text articles were retrieved. Of these, nine met our inclusion criteria (Figure 4.1). As shown in Table 4.2, of the three full evaluations identified (Type I), one could be classified as a CUA and two as a

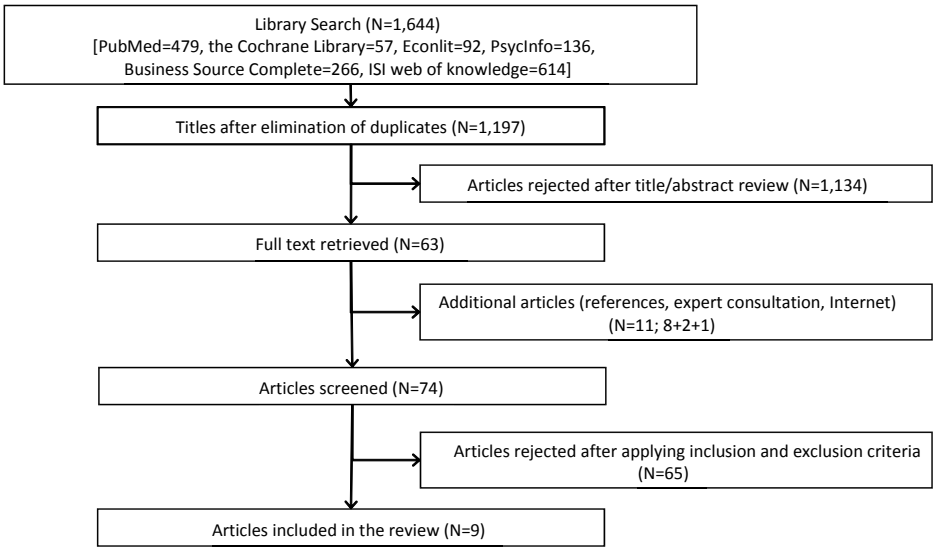


FIGURE 4.1 Search flow and results

TABLE 4.2 Classification and scoring of included studies

Study	Evaluation form	Category	Type [a]	Score [b]
Nahra et al. (2006)	Cost-utility analysis	Full evaluation	I	64.7%
An et al. (2008)	Cost-effectiveness analysis	Full evaluation	I	60.7%
Kouides et al. (1998)	Cost-effectiveness analysis	Full evaluation	I	63.0%
Lee et al. (2010)	Separate cost-effect evaluation	Partial evaluation	II	40.0%
Norton (1992)	Separate cost-effect evaluation	Partial evaluation	II	43.3%
Rosenthal et al. (2009)	Separate cost-effect evaluation	Partial evaluation	II	63.0%
Ryan (2009b)	Separate cost-effect evaluation	Partial evaluation	II	63.3%
Curtin et al. (2006)	Cost-comparison	Partial evaluation	IV	32.3%
Parke (2007)	Cost-comparison	Partial evaluation	IV	57.7%

a. Type I studies consider both costs and effects and explicitly link them to each other. Type II studies consider costs and effects without explicitly linking the two. Type III studies only evaluate costs but allow for inference regarding the (possible) impact on quality. Type IV studies only focus on costs.
b. Unweighted average score based on awarded points for 38 quality criteria (Drummond & Jefferson, 1996). If the response was N/A, which was possible for 17 items, the item was omitted from calculation of the score.

CEA. Six studies were partial evaluations; four evaluated costs and consequences separately (Type II) and two only assessed costs (Type IV). The methodological quality was highly variable with scores between 32 and 65 percent (Appendix 4.4). The number of items scored N/A varied between four and twelve.

4.3.2 Description and comparison of studies

Eight studies were conducted in the US and one in Taiwan. The latter program was implemented by the Bureau of National Health Insurance, the country's sole purchaser of care services providing universal coverage for a defined benefits package. These features are major differences with the US healthcare system. In the US, with the exception of public arrangements such as Medicare and Medicaid, health care and health insurance are largely provided in the private sector. In addition, there is competition among providers and among purchasers. Most P4P-programs in the US have been initiated by private entities, but public purchasers are increasingly enacting programs as well. Among the US studies included in this review, five programs were implemented by private purchasers and three by public purchasers (see Appendix 4.2 for summaries of included studies and Appendix 4.3 for detailed study descriptions and contextual information).

Full economic evaluations (Type I)

Nahra et al. (2006) conducted a CUA from an insurer's perspective to investigate the cost-effectiveness of a P4P-program for improving the quality of cardiovascular care in the hospital setting. Bonus potential for the 85 participating hospitals was 1.2 to 2 percent of the relevant DRG payments. Scores on three process measures were analyzed over a

period of four years and observed improvements were converted into quality-adjusted life years (QALY) gained using existing literature. Results one year after implementation were compared with results after two, three, and four years. The authors estimated a cost per QALY gained between $12,967 and $30,081. Costs included total bonus payments as well as incurred administrative costs for running the program. Although the cost range was still narrow, other studies typically did not include administrative costs. Important limitations of this study are the lack of a control group and the fact that the results from the first year (the comparison year) were possibly already influenced by P4P.

In an RCT with a time frame of ten months, An et al. (2008) investigated the impact of a P4P-program initiated by an insurer to increase referrals of smokers to a quitline. Primary care clinics were randomly assigned to a control (N=25) or an intervention group (N=24). Intervention clinics could receive a $5,000 lumpsum bonus for reaching 50 referrals. In addition, $25 could be earned for each additional referral. In total, intervention clinics made 1,042 additional referrals, resulting in 289 additional enrollees in a quit-smoking program. Analyzed costs include those for development and implementation of the program as well as the bonus payments. For the intervention and control groups, total costs were $95,733 and $8,937, respectively, so the incremental costs-effectiveness of the P4P-program was $83 per additional referral and $300 per additional enrollee.

Kouides et al. (1998) calculated the cost-effectiveness of a P4P-program sponsored by Medicare that focused on increasing influenza immunization rates among the elderly. Primary care practices were randomly assigned to a control (N=27) or an intervention group (N=27). Intervention practices received an additional $0.80 per shot (10 percent of the regular fee) if 70 percent of eligible patients were immunized, and $1.60 per shot for reaching 85 percent. Median bonus potential was $530 per practice, paid out at the end of the study. After four months, scores had improved more in the intervention group (10.3 percent vs. 3.5 percent, P=0.03). Total costs were $4,362, or $3.02 per extra immunization (1,443 more immunizations were observed in intervention practices). However, costs only consisted of the bonus payments. In addition, generalizability of results is limited as the program was implemented in the context of a large demonstration project including an extensive media campaign and extended opening hours of clinics.

Partial economic evaluations: separate cost-effect evaluations (Type II)

Lee et al. (2010) carried out a controlled before-after study to evaluate the impact of a national P4P-program for diabetes care in Taiwan. Hospitals/clinics with specialized physicians could voluntarily participate and enroll patients. Follow-up visits and physical exams and lab tests were encouraged by providing physicians with additional fees on top of regular fee-for-service payments (a detailed description of the program was not provided). Two-year claims data were analyzed for an intervention (N=12,499 patients) and a control group (N=26,172 patients). The intervention group exhibited larger increases in physician visits

and adherence to guidelines, and a smaller increase in hospitalizations. After accounting for the fact that inpatient costs increased in the comparison group but decreased in the intervention group, total costs increased by $104 per patient per year more in the intervention group. However, there was no clear analysis of program costs and no adequate adjustment seems to have been made for potential selection bias.

Norton (1992) conducted an RCT in the early 1980s to examine the impact of a P4P-program for nursing homes on access, residents' health status, and Medicaid expenditures. Nursing homes were assigned to a control (N=18) or an intervention group (N=18) and observed for over twelve months. The program consisted of three financial incentives: an admission incentive to admit sicker people ($2.5 to $28 per diem), an outcome incentive to improve residents' health (lumpsum; $126 to $370), and a bonus for timely discharges (lumpsum; $60 to $230). Using a Markov model, Norton found that the probability of death and hospitalization fell for most residents. Costs were up to 20 percent lower in the intervention group for all health states (except the most severe), mostly due to shorter length of stay (the average saving per stay was $3,000). Yet average daily costs increased by 5 percent in due to program costs and a larger variety of patients. In addition, program costs were not reported in detail and generalizability of results is likely limited.

Rosenthal et al. (2009) investigated the impact of a P4P-program in which both patients and their obstetricians or midwives were incentivized. Both could receive a $100 lumpsum bonus after delivery in case the patient entered care during the first trimester of pregnancy and completed regular visit thereafter. This controlled before-after study, conducted from a private insurer's perspective, evaluated the impact on low birth weight, neonatal intensive care admissions, and spending in the first year of life. After adjustment for unobserved selection, participation was associated with lower odds of admission and spending, but not with low birth weight. The authors estimated a reduction in spending in the first year of life of $235, but program costs were not taken into account. In addition, generalizability of findings is limited because study participants had much to gain from prenatal care and were likely more sensitive to offered bonus payments than other populations. Finally, the relative effects of patient incentives and provider incentives are unknown.

In the retrospective cohort study published by Ryan (2009b), the objective was to estimate the effect of the Hospital Quality Incentive Demonstration (HQID) on 30-day mortality and 60-day cost for acute myocardial infarction, heart failure, coronary artery bypass grafting, and pneumonia. The program was voluntary and contained 33 quality measures, including seven outcomes. For each condition, hospitals received a 2 percent add-on to the relevant annual DRG payment if they were in the top decile of hospitals; a 1 percent add-on was paid to hospitals in the second highest decile. In the third year, a penalty was added for underperforming hospitals. Data were available over the period 2001-2006 (the HQID started in 2003), and results were calculated by analyzing differences between participating, eligible but not participating, and ineligible hospitals. After adjusting for unobserved

selection, Ryan found no effect on risk-adjusted 30-day mortality and risk-adjusted 60-day Medicare costs. As costs of program administration were not included, the HQID was inefficient regarding reducing 30-day mortality.

Partial economic evaluations: cost comparisons (Type IV)

Curtin et al. (2006) conducted a before-after study to evaluate the financial impact of a P4P-program for diabetes care. The program varied year-end payments based on physicians' scores on quality, patient satisfaction, and efficiency measures, which were also publicly reported. To generate funds for the program, 10 percent of providers' capitation payments were withheld. From an insurer's perspective, return on investment (ROI) was calculated by comparing program costs with achieved cost savings. Cost data in the pre-intervention years (2000-2002) were used to estimate the costs that would have been incurred without the program in 2003 and 2004. Savings amounted to $1.9 million in 2003 and $2.9 million in 2004. Program cost totaled $1.15 million annually, so a positive ROI of 1.6 to 2.5 was estimated. Although the cost range was quite wide, costs were not described in detail. In addition, the study lacked a control group and it was not possible to disentangle the effects of P4P and public reporting. Finally, changes in quality were not analyzed.

Parke (2007) investigated whether P4P can lead to cost savings within a year. From a combined employer and plan member perspective, annual costs were compared before and after P4P implementation. Cost categories analyzed included hospital, doctor, pharmacy, administration, and other. The program incentivized both providers (up to 20 percent increase in payment) and their patients (rebate on copayment of $25 per visit with a maximum of $100 per year). Physicians were rewarded for using a web-based tool incorporating evidence-based guidelines, and for adherence to recommended care processes. Patients were rewarded for adhering to recommended treatment. After excluding outlier patients with costs over $30,000, savings totaled $166,272 (9.2 percent). The largest savings were achieved in hospital and pharmacy care, and total administration costs did not increase significantly. However, the study lacked a control group, did not take into account secular trends, and did not assess the impact on quality. In addition, it was not possible to disentangle the effects of physician incentives and patient incentives.

4.4 DISCUSSION

The goal of this review was to provide an overview of the evidence on the cost-effectiveness of P4P. Therefore, we conducted a systematic review of the literature. Nine studies were identified, three of which were full economic evaluations. Among the partial evaluations, four were separate cost-effect studies and two were simple cost comparisons. The results indicate that P4P has the potential to be cost-effective. Type I studies showed improvements

in quality against increases in costs. Type II studies evaluated broader sets of measures and more often analyzed changes in outcomes. Two studies showed quality improvements as well as cost increases. One study showed that savings may be possible while improving quality, while another possibly demonstrated P4P inefficiency. Both Type IV studies reported cost savings. Overall, six studies showed that P4P could lead to improved quality of care. In these cases, the question whether or not sufficient value for money was generated is to be answered by relevant decision makers.

Studies typically failed to include all relevant types of costs and/or effects or did not report in detail about them. Also, methodological flaws impede strong conclusions on cost-effectiveness. The three Type I studies (partially) examined program costs, but not the *impact* of P4P on cost of care. In addition, two of these studies evaluated only one process measure and one lacked a control group. Among the four Type II studies, two analyzed both the impact on cost of care and program costs, but details were not provided. Another Type II study (not incorporating program costs) failed to demonstrate decreased mortality or cost savings. The three studies finding cost savings are all partial evaluations, with one also finding a positive impact on quality (but again did not include the costs of program administration) and the other two only assessing the financial impact. Finally, it is often unclear to what extent evaluated measures were also influenced by other improvement efforts, such as financial incentives for patients, public reporting, and feedback to providers.

4.4.1 Influence of contextual factors and P4P-program design

The studies differ greatly regarding the setting, design of the evaluated P4P-program, and study design, which hampers drawing strong conclusions. Some studies evaluated programs with a time frame of less than a year (Kouides et al., 1998; An et al., 2008), while others observed effects over several years (Nahra et al., 2006; Ryan, 2009b; Rosenthal et al., 2009). Studies typically adopted the purchaser/insurer perspective, which was either public (Kouides et al., 1998; Ryan, 2009b; Lee et al., 2010; Norton, 1992) or private. Three programs targeted a chronic disease (Nahra et al., 2006; Lee et al., 2010; Curtin et al., 2006), while Ryan (2009b) evaluated a program with a focus on acute care. Other studies focused on primary prevention (Kouides et al., 1998; An et al., 2008), while the study from Rosenthal et al. (2009) also focused on secondary prevention. Providers in all but three programs (Norton, 1992; Rosenthal et al., 2009; Curtin et al., 2006) were paid on a fee-for-service basis for the care targeted, and at least in part also for other care. Regarding these factors, studies with favorable results do not seem to differ systematically from studies with less favorable results. Due to the small number and methodological limitations of included studies, it is not possible to identify contextual factors that contribute to P4P cost-effectiveness.

Regarding P4P-program design, authors often did not report relevant details. For example, program duration and whether or not other improvement efforts were in place could not always be derived. For aspects that were reported, the following can be observed.

There is much variation regarding the targeted quality measures; no study focused on a measure that was also evaluated in another study. Some studies focused only on process quality (Nahra et al., 2006; Kouides et al., 2008; An et al., 2008), and one study analyzed both process and (intermediate) outcome measures (Lee et al., 2010). Process measures are ideally included only in case of a positive relationship with outcomes (Donabedian, 1988; Roeg, 2005). One study failed to demonstrate such a link (Ryan, 2009b), while another did not describe it in detail (Lee et al., 2010). Despite the fact that outcomes are more difficult to influence by providers than process measures, the analysis did not show more favorable results among studies evaluating only process measures. Institutions (i.e., hospitals or nursing homes) were targeted in three programs (Norton, 1992; Nahra et al., 2006; Ryan, 2009b). Primary care clinics were targeted in one study (An et al., 2008), and the remaining five programs incentivized individual physicians. Of the programs that did not exclusively focus on institutions, P4P occurred often (but not always) through additional fees for each appropriately managed patient. Reliability analyses were typically not conducted or referred to. Overall, it is not possible to indicate the impact of the targeted entity on P4P cost-effectiveness (it is noteworthy that the programs that also rewarded patients for adhering to recommendations [Parke, 2007; Rosenthal et al., 2009] were relatively successful). In terms of incentive structure, payment size was small in some programs (1 to 2 percent of base reimbursement; Nahra et al., 2006; Ryan, 2009b), while in others it was quite significant (Kouides et al., 1998; Curtin et al., 2006; Parke, 2007; An et al., 2008). There is some weak evidence that larger payments increase (cost-)effectiveness, but comparison is difficult. Most programs rewarded absolute performance as well as improvement, either through enhanced fee-for-service (Parke, 2007; Rosenthal et al., 2009; Lee et al., 2010) or by providing larger rewards for reaching higher thresholds (Nahra et al., 2006; Kouides et al., 1998; Ryan, 2009b; An et al., 2008). The program that only used relative targets combined with penalties for low performers (Ryan, 2009b) was not successful in reducing 30-day mortality and 60-day costs. However, other studies of the same program found improved process quality that would not likely have occurred without the program (Ryan, 2009a). This discrepancy could be the result of a lacking link between (incentivized) processes and mortality, insufficient improvement in process quality, and/or improved processes as a result of better record keeping or gaming (Ryan, 2009b). There is no evidence suggesting that the potential penalty contributed to the negative finding because the other study using penalties (Curtin et al., 2006) found considerable savings. Finally, frequency of payments and program duration varied across studies. Payment frequency ranged from directly after care delivery (Parke, 2007; Lee et al., 2010; Rosenthal et al., 2009) to within a year (Kouides et al., 1998; An et al., 2008) to delays of over a year (Nahra et al., 2006; Ryan, 2009b; Curtin et al., 2006). Duration varied from four months (Kouides et al., 1998) to six years (Ryan, 2009b). There seems to be no clear relationship between frequency and duration of the

incentive and P4P cost-effectiveness, although the three programs without much delay in bonus payments were all relatively successful.

Overall, although variation in evaluated performance measures hampers a meaningful comparison of results, our findings tentatively indicate that increasing incentive size, rewarding absolute performance and/or improvement, simultaneously incentivizing patients, and minimizing the delay between care delivery and payment may contribute to P4P cost-effectiveness,

4.4.2 Implications

Although P4P has become a common part of provider payment systems, convincing evidence of its cost-effectiveness is lacking. Strong conclusions cannot be drawn due to variable methodological quality and differences among studies and evaluated programs in context and design. To enable more effective comparison, sound economic evaluations incorporating broad ranges of costs and effects and assessing programs over several years are required. In particular, economic evaluation of large P4P initiatives such as the Quality and Outcomes Framework (QOF) in the UK (Roland, 2004; Doran et al., 2006) and the Integrated Healthcare Association's program in California (Williams et al., 2009) would provide valuable information. Although it is likely that some QOF measures are cost-effective (Walker et al., 2010), to our knowledge there has been no comprehensive investigation (including an assessment of administrative costs) of the entire program. Finally, while P4P may be effective (although convincing evidence is still lacking), policymakers should keep in mind that other improvement strategies, such as disease management (Smith, 2007; Mattke et al., 2007; Steuten et al., 2007), performance feedback, and public reporting (Robinowitz & Dudley, 2006; Lindenauer et al., 2007; Werner & Bradlow, 2010), may provide more value for money.

4.4.3 Limitations

This review has several limitations. First, most studies were not conducted as economic evaluations in the common sense (Drummond et al., 2005). In part, this may be an explanation for the fact that cost ranges were typically narrow and that some studies has poor methodological quality scores. Second, the heterogeneity in methodological quality and study characteristics made comparison of results across P4P-programs problematic and hampered drawing strong conclusions. Third, as in any literature review, our study may suffer from publication bias. For example, sponsors of inefficient P4P-programs may have obstructed publishing the results. Finally, only peer-reviewed literature was included, so we may have missed potentially relevant work.

4.5 CONCLUSION

Without a doubt, P4P will continue to be a popular improvement strategy in health care. In addition to the US and the UK, public and private purchasers in other countries have begun or intent to implement P4P-programs. However, based on the available evidence, a definitive conclusion on P4P cost-effectiveness cannot be made. Economic evaluations incorporating broad ranges of costs and effects are needed to determine if P4P is worth the credit as suggested by its popularity in practice.

APPENDICES

Appendix 4.1 Literature search history

PubMed

#1: 10.05.2010; N=2,094

Search ("pay for performance" OR P4P OR PFP OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND quality) OR ("financial incentive" AND quality) OR ("monetary incentives" AND quality) OR ("monetary incentive" AND quality) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive") Limits: English, German, Spanish, published in the last 10 years Field: Title/Abstract

#2: 10.05.2010; N=311,046

Search ("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "return on investment" OR "roi" OR "rate of return" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros") Limits: English, German, Spanish, published in the last 10 years Field: Title/Abstract

#3: 10.05.2010; N=4,852

Search "Costs and Cost Analysis/economics"[mesh] OR "Cost Control/economics"[mesh] OR "Cost Savings/economics"[mesh] OR "Cost-Benefit Analysis/economics"[mesh] OR "Program Evaluation/economics"[mesh] OR "Health Services Research/economics"[mesh] OR "Utilization Review/economics"[mesh] OR "Efficiency, Organizational/economics"[mesh]

#4: 10.05.2010; N=479

Search #1 AND (#2 OR #3)

The Cochrane Library

#1: 10.05.2010; N=57 (CR=3; OR=1, CT=47, MS=0, TA=2, NHS EED=4, CC=0)

Search (pay NEXT for NEXT performance) OR (P4P) OR (PFP) OR (pay NEXT for NEXT value) OR (pay NEXT for NEXT quality) OR (payment* NEXT for NEXT quality) OR (value NEXT based NEXT purchasing) OR ("financial incentive*" NEAR/9 quality) OR ("monetary incentive*" NEAR/9 quality) OR (bonus NEAR/9 quality) OR (reward* NEAR/9 quality) OR (performance NEXT based NEXT payment*) OR (performance NEXT based NEXT reimbursement) OR (performance NEXT based NEXT contracting) OR (output NEXT based NEXT payment*) OR "incentive reimbursement" OR (quality NEXT based NEXT purchasing) OR ("quality incentive") OR ("quality incentives") OR ("quality-payment") OR ("quality-payments") OR (quality NEXT based NEXT payment*) OR (quality NEXT adjusted AND capitation) OR (payments NEXT for NEXT quality) OR "provider profiling" OR "value profiling" OR (value NEXT of NEXT care)

OR (value NEXT driven NEXT health NEXT care) OR (performance NEXT related NEXT payment*) OR (quality NEXT incentive NEXT payment*) OR ("performance contracting")

Econlit

#1: 10.05.2010; N=277

AB (("pay for performance" OR "p4p" OR "pfp" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND "quality") OR ("financial incentive" AND "quality") OR ("monetary incentives" AND "quality") OR ("monetary incentive" AND "quality") OR ("bonus" AND "quality") OR ("reward" AND "quality") OR ("rewards" AND "quality") OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive")) or TI (("pay for performance" OR "p4p" OR "pfp" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND "quality") OR ("financial incentive" AND "quality") OR ("monetary incentives" AND "quality") OR ("monetary incentive" AND "quality") OR ("bonus" AND "quality") OR ("reward" AND "quality") OR ("rewards" AND "quality") OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive"))

#2: 10.05.2010; N=93,659

AB (("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "return on investment" OR "roi" OR "rate of return" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros")) or TI (("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "roi" OR "rate of return" OR "return on investment" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros")

#3: 10.05.2010; N=92

#1 AND #2

PsycInfo

#1: 10.05.2010; N=1,287

TI (("pay for performance" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND quality) OR ("financial incentive" AND quality) OR ("monetary incentives" AND quality) OR ("monetary incentive" AND quality) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive")) or

AB (("pay for performance" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND quality) OR ("financial incentive" AND quality) OR ("monetary incentives" AND quality) OR ("monetary incentive" AND quality) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive"))TI (("pay for performance" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND quality) OR ("financial incentive" AND quality) OR ("monetary incentives" AND quality) OR ("monetary incentive" AND quality) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive"))

#2: 10.05.2010; N=86,455

TI (("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "return on investment" OR "roi" OR "rate of return" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros")) or AB (("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "roi" OR "rate of return" OR "return on investment" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros")

#3: 10.05.2010; N=136

#1 AND #2

Business source complete

#1: 10.05.2010; N=1,748

TI (("pay for performance" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND quality) OR ("financial incentive" AND quality) OR ("monetary incentives" AND quality) OR ("monetary incentive" AND quality) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive")) or AB (("pay for performance" OR "pay for value" OR "pay for quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND quality) OR ("financial incentive" AND quality) OR ("monetary incentives" AND quality) OR ("monetary incentive" AND quality) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based pay*" OR "performance-based reimbursement" OR "performance-based contracting" OR "output-based pay*" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive"))

#2: 10.05.2010; N=219,861

TI (("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "return on investment" OR "roi" OR "rate of return" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros")) or AB (("cost analysis" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation" OR "program evaluations" OR "economic evaluations" OR "economic evaluation" OR "financial analysis" OR "cost saving" OR "saving" OR "savings" OR "cost" OR "costs" OR "cost-efficiency" OR "cost-efficient" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "roi" OR "rate of return" OR "return on investment" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "qaly" OR "qalys" OR "dollars" OR "dollar" OR "pounds" OR "pound" OR "yen" OR "yens" OR "euro" OR "euros"))

#3: 10.05.2010; N=266

#1 AND #2

ISI web of knowledge

#1: 10.05.2010; N=2,298

TS=("pay for performance" OR "P4P" OR "PFP" OR "pay for value" OR "pay-for-quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND "quality") OR ("financial incentive" AND "quality") OR ("monetary incentives" AND "quality") OR ("monetary incentive" AND "quality") OR ("bonus" AND "quality") OR ("reward" AND "quality") OR ("rewards" AND "quality") OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive") AND Language=(English OR German OR Spanish OR Turkish)

#2: 10.05.2010; N=449

TI=("pay for performance" OR "P4P" OR "PFP" OR "pay for value" OR "pay-for-quality" OR "payment for quality" OR "value-based purchasing" OR ("financial incentives" AND "quality") OR ("financial incentive" AND "quality") OR ("monetary incentives" AND "quality") OR ("monetary incentive" AND "quality") OR ("bonus" AND "quality") OR ("reward" AND "quality") OR ("rewards" AND "quality") OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive") AND Language=(English OR German OR Spanish OR Turkish)

#3: 10.05.2010; N>100,000

TS=("cost and cost analysis" OR "cost control" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation$" OR "health services research" OR "utilization review" OR ("efficiency" AND "organizational") OR "economic evaluation$" OR "financial analysis" OR "cost analysis" OR "cost saving$" OR "budgetary analysis" OR "saving$" OR "cost off-sets" OR "cost$" OR "cost-efficiency" OR "cost-efficient" OR "revenue" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "benefit-cost" OR "ROI" OR "rate of return" OR "return on investment" OR "net present

value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "QALY$" OR "dollar$" OR "pound$" OR "yen" OR "euro$") AND Language=(English OR German OR Spanish OR Turkish)

#4: 10.05.2010; N>100,000

TI=("cost and cost analysis" OR "cost control" OR "cost savings" OR "cost-benefit analysis" OR "program evaluation$" OR "health services research" OR "utilization review" OR ("efficiency" AND "organizational") OR "economic evaluation$" OR "financial analysis" OR "cost analysis" OR "cost saving$" OR "budgetary analysis" OR "saving$" OR "cost off-sets" OR "cost$" OR "cost-efficiency" OR "cost-efficient" OR "revenue" OR "profit" OR "efficiency" OR "efficient" OR "cost-effectiveness" OR "cost-effective" OR "cost-utility" OR "cost-benefit" OR "benefit-cost" OR "ROI" OR "rate of return" OR "return on investment" OR "net present value" OR "benefit-cost ratio" OR "business case" OR "economic case" OR "social case" OR "quality-adjusted life years" OR "QALY$" OR "dollar$" OR "pound$" OR "yen" OR "euro$") AND Language=(English OR German OR Spanish OR Turkish)

#5: 10.05.2010; N=614

(#1 OR #2) AND (#3 OR #4)

Appendix 4.2 Summaries of included studies

Nahra et al. (2006) investigated the cost-effectiveness of a P4P-program focusing on the improvement of heart-care in the hospital setting. The program, initiated by Blue Cross Blue Shield of Michigan (BCBSM, a non-profit insurer) in 2000, provided financial incentives to 85 Michigan hospitals for reaching minimum levels of adherence to clinical guidelines. Payments were calculated as a percentage add-on to hospitals' inpatient DRG-reimbursements from BCBSM (with a maximum of 1.2 percent in 2000-2002 and 2 percent in 2003). In this study, only costs directly related to the P4P-program were considered, which consisted of total incentive payments and administrative costs. Quality effects were based on scores on three process measures, two for patients diagnosed with acute myocardial infarction (AMI) and one for patients diagnosed with congestive heart failure. Hospital performance was measured as the proportion of all eligible patients (not just patients insured with BCBSM) receiving recommended treatment. Based on existing literature, improvements in compliance were converted into QALYs gained over a 4-year period (2000-2003). 24,418 patients ultimately received improved care. Observed improvements in compliance were: aspirin for patients discharged after AMI from 87 percent to 95 percent, beta blockers for patients discharged after AMI from 81 to 93 percent, and ACE inhibitors for discharged patients after heart failure from 70 to 80 percent. Over this period regarding these measures, BCBSM paid out $21 million. Total program costs added up to over $22 million so about 5 percent reflected administrative costs of the program. Subsequently, the lower- and upper-bound discounted QALY estimates (733.3 and 1701.2, respectively) were related to total costs, resulting in a cost per QALY range of $30,081 to $12,967.

An et al. (2008) evaluated a P4P-program initiated by Blue Cross Blue Shield of Minnesota aimed at increasing clinician referral of smokers to a state tobacco quit-line. Forty-nine adult primary care clinics, all part of a large multispecialty physician network, were randomly assigned to a control group (N=25) and an intervention group (N=24). Intervention clinics received a $5,000 lump sum payment in case 50 eligible smokers were referred. In addition, for each referral exceeding the 50th, clinics received an extra $25. Statistical analysis showed no differences between both groups regarding clinic type, number of physicians, whether or not an EMR was in place, engagement with quality improvement, total number of patients seen, estimated prevalence of smoking among registered patients, and smokers seen. Only intervention group and quality improvement history were found to be independent predictors of referral rates. Overall, intervention clinics referred 11.4 percent of smokers seen while controls referred 4.2 percent (P=.001). Sub-group

analysis showed that differences in referral rates were only statistically significant in clinics with a history of being engaged (N=22, 10.1 percent vs. 3.0 percent, P=.001) or less engaged (N=18, 10.1 percent vs. 1.1 percent, P=.02) with quality improvement. Overall, 60.2 percent of referrals resulted in contact between patients and quit-smoking counsellors. Of these, 49.4 percent actually resulted in patient enrolment, corresponding to a mean enrolment rate of 27 percent. The overall percentage of smokers who were referred and enrolled in quit-line services was significantly higher in intervention clinics (3.0 percent vs. 1.3 percent, P=.005). More inappropriate referrals in the intervention group was not likely because rates of contact with referred smokers and subsequent enrolment did not differ with those observed in controls. Eleven intervention clinics received the $5,000 bonus and $25 was paid out for 619 additional referrals. This adds up to a total amount of remitted payments of $70,475. For intervention clinics, total costs, costs per referral, and cost per enrolee were $95,733, $65, and $232, respectively. For control clinics, these figures were $8,937, $20, and $72. For these extra costs, intervention clinics made 1,042 extra referrals (1,483 vs. 441) resulting in 289 additional enrolees. Thus, the incremental costs-effectiveness of the program was $83 per additional referral and $300 per additional enrolee.

Kouides et al. (1998) conducted a randomized controlled trial to investigate the effects of a P4P-program directed at primary care practices participating in the Monroe County Medicare Influenza Vaccination Project, developed to increase influenza immunization rates in the elderly. In this project, physicians tracked their own immunization rates on a weekly basis (from September 1991 to January 1992). In total, 54 practices participated, constituting about 30 percent of all primary care physicians in the community. Practices randomized to the intervention group (27 practices) received an additional $0.80 per shot (10 percent of the regular fee) in case a 70 percent of eligible patients had been immunized and $1.60 per shot in case 85 percent was reached. There were no statistically significant differences between intervention and control group regarding the distribution of the number of physicians in the practices, the median number of elderly patients during the study, the mean immunization rates at baseline (in 1990, about 58 percent), type of practice, patient demographics, estimated percentage of Medicaid patients, and reminder systems already in place. For intervention practices, the mean and overall immunization rates were higher compared to controls: 68.6 percent vs. 62.7 percent (P=.22) and 66.9 percent vs. 60.1 percent, respectively. Intervention groups were able to improve rates by 10.3 percent vs. 3.5 percent in controls (P=.03). In the intervention group, 52 percent of the practices reached the 70 percent threshold and 15 percent reached the 85 percent rate. For controls, these figures were 44 percent and 7 percent. A multiple regression analysis (R^2=.41) showed that only intervention group (7.1 percent increase) and baseline rate (+ 10 percent would result in a 4.6 percent decrease) were independent predictors of the change in immunization rate. Total performance payments added up to $4,362. Since about 1,443 more immunizations were observed in the intervention groups as compared to the control group, the incremental cost-effectiveness ratio was calculated as $3.02 per additional immunization.

Rosenthal et al. (2009) evaluated the impact of a P4P-program on low birth weight, neonatal intensive care unit (NICU) use, and healthcare spending in the first year of life. The program, initiated in 1999 by a private health insurer in Las Vegas covering about 135,000 lives with relatively low socio-economic status, sought to improve neonatal health and reduce the spending associated with NICU admission and sequelae of low birth weight. Both pregnant members and obstetricians or midwifes received a $100 bonus after delivery in case the patient entered care during the first trimester of pregnancy and completed regular visits thereafter. At baseline, 14 percent of the members received prenatal care in the first trimester, and the NICU use rate was above 10 percent. Over the study-period (1998-2001), a fivefold increase in adherence to recommended prenatal care to 73 percent of deliveries was observed. After having increased in 1998, overall rates of low birth weight and admissions declined from 5.7 percent and 7.7 percent in 1999 to 5.2 percent and 5.0 percent in 2001. In total, 3,590 deliveries were observed of which 1,436 were from program participants. Participants were less healthy than nonparticipants (measured by the presence of comorbidities), although overall

prevalence rates were low. Other maternal characteristics (e.g., maternal age, multiple births) did not differ between both groups, although some relevant information remained unobserved (e.g., smoking status and substance use). Unadjusted comparisons showed that participants had lower rates of low birth weight (4.5 percent vs. 6.0 percent, P=.04) and admissions (5.2 percent vs. 7.5 percent, P=.01) than nonparticipants. No difference was found in mean spending. The authors' instrumental variables model adjusting for the impact of voluntary program participation on birth outcomes and spending showed that participation was associated with lower odds of admission (0.45, CI=0.23 to 0.88) and spending (elasticity=-0.07, CI=-0.12 to -0.01), but not with low birth weight (0.53, CI=0.23 to 1.18). Other explanatory variables had the expected signs, but many were insignificant due to low prevalence of risk factors in the study population. Inclusion of hospital or physician fixed effects in the models resulted in only a significant effect of participation on spending (elasticity=-0.05, CI=-0.09 to -0.01). Although data on program costs were not available, the authors argue that the overall financial consequences may have been favorable from a payer's perspective because the bonus per delivery was $200 and the results suggest a reduction in spending in the first year of life of $235. In addition, payers may also benefit from higher rates of well-child visits and immunizations beyond the first year of life, which have been associated with prenatal care use in previous studies.

Ryan (2009b) investigated the effects of the Hospital Quality Incentive Demonstration (HQID) on Medicare patient mortality, costs, and outlier classification for AMI, heart failure, pneumonia, and CABG. The program, a voluntary public reporting and P4P-program for US hospitals, was a collaboration between Premier Inc. and CMS that started in October 2003. Of the 421 hospitals asked to participate, 266 decided to do so. For each included clinical condition, hospitals in the top decile of a composite quality measure received a 2 percent add-on on Medicare reimbursement rates; hospitals in the second highest decile received an additional 1 percent. Penalties for successive low performers since 2003 were imposed in 2006. Previous studies found a positive impact of the program on most process measures, but evidence on outcomes (e.g., mortality) and costs was inconclusive or lacking. In this study, data were available over six years (2001-6) on 3,570 acute care hospitals, of which 155 were eligible but refused to participate. HQID hospitals were different with regard to structural characteristics and the study outcomes than the other hospital cohorts, particularly non-eligible hospitals. Because differences in pre-intervention trends in risk-adjusted (RA) outcomes between HQID hospitals and comparison hospitals were small and not statistically significant, comparison hospitals could be viewed as counterfactuals for HQID hospitals. Using three econometric approaches to adjust for unobserved (time-varying and time-invariant) selection and other confounds, no significant effect was found on RA 30-day mortality for AMI, heart failure, pneumonia, or CABG. In addition, evidence that the program had a causal impact on 60-day cost was found to be very weak. Finally, while the effect on outlier classification was large and significant for heart failure and pneumonia in the models estimated among all hospitals (but not large enough to be reflected in an effect on 60-day cost), the other two estimators were non-significant for all conditions evaluated. The author concludes that although only one health outcome for only Medicare patients was analysed, study findings indicate that by not reducing mortality and cost growth, the program has not increased the value of inpatient care purchased by Medicare.

Lee et al. (2010) examined a national P4P-program for diabetes in Taiwan. The program was implemented in 2001 by the Bureau of National Health Insurance (NHI) in addition to four other P4P-programs for tuberculosis, breast and cervical cancer, and asthma. The program focused on hospitals and clinics with physicians qualified in metabolic specialty, who could voluntarily participate and enroll patients. Increases in comprehensive follow-up visits including self-care education and regular physical exams and lab tests were encouraged by providing physicians with 'enlarged physician fees' and 'case management fees', in addition to fee-for-service payments. Required and recommended services within these visits were delineated by the program. The study analyzed utilization and cost data over 2005 and 2006 for a group of diabetics that first participated in the program in 2006 (N=12,499) compared to a group of patients that never participated in

the program (N=26,172). The groups differed significantly on age, gender, and comorbidity; on average the intervention group was unlikely to be healthier than the control group. Regarding the number of essential exams/tests, the net effect (difference-in-difference) of the program on the completion of all 7 essential exams/tests was 2.5 (P<.001; the number of tests in the intervention group increased from 3.8 to 6.4). For the number of physician visits, the net effect was two (P<.001) and the intervention group had more visits (17.5) than the year before (fifteen). Regarding hospitalizations, the net effect was -0.027 (P<.003; observed increase not significant in intervention group, but significant for controls). Regarding costs, for physician visits the expenses in the intervention group increased by NT$8,462, resulting in a difference of NT$7,191 between the groups (P<.001; NT$1,270 of this amount was due to the program's management fees for the initial enrolment, follow-up, and annual evaluation visits). For diabetes-related inpatient services, expenses decreased in the intervention group and increased in the control group, which resulted in a net decrease of NT$3,878 (P<.001). Total diabetes-related costs (excluding and nursing home and home health care) were NT$3,312 higher in the intervention group (P<.001), or US$104 per patient per year.

Norton (1992) evaluated a P4P-program for nursing homes implemented in San Diego in the early 1980s. The objectives were achieving improvements in efficiency and in access and health of Medicaid clients. In a randomized controlled trial, eighteen nursing homes were given three financial incentives: a change in daily reimbursement rates from historical-cost based to case-mix based, bonuses for improved health within 90 days while admitted, and bonuses for timely discharges. Residents' health was assessed periodically by registered nurses over a 30-month period. In total, 3,215 observations were analyzed, of which 2,135 in the control group. Using a Markov model, the effects of the incentives were evaluated by estimating the effect on probabilities of admission and to go to other (health) states, length of stay, and cost savings for Medicaid. The program had beneficial effects on access, quality, and cost. More disabled people were admitted (controlling for age, sex, race, and marital status), mean and median length of stay decreased, and probability of death and hospitalization fell for most residents. Transition probability matrices differed between both groups (P<.001). In addition, savings added up to 20 percent per stay, primarily as a result of shorter stays; correcting for the distribution of health states at admission, the average saving per stay was $3,000. Because of excess demand, however, occupancy rates for nursing homes remained high, leading to an estimated increase of average daily Medicaid costs by about 5 percent due to program costs and a larger variety of patients. However, substantial savings in other sectors would likely be realized via reduced hospitalizations and excess demand for Medicaid patients as a result of the shorter length of stay. Although not explicated in detail, the author asserted that program costs were probably small compared to the gains in improved health and reduced hospital expenditures.

Curtin et al. (2006) calculated return on investment (ROI) of a P4P-program implemented in the context of a partnership (2000-4) between a non-profit health plan and an independent practice association in upstate New York. The program varied year-end payments to physicians based on their scores on quality, patient satisfaction, and efficiency measures. Scores were also publicly reported and physicians received feedback on their performance. Each year $12 to $15 million was distributed to about 3,700 physicians (specialists and generalists). This pool was filled mainly with funds generated by withholding about 10 percent of the IPA's capitation payments. The distribution for the average physician ranged between $6,000-18,000 per year. Although ROI was evaluated only for diabetes, the program also focused on other diseases and specialties. The authors used cost data from the pre-intervention years (2000-2) to estimate the costs that would have been incurred in 2003 and 2004 had the program not been implemented. These estimations were compared with actual costs in these years. Total cost was defined as the amount paid to all providers in all categories (inpatient hospital, outpatient facility, pharmacy, physician services, other) by both the plan and its members. In total, the cost of the program amounted to $1.15 million annually. Savings, however, added up to $1.9 million in 2003 and $2.9 million in 2004, yielding a positive return on investment of 1.6 to 1.0 and 2.5 to 1.0

in 2003 and 2004, respectively. Sub-analysis showed that cost decreased for all care components analyzed, and that the decrease was most significant for hospital costs (6.8 percent). Because the program encouraged increases in utilization and because savings (sufficient to cover the cost of the entire program) resulted from shifts in care for a single chronic disease, the authors characterize the results as impressive and the ROI-estimates as conservative. Unfortunately, however, the impact on quality of care and patient satisfaction was not evaluated.

Parke (2007) evaluated a P4P-program (implemented in 2004) for providers (enhanced fee-for-service) and their patients (rebate on copayments) on its impact on total spending. Savings were calculated from a combined employer (health plan) and member perspective. Parke hypothesized that financial incentives and web-based support for providers and patients could reduce utilization and encourage healthy behaviour. Providers were spurred to adhere to evidence-based guidelines and to provide health information to patients; patients were encouraged to follow recommended treatments. With the pre-intervention year serving as baseline, total costs decreased from $2,049,780 to $2,316,929 in one year (11.5 percent). After excluding outlier cases with costs higher than $30,000, savings totalled $166,272 (9.2 percent). The vast majority of savings were achieved through lower costs related to hospital and pharmacy care (about 50 percent and 16 percent of total costs at baseline, respectively). Total administration costs increased as a result of the program, but were negligible compared to achieved savings. As anticipated, cost related to physician services increased. Savings were also calculated for eighteen ICD-9 disease groupings. As cost reductions were observed in thirteen of these categories, Parke concluded that the overall cost reduction was not a result of outliers or random chance. Interestingly, for the city's perspective overall cost savings were achieved despite higher reimbursements per unit of service, rebates on copayments, increases in reimbursement per unit as a result of changing provider networks, and the fee paid in order to use the program. On the other hand, patient cost-sharing was expanded in the intervention year, which presumably reduced overall costs. Although it was not possible to disentangle the effects of all these changes, the program seems to have contributed considerably to the achieved cost reductions.

Appendix 4.3 Detailed study information

Study information	Context	Intervention design	Study design	Analysis
Reference: Nahra et al. (2006) *Study-period*: 2000-2003 *Critical remarks*: - many parameters based on existing literature - no differentiation possible between ages (e.g., regarding compliance) - narrowly defined costs and effects (e.g., indirect cost savings not estimated); - effects cannot be attributed to the P4P-program - cost-effectiveness ratios overstated as improvements probably also reflect secular trends	*Country/region*: USA, Michigan *Setting*: Inpatient care; private, non-profit insurer *Range of services*: Tertiary prevention, post-discharge chronic care *Parties involved*: - Blue Cross Blue Shield (BCBS) - hospitals and affiliated physicians - Robert Wood Johnson foundation - California Healthcare Foundation - Agency for Health care Research and Quality *Base payment system*: Diagnosis-related groups (DRG) *Background information*: - P4P-program was part of national Rewarding Results demonstration - BCBS subjected to charter form Michigan state specifying its governance structure	*Targeted entity*: hospitals *(Potential) payment size*: Maximally 1.2% to 2% of annual DRG payments by BCBS *Quality dimensions*: Impact on 3 process measures evaluated: % of eligible acute myocardial infarction (AMI) patients receiving aspirin at discharge; % of eligible AMI patients receiving beta blockers at discharge; % of eligible congestive heart failure (CHF) patients receiving ACE inhibitors at discharge *Structure:* - rewards calculated by multiplying performance scores by 1.2-2% of DRG payments - before 2002, hospitals were ranked on their scores and divided in quartiles with fixed payments. In 2002-2003, hospitals received a bonus if they reached the median of all hospitals. Improvement encouraged by shifting thresholds upwards *Financing:* New money *Frequency/duration*: Annual payments over a 4-year period *Other improvement efforts*: Multiple other incentive programs targeted the 3 evaluated measures	*Time-frame*: 4 years *Type of study*: Uncontrolled before-after study *Characteristics of study participants*: - issuer of incentive: 1 non-profit insurer (board comprised of customers (65%), care providers, and the public) - recipients of incentive: 85 Michigan hospitals - study population: hospitalized for AMI (on average 23,000 for all years for all hospitals combined) or CHF (38,000 patients). Hospitals treating more than 70 AMI or CHF patients were allowed to submit scores for a sample of patients *Data collection/source*: - costs: claims data - performance: self-reported, subjected to random audit - patient volumes and demographics: claims data and all-payer database maintained by Michigan Health and Hospital Association. - number of patients per diagnosis for 2001-3: via extrapolation - life years gained and QALYs: estimated using existing literature	*Type of analysis*: Cost-utility analysis *Perspective*: Insurer *(Non-)financial effects*: - financial: costs of P4P payments and program administration. - non-financial: number of patients receiving improved discharge orders in years 2001-2003 vs. 2000, total life year gains across the four years, and discounted QALYs gained *Measurement*: - life year gains: individual life-year gains from literature multiplied by estimated number of compliant patients receiving better care - QALYs: estimated using literature-based time-trade-off values of life years for patients who survived AMI or CHF *(Statistical) methods used*: - gains discounted by a 5% rate - sensitivity analysis to obtain lower- and upper-bound QALYs based on extremes found in literature - annual decline by 10% in post-discharge medication compliance rate was assumed, starting with 50%. A 20% rate was also applied as additional sensitivity analysis *Price-year*: 2000 US$

Study information	Context	Intervention design	Study design	Analysis
Reference: An et al. (2008) *Study-period*: September 1 2005 to June 31 2006 *Critical remarks*: - health effects and savings in (indirect) costs of tobacco counselling not analysed (literature finding health and economic benefits of tobacco cessation cited) - focus should be on increasing enrolment after referral to quitline - not possible to disentangle effects of financial incentives, goal-setting, and feedback - impact of relevant patient features not reported - no sensitivity analyses - costs per fax referral not justified	*Country/region*: USA, Minnesota *Setting*: Primary care; private, non-profit insurer *Range of services*: Primary prevention *Parties involved*: - Blue Cross Blue Shield (BCBS) - 5 major health plans - Clearway Minnesota (non-profit institute promoting quitline) - FPA, a multi-specialty network of physicians - primary care clinics *Base payment system*: Not reported *Background information*: - health plans and Clearway jointly created a referral system - health plans closely collaborated in addressing issues possibly limiting the effectiveness of the incentive (e.g., focus on all patients regardless of health plan coverage) - BCBS subjected to charter from state of Minnesota specifying its governance structure	*Targeted entity*: Adult primary care clinics *(Potential) payment size:* $5,000 for reaching 50 referrals and $25 for each referral beyond the 50th *Quality dimensions*: 1 process measure: rate of referrals to quit-line (number of unique referrals divided by estimated number of eligible patients) *Structure*: Reward for reaching absolute threshold (50 referrals) and additional referrals beyond 50th *Financing*: New money *Frequency/duration*: Single payment at end of the contract period, monthly score updates provided only to clinic administrators *Other improvement efforts*: Goal-setting and performance feedback simultaneously in place	*Time-frame*: 10 months *Type of study*: Randomized controlled trial *Characteristics* of study *participants*: - issuer of incentive: a non-profit insurer - recipient of incentive: 49 primary care clinics, 24 in intervention group. 49% family practice, rest internal medicine, obstetrics, multispecialty. 32 clinics used electronic health record, 9 clinics deemed very engaged with quality improvement - patients eligible for referral: smoking with intention to quit within 30 days, and ≥18 years of age. Intervention and control groups similar regarding no. of patients seen, smokers seen, and smoking prevalence. *Data collection/source:* - costs: claims data and researchers' tracking - smoking prevalence: patient survey - referrals: administrative data from quitlines - patient volumes, clinic features: claims data - engagement with quality improvement: FPA's subjective rating	*Type of analysis*: Cost-effectiveness analysis *Perspective*: Insurer *(Non-)financial effects*: - financial: costs of development, implementation, referrals, and incentive payments - non-financial: eligible smokers referred to quitlines, rates of contacted smokers, and rates of enrolment *Measurement*: - costs: development (time of physicians and staff of project, FPA and health plan) + implementation (information packages to clinics, feedback efforts to intervention clinics) + referrals (triage fees, staff time) + incentive payments - costs per fax referral: not reported - fixed costs affecting both groups divided equally across both groups - smokers per clinic: unique patients seen multiplied by estimated smoking prevalence in each clinic - hourly pay rates: based on annual salaries for participating staff *(Statistical) methods used*: - statistical testing performed at clinic level, comparison of outcomes using 2-sample t-tests - predictors of referral examined in one- and two-way ANOVA analyses - confidence intervals around estimates and P-values reported *Price-year*: US$ (year not reported)

Study information	Context	Intervention design	Study design	Analysis
Reference: Kouides et al. (1998) *Study-period*: September 1990 – January 1st 1992 *Critical remarks*: - only bonuses used as costs - low study power - generalizability low: other project in place and baseline levels high compared to national levels - study personnel not blinded; P4P groups were not contacted more often than control groups, but contact may have differed qualitatively - potential bias: underreporting of number of eligible patients - controls were aware they served as controls - no information on impact of patient features/not reported in detail - no discounting and sensitivity analysis	*Country/region*: USA, New York, Monroe County *Setting*: Primary, community-based care, Medicare *Range of services*: Primary prevention *Parties involved*: - primary care practices and staff - Medicare, staff of the Demonstration Project - study staff *Base payment system*: $8 fee per influenza immunization. System for other services not reported *Background information*: Providers participated in Medicare Influenza Demonstration, including media campaign, letters to patients, and extended opening hours of vaccination clinics. 1991-92 was the fourth year of the project. All physicians in this study participated and used a target-based poster model to track immunization rates among elderly on a weekly basis	*Targeted entity*: primary care practices/ physicians *(Potential) payment size*: $.80 or $1.60 per shot given in physicians' offices. Median patient number in 1991: 331. Median bonus potential: $530 per practice *Quality dimensions*: 1 process measure: final immunization rates (total number of shots per practice, wherever given, divided by number of eligible patients) *Structure*: For final immunization rates above an absolute threshold of 70%, a bonus of $0.80/shot. For 85% attainment, a bonus of $1.60/shot. Based on performance over 4 months (last quarter of 1991), and compared to prior year's scores *Financing*: New money *Frequency/duration*: Lumpsum bonus at the end of the study-period *Other improvement efforts*: Unknown	*Time-frame*: 16 months *Type of study*: Randomized controlled trial *Characteristics* of study *participants*: - issuer of incentive: Medicare - recipients of incentive: primary care practices (28 solo and 26 group, 47 private) with at least 50 patients, participating in the demonstration, that tracked vaccination rates during the 1990 flu season, and that had not participated in P4P before - patient population: non-institutionalized patients ≥65 years with office visit in 1990 or 1991. 22,000 patients in intervention group (median 325) and 17,600 patients in control group (median 432 in 1990 and 495 in 1991) *Data collection/source*: - data on target population: self-reported - No. of physicians per practice/ specialty, no. of patients, insurance types, % patients in Medicaid: staff interviews - performance: self-reported, checked with Medicare claims data	*Type of analysis*: Cost-effectiveness analysis *Perspective*: Medicare *(Non-)financial effects*: - financial: total incentive payments - non-financial: total additional immunizations *Measurement*: - in practices that used posters for each physician, group practice rate were calculated using a weighted average of individual rates (weights based on physicians' practice size) - shots given outside the office used for rate calculation, but not for bonus calculation *(Statistical) methods used*: - unit of randomization was practice, practices stratified by number of elderly patients (3 categories) - t-tests, Wilcoxon Rank Sum tests, chi-square tests, Fisher's Exact test, Shapiro-Wilcoxon test, and multiple linear regression. Regression model used change in % immunized as de-pendent variable and intervention group and potential confounders (e.g., % immunized in baseline year) as independent variables - P-values and 95% CIs reported. Residuals not normally distributed; alternatives (e.g., bootstrapping) yielded similar coefficients and significance estimates *Price-year*: US$ (year not reported)

Study information	Context	Intervention design	Study design	Analysis
Reference: Rosenthal et al. (2009) *Study-period*: 1998-2001 *Critical remarks*: - relative impact of patient and provider incentives unknown - speculative estimates on cost savings (e.g., no program costs) - results biased since utilization for non-participants could not be measured - chance of bias remains despite instrumental variables; omitted factors (e.g., ethnicity, smoking, substance use) may be are associated with participation - generalizability low; low-income, largely Hispanic population with much to gain from prenatal care and which may be more sensitive to incentives than other populations	*Country/region*: USA, Nevada, Las Vegas *Setting*: In- and outpatient care, 1 private health insurer *Range of services*: Prenatal/obstetric care, primary and secondary prevention *Parties involved*: - Culinary Health Fund Las Vegas - patients and their obstetricians/midwifes - the Commonwealth Fund (grant) *Base payment system*: Obstetric care paid using bundled case rates varying only by type of delivery *Background information*: - program intensively advertised to patients and providers by mail, phone, and newsletters - insurer already had a high-risk maternity management program in place	*Targeted entity*: Pregnant women and obstetricians/ midwifes *(Potential) payment size*: $100 bonus after delivery per appropriately treated member *Quality dimensions*: Process measures: yes/no entered care, yes/no visits thereafter *Structure*: Voluntary program offering rewards to pregnant women and their obstetricians or midwives after delivery if the patient entered care in first trimester of pregnancy and had regular visits thereafter *Financing*: New money *Frequency/duration*: 1 lumpsum bonus after delivery *Other improvement efforts*: Not reported, but likely none	*Time-frame*: 4 years *Type of study*: Controlled before-after study (natural quasi-experiment) *Characteristics* of study *participants*: - issuer of incentive: 1 union-sponsored insurer covering 135,000 lives - recipient of incentive: pregnant women and obstetricians/midwifes - patient population: privately insured, low-income, mostly Hispanic, relatively low comorbidity rates. Mean maternal age was 30. Of the 3,590 deliveries, 40% participated in study *Data collection/source*: - baseline levels: medical management claims data - CPT-4 and ICD-9 diagnosis codes for identifying deliveries, low birth weight, pregnancy-related and -unrelated comorbidity: claims data - Elixhauser index: diagnosis data from in- and outpatient sources - admissions: claims data on room type, CPT, DRG, and revenue codes, and birth and admission dates	*Type of analysis*: Separate cost-effect evaluation *Perspective*: Insurer *(Non-)financial effects*: - financial: spending in 1st year of life - non-financial: rates of low birth weight and admission rates *Measurement*: - admissions <1 day excluded - analyses restricted to cases where claims/ enrolment data for mothers could be matched with claims/ enrolment data for infants - covariates in regression analyses selected based on literature (e.g., Elixhauser index for comorbidity) *(Statistical) methods used*: - unadjusted differences analysed using t-tests and chi-squared tests - linear regression to analyse outcomes as function of risk factors - logistic regression to examine low birth weight and admissions - P-values and CIs reported - spending: generalized linear model with log link and gamma distribution - models adopted linear time trend and adjustments for clustering - two-stage instrumental variables approach to account for selection bias. Extra checks using models with a hospital or physician fixed effect *Price-year*: US$ (year not reported)

Study information	Context	Intervention design	Study design	Analysis
Reference: Ryan (2009b) *Study-period*: September 2000-September 2006 *Critical remarks*: - one outcome for Medicare patients - time-frame for mortality and costs narrow, although adherence to process measures may have short-term impact on mortality and costs - if unobserved patient severity was not constant over time, the risk-adjusters may have been too limited - the specification assumed that P4P had no effect on non-targeted conditions - small number of eligible hospitals limited power to detect small effects - no information on program costs	*Country/region*: USA *Setting*: Inpatient care, public insurer (Medicare) *Range of services*: Acute care, tertiary prevention *Parties involved*: - Premier Inc. - CMS - AHRQ (grant) - Jewish Healthcare Foundation (grant) *Base payment system*: Medicare Diagnostic-related groups (DRG) *Background information*: - Medicare has been implementing P4P since the early 2000s - program also applied public reporting of quality scores	*Targeted entity*: Acute care hospitals *(Potential) payment size*: 2% of DRG-payments if in top decile, 1% of payments if in second highest decile *Quality dimensions*: 33 measures, 7 of which are outcomes. Composite measures for heart failure, acute myocardial infarction (AMI), pneumonia, coronary artery bypass grafting (CABG), and hip and knee replacement (latter not evaluated) *Structure*: Relative scheme; 2% additional payments for top decile and 1 % for second highest decile. 1% penalty for structurally low performers. Also applicable to non-Medicare patients *Financing*: New money *Frequency/duration*: Started in October 2003, ended in 2009. Bonus payments paid out annually as add-on to DRG-payments *Other improvement efforts*: Unknown	*Time-frame*: 6 years *Type of study*: Controlled before-after study (natural quasi-experiment) *Characteristics* of study *participants*: - issuer of incentive: CMS and Premier. - recipient of incentive: 266 hospitals. Data available on 3,570 unique hospitals. Participating hospitals differed from other cohorts, particularly non-eligible hospitals, both regarding structural characteristics (number of beds, region, medical school affiliation, specialty units) and outcomes - patient population: 11,232,452 admissions from 6,713,928 patients *Data collection/source*: All data from Medicare: - inpatient claims (principal diagnoses, secondary diagnoses, type of admission for risk adjustment, cost data, and discharge status to exclude transfer patients) - Medicare Denominator File (to add risk-adjusters and to determine 30-day mortality) - Medicare Provider of Service file (hospital characteristics)	*Type of analysis*: Separate cost-effect evaluation *Perspective*: Medicare *(Non-)financial effects*: - financial: Medicare 60-day costs - non-financial: Medicare patient 30-day mortality *Measurement*: - 60-day cost per admission: hospital costs incurred in the 60 days post admission. Costs attributed to hospital that admitted the patient - Risk-adjusted outcomes: observed divided by expected per hospital per condition multiplied by overall mean for condition. Risk-adjusters: age, sex, race, 30 comorbidities, and type and season of admission. *(Statistical) methods used*: - expected outcomes: patient-level logistic and linear regression - econometric approaches: (1) hospital fixed effects (if unobserved selection constant), (2) time-effects: effect of program relative to hospitals interested in improving quality, (3) in-hospital outcome differences between clinical conditions included/not included before/after start of the program - fixed-effect linear trend models - sensitivity analysis: excluding 10% patients with highest predicted mortality probability - heteroskedasticity taken into account by estimating hospital-level cluster-robust standard errors *Price-year*: not reported (US$)

Study information	Context	Intervention design	Study design	Analysis
Reference: Lee et al. (2010) *Study-period*: 2005-2006 *Critical remarks*: - selection effects not accounted for - lack of a detailed time record of enrolment limits program effect - unclear if more utilization resulted in better outcomes; lab results on e.g., HbA1C and lipid profile could not be obtained - no pre-intervention trend data analysed - generalizability to other healthcare systems limited - source of data not always reported - increased hospitalizations in comparison group unexplained - no discounting/ sensitivity analysis - unclear if physicians in intervention group also treated control patients - program costs not analysed - aspects of P4P design unknown	*Country/region*: Taiwan *Setting*: in- and outpatient care, public insurer *Range of services*: diabetes care; primary, secondary, and tertiary prevention *Parties involved*: - bureau of National Health Insurance - hospitals and community clinics *Base payment system*: mostly fee-for-service; 53 procedures reimbursed via fixed payment schedules *Background information*: - Universal insurance system implemented in 1995. It is the sole payer of healthcare services; more than 90% of all providers providing care to >99% of population contract with NHI - the average number of physician visits per person per year is 15 - there is no referral system in place - 4 other P4P schemes in place for 4 diseases: tuberculosis, breast cancer, cervical cancer, and asthma	*Targeted entity*: Physicians in hospitals or community clinics *(Potential) payment size*: not reported *Quality dimensions*: process measures: increased number of follow-up visits (e.g.,, enhanced self-care education) and increased number of annual physical exams (e.g.,, eye exams and lab tests) *Structure*: voluntary, implemented by NHI in 2001. Physicians enrol patients into the program. In addition to base payments, physicians are paid additional 'enlarged physician fees' and 'case management fees'. The latter cover initial enrolment, comprehensive follow-up and annual evaluation visits. Required/ recommended services included in these initial and follow-up visits defined by the program *Financing*: not reported *Frequency/duration*: not reported *Other improvement efforts*: 4 other P4P-programs in place for 4 other conditions	*Time-frame*: 2 years *Type of study*: controlled before-after study (natural quasi-experiment) *Characteristics* of study *participants*: - issuer of incentive: National Health Insurance (>90% of care) - recipient of incentive: hospitals and clinics with physicians qualified in metabolic speciality - patient population: enrolled in 2006, diagnosed with diabetes each year between 2004-6 and filed drug claims for at least 3 months in each year; N=12,499. Diabetics who never joined since 2001 randomly sampled to create control group (N=26,172). Intervention group more often female, younger, and had more comorbidities compared to controls. Overall mean age was 62.8 and over 50% of all patients had a Charlson Comorbidity Index of ≥ 2. *Data collection/source*: - utilization and expenses: NHI claims database - diabetics not in program since 2001: not reported - patient information: not reported	*Type of analysis*: separate cost-effect evaluation *Perspective*: Bureau of National Health Insurance *(Non-)financial effects*: - financial: total costs and annual costs for diabetes-related physician visits and inpatient services - non-financial: annual number of essential tests, diabetes-related physician visits, and hospital admissions *Measurement*: - costs: fees for medication, physicians, exam-tests for outpatient visits, and other services (e.g., inpatient rehabilitation services). Nursing home and home health care excluded - 7 essential tests, e.g., for blood glucose, HbA1C, and lipid profile *(Statistical) models/methods used*: - difference-in-difference models. Distributions per outcome: number of tests, Poisson; number of physician visits and hospitalizations, negative binominal; costs, normal - generalized estimating equations with proper distribution, taking into account the correlation between repeated measures for each patient - standard errors of the differences and the difference-in-difference of predictions estimated by bootstrap technique. Estimates obtained via 100 replications with repeated samples of same size *Price-year*: 2006 New Taiwan $ (1US$=32NT$)

Study information	Context	Intervention design	Study design	Analysis
Reference: Norton (1992) *Study-period*: November 1980-April 1983 *Critical remarks*: - Markov model: strong assumptions (e.g., constant probabilities) - no pre-P4P data - data on program costs not reported - no sensitivity analysis, although results based on full sample may be understated as incentive effects were found to be larger when based only on data from second half of experiment (i.e., after the homes had adjusted to new system) - generalizability not discussed and likely limited - data source not always reported - sensitivity analysis could have been extended to test the impact of model assumptions	*Country/region*: USA, California, San Diego *Setting*: Inpatient care, for-profit nursing homes, public insurer *Range of services*: Nursing home care, elderly care, tertiary prevention *Parties involved*: - Medicaid - nursing homes - registered nurses - Agency for Health Care Policy Research - external research bureau (data collection and study supervision) *Base payment system*: Case-mix and outcome independent per diem rate of $36 (in 1981$) *Background information*: - first study to assess whether more efficient use of nursing homes can reduce total healthcare expenditures - prior studies already looked at effect of each incentive in isolation and found small positive effects on admissions, but no difference in outcomes and discharges	*Targeted entity*: Nursing homes *(Potential) payment size*: - admission: $2.5-28 per case - outcome: $126-370 per case - discharge: $60-230 - Only Medicaid patients *Quality dimensions*: 1 outcome (health status), 1 process (timely discharge), and 1 access measure (admission of severely dependent patients) *Structure*: - admission incentive: per diem bonus/malus based on wages and costs for increased nursing care coverage. Increased if health declined unavoidably - outcome incentive: bonus for attainment of health goals within 90 days during stay. Based on costs and wages for additional care - discharge incentive: bonus for timely dis-charge, based on costs of discharge and maintaining a free bed *Financing*: New money *Frequency/duration*: Admission bonus continued for up to 4 years; others only in study period. Frequency not reported *Other improvement efforts*: Not reported, but likely none	*Time-frame*: 30 months *Type of study*: Randomized controlled trial with Markov model *Characteristics* of study *participants*: - issuer of incentive: Medicaid - recipient of incentive: 36 for-profit, Medicaid certified, nursing homes with ≥30 beds. Homes matched according to location and number of admissions, and randomly assigned to intervention or control groups. Controls received a nominal payment to cover bookkeeping costs - patient population: 11,389 residents, 6,621 of which covered by Medicaid. Complete data for 3,215 residents with debilitating acute and chronic conditions. Mean age was 80 and women, whites, and married people were overrepresented *Data collection/source*: - residents' health states and demographics: recorded by nurses - costs: not reported - other data (e.g., length of stay): estimated	*Type of analysis*: Separate cost-effect evaluation *Perspective*: Medicaid *(Non-)financial effects*: - financial: Medicaid costs in terms of average daily rate per case - non-financial: changes in health, lengths of stay, mortality rates, timely discharges, and admissions of severely dependent patients *Measurement*: - health: nurses classified residents' health based on 6 activities of daily living, if discharge was likely in 90 days, and if special care was needed - expected costs per person per stay: program-related costs not included - average daily costs per resident: total costs including program costs divided by the no. of Medicaid days *(Statistical) models/methods used*: - Markov model to estimate effect of incentives simultaneously. 9 health states were possible. - Wald test to compare transition probability matrices for both groups - measurement errors did not change estimates of probabilities of leaving the nursing home - multinomial logit with age, race, sex, and marital status as covariates, 5 health states as outcomes *Price-year*: 1982 US$

Study information	Context	Intervention design	Study design	Analysis
Reference: Curtin et al. (2006) *Study-period*: 1999-2004 *Critical remarks*: - no data on impact on quality and patient satisfaction - no control group - generalizability not discussed and likely limited - methods for cost calculation not motivated - only averages, no details on variation - declining HMO population 2000-4; no comparison between members who stayed and who left - unexpected and -explained variation in trends; savings may have been a result of unexplained variation in costs instead of P4P - no analysis on impact of plan, physician, and patient features - no discounting/ sensitivity analysis - information on structure of intervention (e.g., relative vs. absolute targets) not reported	*Country/region*: USA, New York, Rochester *Setting*: Non-profit insurer, commercial IPA HMO *Range of services*: Diabetes care, primary/ secondary prevention *Parties involved*: - RIPA - Excellus Health Plan - Rewarding Results participants *Base payment system*: RIPA paid by capitation. Individual practitioners: unknown *Background information*: - Rewarding Results provided grant to the Excellus-RIPA partnership with purpose of expanding the incentive program to include performance measurement of care for chronic diseases. Physician profiles developed/ disseminated. Adherence to profiles supported by P4P program in place since 2003 and applicable to several specialties incl. diabetes, coronary disease, asthma, and internal medicine. - data periodically reviewed with practitioners and provided assistance with improvement efforts	*Targeted entity*: Primary care physicians (PCP) (this study) *(Potential) payment size*: For the average PCP, the distribution ranged between $6,000 and $18,000 per year. Annual bonus pool between $12-15 million. *Quality dimensions*: Profiles: scores on cost, quality, and satisfaction measures. Diabetes: 5 process measures *Structure*: 10% of withheld funds distributed to practitioners at year-end based on scores on specialty-specific measures. Withholds constituted the majority of funds but also included funds from gain-sharing between plan and IPA. Physician participation nearly guaranteed by withholding and gain-sharing incentives *Financing*: Old and new money *Frequency/duration*: Performance feedback three times a year; annual bonuses *Other improvement efforts*: Public reporting, but no new disease-management program, (nor did other agencies initiate new programs for improving diabetes outcomes)	*Time-frame*: 6 years *Type of study*: Uncontrolled before-after study *Characteristics* of study *participants*: - issuer of incentive: Excellus (non-profit health plan with 2 million enrolees in New York) - recipient of incentive: RIPA, a physician organization with 3,700 practitioners in the Rochester area (1,000 PCPs; 1,700 specialists; 1,000 psychologists, social workers, and physiotherapists) providing care to 345,000 HMO members in 2002 - study population: HMO members with diabetes, between 18-75 years old, continuously enrolled for 2 years, with any of the following criteria: ≥2 ambulatory visits with diabetes diagnosis, ≥1 emergency room visits for diabetes, ≥1 inpatient admissions for diabetes, ≥1 prescriptions for insulin, hypoglycaemic medication, or anti-hyperglycaemic medication. Diabetic member months ranged from 173,900 (2004) to 245,062 (2001) *Data collection/source*: - costs: claims data - other improvement efforts: survey data	*Type of analysis*: Cost-comparison *Perspective*: Insurer *(Non-)financial effects*: - financial: cost of program and care - non-financial: none *Measurement*: - program costs: resources for measurement; staffing to produce reports; involving physicians in design and maintenance of measurement; supporting efforts; software and staff time in leadership, analysis, provider relations, and quality management - cost of diabetes care: physician, inpatient, outpatient, pharmacy, and "other"; calculated as allowed amount (including patient cost-sharing) and actual amount. Costs per member per month (PMPM) per care type calculated by dividing allowed amount by member-months *(Statistical) methods used*: Cost projection made for 2003-2004 based on 2000-2002. Trends calculated by dividing annual PMPM by previous years' PMPM. Total annual trend: trends of each care type added, with weights based on the share in total PMPM over the 4 years. Average 2-year trends then calculated to account for volatility in pre- and post-intervention years; this was used to predict expected cost of next year by multiplying by actual PMPM of previous year. That value was compared to that year's actual value. Cases with annual cost ≥$100,000 were excluded. *Price-year*: Not reported

Study information	Context	Intervention design	Study design	Analysis
Reference: Parke (2007) *Study-period*: August 1 2003-July 31 2005 *Critical remarks*: - no evaluation of impact on quality - not possible to disentangle effects of physician incentives, patient incentives, web-based tool, and other concurrent changed (e.g., increases in patient cost-sharing) - no control group - no descriptive information key stakeholders - pre-intervention trend not analysed - generalizability likely limited - unclear if all cost savings can be attributed to P4P - no discussion of limitations - no analysis on the impact of plan, physician, and patient features - limited sensitivity analysis	*Country/region*: USA, Oklahoma, Duncan *Setting*: City (employer) health plan (commercial) *Range of services*: Not reported *Parties involved*: - city of Duncan - third-party administrator - program staff - University of Oklahoma (grant/support) *Base payment system*: Fee-for-service *Background information*: - program designed by a preferred provider organization founded by the author in 2000 - program is internet-based with evidence-based guidelines and patient information - interested plans/ employers pay $2.50 per member per month - implemented as part of a benefits program for employees and their dependents - prior to the program, costs had increased by 10% and physician payment per unit had decreased by 5% to 10%	*Targeted entity*: Individual providers (hospitals, physicians) *(Potential) size*: 10% of income. 50% difference between fees for compliance and non-compliance *Quality dimensions*: Adherence to 117 evidence-based guidelines *Structure*: Voluntary program with rewards and penalties for providers (enhanced fees) and patients (rebate on copayment of $25/visit, max. $100/year) to comply with recommendations. *Financing*: New money *Frequency/duration*: Frequency high (enhanced fee-for-service); duration unknown *Other improvement efforts*: - deductible increase for health (from $250 to $600) and pharmacy benefits (70-110%) - addition of office visit to benefits plan ($25/visit copayment) - reduced in-network providers to which 80% of allowed services is paid; increased out-of-network providers to which 50% is paid	*Time-frame*: 2 years *Type of study*: Uncontrolled before-after study *Characteristics* of study *participants*: - issuer of incentive: city of Duncan (self-insured) - recipient of incentive: no detailed provided - study population: non-Medicare, relatively old, often chronic disease with medical crisis. 462 of 1,054 patients who were given opportunity to use program did so. Total no. of claims in baseline and intervention year 8,262 and 8,505, respectively. *Data collection/source*: - costs: claims data - participation statistics, total prescriptions, diagnosis and specialty information, and number of patient rebates: program data system - qualitative data: telephone interviews with patients, physicians, and administrators	*Type of analysis*: Cost-comparison *Perspective*: Both employer (city) and members *(Non-)financial effects*: - financial: program costs and total healthcare costs - non-financial: none *Measurement*: - radiology chosen as surrogate for practice of "defensive medicine" - program costs: administration (e.g., operation and monitoring), link to payment, training of patients and physicians, and validation of compliance; provider and patient payments; license fees - total care costs: city's and patients' spending on: physician, hospital, dental services; drugs; supplies; pharmacy benefits. Costs classified by surgery, radiology, pathology, and e.g., visits, and segmented by sums paid by city and members - savings: total claims in intervention year (2004-2005) minus total claims in baseline year (2003-2004) *(Statistical) methods used*: - adjustment for changes in employment - comparison also for data in which members exceeding the stop-loss limit ($30,000/year) were excluded - ICD-9 ranges used to see if savings were referable to certain diseases *Price-year*: Not reported

Appendix 4.4 Methodological quality assessment results

Checklist criteria	Possible scores	Nahra et al. (2006)	An et al. (2008)	Kouides et al. (1998)	Rosenthal et al. (2009)	Ryan (2009b)	Lee et al. (2010)	Norton (1992)	Curtin et al. (2006)	Parke (2007)
1. Study question										
The research question is stated (e.g., are alternatives compared?)	0;1	1	1	1	1	1	1	1	1	1
The (economic) importance of research question is stated (both cost and effects?)	0;1	1	1	1	1	1	0	1	0	1
The viewpoint(s) of the analysis are clearly stated and justified	0;1	1	0	1	1	0	1	1	0	1
2. Selection of alternatives										
The rationale for choosing the alternatives compared is stated (alternatives omitted?)	0;1	0	0	0	0	1	1	1	0	0
The alternatives being compared are clearly described	0;1	0	1	1	1	1	0	1	1	1
3. Form of evaluation										
The form of economic evaluation used is stated	0;1	1	0	0	0	0	0	0	0	0
The type of economic evaluation is justified in relation to the questions addressed	0;1	1	0	0	0	0	0	0	0	0
4. Effectiveness data										
The source(s) of effectiveness estimates used are stated (effectiveness established?)	0;1	1	1	1	1	1	1	1	0	0
Details of design and results of effectiveness study are given (if based on single study)	0;1; N/A	0	0	0	1	1	0	1	N/A	N/A
Details of method of synthesis of estimates are given (if applicable)	0;1; N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5. Benefit measurement and valuation										
The primary outcome measure(s) for the economic evaluation are clearly stated	0;1; N/A	1	1	1	1	1	1	0	N/A	N/A
Methods to value health states and other benefits are stated	0;1; N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Details of the subjects from whom valuations were obtained are given	0;1; N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Checklist criteria	Possible scores	Nahra et al. (2006)	An et al. (2008)	Kouides et al. (1998)	Rosenthal et al. (2009)	Ryan (2009b)	Lee et al. (2010)	Norton (1992)	Curtin et al. (2006)	Parke (2007)
Productivity changes (if included) are reported separately	0;1; N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
The relevance of productivity changes to the study question is discussed	0;1; N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6. Costing										
Cost range is wide enough	0;1	1	1	0	0	0	0	1	1	1
Quantities of resources are reported separately from their unit costs	0;1	0	1	0	0	0	0	0	0	0
Methods for the estimation of quantities and unit costs are described	0;1	0	1	1	0	0	0	0	0	0
Currency and price data are recorded	0;1	0	0	0	0	0	1	0	0	0
Details of currency of price adjustments for inflation or currency conversion are given	0;1	0	0	0	0	0	1	0	0	0
7. Modeling										
Details of any model used are given	0;1; N/A	1	N/A	1	1	1	0	1	1	N/A
The choice of model used and the key parameters on which it is based are justified	0;1; N/A	1	N/A	1	1	1	0	1	0	N/A
8. Adjustments for timing of costs and benefits										
Time horizon of costs and benefits is stated	0;1	1	1	1	1	1	1	1	1	1
The discount rate(s) is stated	0;1; N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	0	N/A
The choice of discount rate(s) is justified	0;1; N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	0	N/A
An explanation is given if costs or benefits are not discounted	0;1; N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0	N/A
9. Allowance for uncertainty										
Details of statistical tests are given for stochastic data (e.g., p-values)	0;1; N/A	0	1	1	1	1	1	0	0	0
The approach to sensitivity analysis is given	0;1; N/A	1	0	N/A	N/A	1	0	0	0	1
The choice of variables for sensitivity analysis is justified	0;1; N/A	1	0	N/A	N/A	1	0	0	0	1
The ranges over which the variables are varied are stated	0;1; N/A	1	0	N/A	N/A	1	0	0	0	1

Checklist criteria	Possible scores	Nahra et al. (2006)	An et al. (2008)	Kouides et al. (1998)	Rosenthal et al. (2009)	Ryan (2009b)	Lee et al. (2010)	Norton (1992)	Curtin et al. (2006)	Parke (2007)
10. Presentation of results and discussion										
Relevant alternatives are compared	0;1	1	1	1	1	1	1	1	1	1
Incremental analysis is reported	0;1; N/A	0	1	1	0	0	0	0	1	1
Major outcomes are presented in a disaggregated as well as aggregated form	0;1	0	1	0	0	0	0	0	0	1
The answer to the study question is given	0;1	1	1	1	1	1	1	1	1	1
Conclusions follow from the data reported	0;1	1	1	1	1	1	1	0	1	0
Comparison made with relevant other studies (regarding costs, effects, etc.)	0;1	1	1	0	1	1	0	0	0	0
Discussion on generalizability of results	0;1	0	0	1	1	0	0	0	0	1
Conclusions are accompanied by appropriate caveats and other important factors	0;1	1	1	1	1	1	0	0	1	1
Results										
Denominator	38	34	28	27	27	30	30	30	31	26
N/A	17	4	10	11	11	8	8	8	7	12
Numerator	38	22	17	17	17	19	12	13	10	15
Total score		64,7%	60,7%	63,0%	63,0%	63,3%	40,0%	43,3%	32,3%	57,7%

CHAPTER 5

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

With Martin Emmert, Manfred Scheppach, and Oliver Schöffski

Health Policy, 2013, 10(2-3): 115-130.

ABSTRACT

Background: A vast amount of literature on effects of pay-for-performance (P4P) in health care has been published. However, the evidence has become fragmented and it has become challenging to grasp the information included in it.

Objectives: To provide a comprehensive overview of effects of P4P in a broad sense by synthesizing findings from published systematic reviews.

Methods: Systematic literature search in five electronic databases for English, Spanish, and German language literature published between January 2000 and June 2011, supplemented by reference tracking and Internet searches. Two authors independently reviewed all titles, assessed articles' eligibility for inclusion, determined a methodological quality score for each included article, and extracted relevant data.

Results: Twenty-two reviews contain evidence on a wide variety of effects. Findings suggest that P4P can potentially be (cost-)effective, but the evidence is not convincing; many studies failed to find an effect and there are still few studies that convincingly disentangled the P4P effect from the effect of other improvement initiatives. Inequalities among socioeconomic groups have been attenuated, but other inequalities have largely persisted. There is some evidence of unintended consequences, including spillover effects on unincentivized care. Several design features appear important in reaching desired effects.

Conclusion: Although data are available on a wide variety of effects, strong conclusions cannot be drawn due to a limited number of studies with strong designs. In addition, relevant evidence on particular effects may have been missed because no review has explicitly focused on these effects. More research is necessary on the relative merits of P4P and other types of incentives, as well as on the long-term impact on patient health outcomes and costs.

5.1 INTRODUCTION

In many countries, healthcare delivery is suboptimal. For example, adherence to professional medical guidelines is often low (Grol 2001; McGlynn et al., 2003; Steel et al., 2007), while costs of care continue to rise (OECD, 2012). Pay-for-performance (P4P) has become a popular approach to increase efficiency in health care. In addition to the United States where P4P has become widespread, P4P-programs are being implemented in many other countries, including in the United Kingdom, Australia, Canada, Taiwan, Israel, and Germany (see chapter 3). In P4P, healthcare providers receive explicit financial incentives based on their scores on predefined performance measures that may pertain to clinical quality, resource use, and patient-reported outcomes. Along with the dissemination of P4P, the literature on the effects of P4P has expanded rapidly over the past fifteen years. Although this is a desirable development, the evidence has become fragmented. Several systematic reviews have synthesized available evidence, but they all had different foci (e.g., only including experimental studies, only focusing on preventive services, not addressing other potential P4P effects besides impact on incentivized performance, etc.) and hence different conclusions. Consequently, it is still challenging to comprehend this evidence and to extract success factors and pitfalls when it comes to implementing P4P.

In this paper, we summarize the existing literature on P4P effects in a broad sense by conducting a systematic review of published systematic reviews. The paper adds to the literature by synthesizing key findings from these reviews. The goal is to provide a structured, comprehensive overview of the evidence on P4P effects and mediating factors. We achieve this by addressing the following six questions: to what extent has P4P been (1) effective and (2) cost-effective? (3) Which unintended consequences of P4P have been observed? To what extent has P4P (4) affected inequalities in the quality of care and (5) been more successful when combined with non-financial incentives? (6) Which specific design features contribute to (un)desired effects? To our knowledge, no prior study has provided such an overview. The results will be of interest for policymakers intending to implement a P4P-program as well as those who have already done so. The next section provides a theoretical background on the relevance of the six questions. Next, after describing the search strategy and inclusion and exclusion criteria, the results are presented separately for each question. In the discussion, the results are compared with findings from recent studies not included in any of the identified reviews (if available and relevant). We end with discussing the implications of our findings for research and policy.

5.2 THEORETICAL BACKGROUND

Both economic theory and common sense support the notion that payment for health care should be determined, at least in part, based on meaningful indicators of quality or value (Rosenthal, 2007b). Given notable deficiencies in the quality and efficiency of care, that healthcare providers (be they individual physicians, physician groups, or institutions) are responsive to financial incentives and that improving performance requires changes in their behavior, that many common payment methods (e.g., fee-for-service, capitation) do not explicitly stimulate good performance, and that performance measurements have become more sophisticated and accurate, it seems natural to tie a portion of providers' compensation to their performance. However, although the idea underlying P4P is simple, in practice there are many potential pitfalls.

P4P can be considered *cost-effective* when improved quality is achieved with equal or lower costs or when the same quality is achieved with lower costs. But even if P4P leads to cost increases it may still be viewed as cost-effective, as long as quality improvements are large enough. However, designing and implementing a successful P4P-program is highly complex (see chapter 2). Engaging providers, reaching consensus about program design, collecting and validating data, calculating payments, and maintaining and evaluating the program likely involve high transaction costs. This raises the question whether P4P can be cost-effective.

In theory, P4P may have several *unintended consequences*. First, when casemix differences among providers are not taken into account, providers have an incentive to select healthy/compliant patients and to avoid severely ill/noncompliant patients, especially for clinical outcome and resource use measures. Even sophisticated risk-adjustment models may fail in preventing selection because providers are likely to have superior information about their patients than included in these models (Dranove et al., 2003). Other strategies, such as allowing providers to exclude noncompliant patients from performance calculations (Doran et al., 2008a), may be necessary. Second, P4P may cause providers to focus disproportionately on aspects of care that are incentivized and neglect other important aspects that are not (Holmstrom & Milgrom, 1991). A broad set of measures (e.g., clinical quality, patient satisfaction, continuity of care, resource use) seems therefore important. However, this is often not feasible in practice. Third, P4P may "crowd out" providers' intrinsic motivation to provide high-quality care, especially when the definition of performance is not shared. P4P could then play a trivializing role regarding the non-financial motivation (Christianson et al., 2008), which may have several undesired effects. Finally, to maximize income, providers may manipulate data such that their performance seems better than it is in reality ("gaming").

P4P may narrow, widen, or maintain *inequalities* regarding access to and/or receipt of high-quality care (Chien et al., 2007). Inequalities may widen if P4P encourages risk selec-

tion or results in reduced income for providers serving minority populations (Alshamsan et al., 2010). Providers in deprived areas will typically have lower performance and be less likely to receive incentive payments than providers practicing in affluent areas, for example because their patients are less likely to adhere to recommended treatment (Casalino et al., 2007). By adversely affecting the income of providers practicing in deprived areas, P4P may reduce both the number of providers working in such areas and their ability to invest in performance improvement. Widening inequalities can be prevented by rewarding improvement in performance, adequate risk adjustment, inclusion of measures that are more important for minority patients, and/or directly rewarding reductions in inequalities (Casalino et al., 2007; Alshamsan et al., 2010; Blustein et al., 2011).

Non-financial incentives such as public reporting (PR) of (differences in) performance scores and timely performance feedback to providers may complement P4P. Both PR and P4P reward providers for good performance, but the financial incentive in PR operates indirectly via consumer choice (Chien et al., 2007). Performance feedback and reminders make treatment patterns and performance issues salient and can activate providers to adjust their practice style. Feedback may also create a reputational incentive if reports also include information on the performance of peers.

The *design of P4P* has important consequences for the incentives that providers experience and how they might respond to them (Mehrotra et al., 2010a). Seemingly important design elements are the number and type of included performance measures, risk adjustment, the entity targeted (individual physicians, physician groups, institutions), the type (rewards, penalties) and size of the incentive, the frequency of payment, and the type (absolute, relative, improvement) and number of performance targets (Conrad & Perry, 2009; see also chapter 2). In summarizing the literature, we attempt to infer about preferred design in practice by identifying patterns in the results.

5.3 METHODS

5.3.1 Search and selection procedure

For this review, we adhered to guidelines from the Cochrane Collaboration (Higgins & Green, 2008), the Institute for Quality and Efficiency in Health Care (2008), the Hannoveraner Konsensus (Graf von der Schulenburg et al., 2007), and the NHS Economic Evaluation Database (Craig & Rice, 2009). We searched five databases: Medline (through Pubmed), Embase, ISI web of knowledge, the Cochrane Database of Systematic Reviews, and Scopus (see Appendix 5.1 for the full search history). We also searched the Internet via Google, contacted experts, and reviewed reference lists of retrieved articles. Articles written in English, Spanish, or German published between January 2000 and June 2011 were eligible for inclusion. Two authors independently reviewed all titles generated by the procedure and

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

5.3.2 Methodological quality assessment

To determine the methodological quality of included reviews, we applied the checklist of the German Scientific Working Group, which contains eighteen distinct criteria (Dreier et al., 2010). The items are grouped under five categories: research question, search procedure, evaluation of information, synthesis of information, and presentation of results (see Appendix 5.2 for details). A total score is obtained by adding up awarded points and dividing by the number of points that could maximally be earned. Two authors independently carried out the scoring.

5.3.3 Data extraction and synthesis

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

The findings from included reviews are also compared with findings from several recently published primary studies that are not included in any of the reviews but that do provide relevant information. These studies were not identified from an additional system-

atic review, but from our knowledge of the current evidence base on P4P effects. Although there may be more studies than the ones we discuss, comparing our results with findings from additional studies we are aware of provides additional insight in the effects of P4P and enables us to draw stronger conclusions.

5.4 RESULTS

The initial search identified 2,004 articles (Figure 5.1). After review of titles/abstracts, 487 studies remained for detailed reflection. Reference tracking, Internet searches, and expert consultation yielded 28 additional articles. Of the 515 articles subjected to full-text review, 493 articles met at least one exclusion criterion, leaving 22 articles for inclusion. Table 5.1 presents their main features. The reviews vary considerably by inclusion criteria and focus. For example, some reviews focus only on one (subset of) condition(s) or on one specific sector. Others only include studies with a particular design (e.g., RCTs) while still others had no restrictions at all. The result is a wide range in the number of included studies across the reviews. While most reviews only included studies from the US and the UK, studies conducted in other countries have increasingly been identified (ten in total). Most studies were conducted in primary care, although an increasing number of studies have evaluated P4P in other sectors (e.g., inpatient care). Evidence mainly comes from observational studies and many authors have therefore noted that results must be interpreted with caution. Table 5.2 and the following sections present the key findings for each of the six domains.

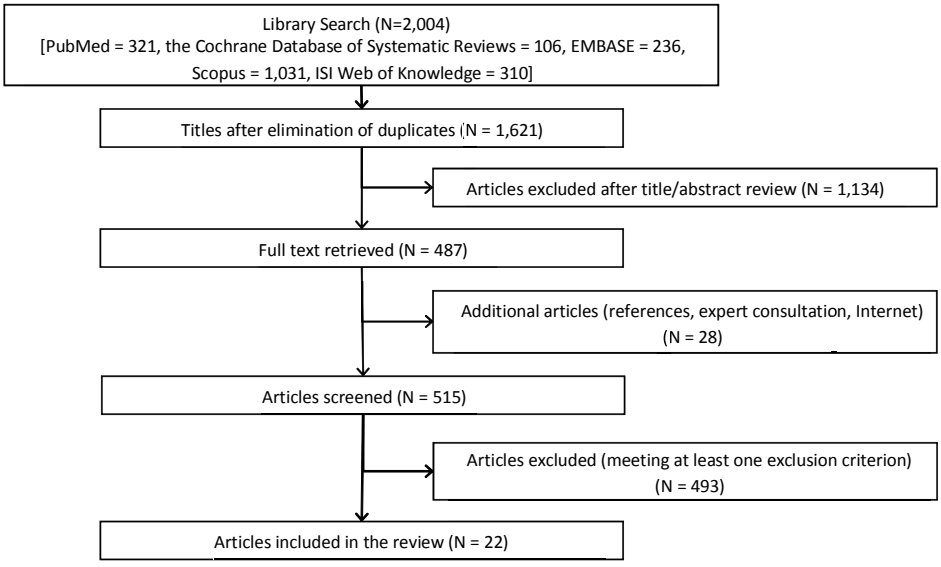


FIGURE 5.1 Search flow and results

TABLE 5.1 General features of included systematic reviews of the literature on effects of P4P

Reference	Score [a]	Search period	Studies	Type of studies [b]	Studies per level	Countries [c]	Sector [d]	Evidence on [e]
Alshamsan et al. (2010)	93%	1980-November 2008	22	UBA (5) UA/CS (17)	Level III: 5 Level IV: 17	UK (21) US (1)	Mostly PC (QOF)	I DF
Armour & Pitts (2003)	73%	1966-December 2001	6	RCT (2) TS (1) UA/CS (3)	Level I: 2 Level II: 1 Level IV: 3	US	PC (5) H (1)	E CE UC DF
Briesacher et al. (2009)	87%	1980-August 2007	4	RCT (1) UA/CS (3)	Level I: 1 Level IV: 3	US	NH	E CE
Chaix-Couturier et al. (2000)	87%	1993-May 1999	2	RCT	Level 1: 2	US	PC	E NFI DF
Christianson et al. (2007)	87%	1988-June 2007	44	R (7) RCT (7) QE (4) CBA (6) TS (2) UBA (4) CA (1) UA/CS (11) Q (2)	Level I: 14 Level II: 12 Level III: 5 Level IV: 11 Others: 2	US (27) UK (7) SP (1) AU (1) TW (1) NA (7)	H (6) NH (1) PC (36) PC+SC (1)	E CE
Christianson et al. (2008)	87%	-August 2007	27	RCT (2) QE (4) CBA (4) TS (1) UBA (4) CA (1) UA/CS (10) Q (1)	Level I: 2 Level II: 9 Level III: 5 Level IV: 10 Others: 1	US (18) UK (7) AU (1) SP (1)	H (6) PC (20) PC/SC (1)	E CE UC I DF
Dudley et al. (2004)	93%	1980-2003	8	RCT	Level 1: 8	US	PC (8) PH (1)	E DF
Eldridge and Palmer (2009)	60%	1990-2008	27	QE (1) CA (1) UA/CS (25)	Level II: 1 Level III: 1 Level IV: 25	8 developing countries	Not reported	E
Emmert et al. (2012)	93%	2000-April 2010	9	RCT (3) CBA (3) UBA (3)	Level I: 3 Level II: 3 Level III: 3	US (8) TW (1)	H (5) PC (4) NH (1)	CE NFI DF
Frølich et al. (2007)	93%	1980-June 2005	8	RCT	Level I: 8	US	Not reported	E DF
Giuffrida et al. (2000)	100%	1966-October 1997	2	RCT TS	Level I: 1 Level II: 1	US UK	PC	E CE
Kane et al. (2004)	93%	1966-October 2002	9	RCT (6) TS (1) UBA (2)	Level I: 6 Level II: 1 Level III: 1	US (8) UK (1)	Prevention	E CE DF
Mehrotra et al. (2009)	87%	1996-June 2007	8	QE (2) CA (1) UA/CS (4) Q (1)	Level II: 2 Level III: 1 Level IV: 4 Others: 1	US	H	E CE UC NFI

TABLE 5.1 (continued)

Reference	Score [a]	Search period	Studies	Type of studies [b]	Studies per level	Countries [c]	Sector [d]	Evidence on [e]
Petersen et al. (2006)	100%	1980-November 2005	17	RCT (9) CBA (4) UA/CS (4)	Level I: 9 Level II: 4 Level IV: 4	Mainly US	Mainly PC	E CE UC
Rosenthal and Frank (2006)	73%	-Fall 2003	6	RCT (4) QE (1) UBA (1)	Level I: 4 Level II: 1 Level III: 1	US	PC	E UC NFI
Sabatino et al. (2008)	80%	-September 2004	3	RCT QE UBA	Level I: 1 Level II: 1 Level III: 1	US	Prevention (cancer)	E
Schatz (2008)	67%	2006-2007	22	RCT (7) CBA (6) UBA (7) UA/CS (2)	Level I: 7 Level II: 6 Level III: 7 Level IV: 2	US (19) UK (3)	Ambulatory care	E UC NFI DF
Scott et al. (2011)	93%	2000-August 2009	6	RCT (3) QE (1) TS (2)	Level I: 3 Level II: 3	US (5) GER (1)	PC	E UC
Sorbero et al. (2006)	73%	1995-April 2006	15	RCT (7) QE (2) UBA (6)	Level I: 7 Level II: 2 Level III: 6	US	PC (physicians)	E NFI DF
Steel and Willems (2010)	78%	-January 2010	34	TS (4) UBA (8) CS/UA (17) M (1) Q (4)	Level II: 4 Level III: 8 Level IV: 17 Others: 5	UK	PC (QOF)	E CE UC I
Town et al. (2005)	73%	1966-2002	6	RCT	Level I: 6	US	PC (prevention)	E CE NFI DF
van Herck et al. (2010)	100%	1990-July 2009	128	RCT (10) QE (4) CBA (17) TS (6) UBA (30) UA/CS (57) M (4)	Level I: 10 Level II: 27 Level III: 30 Level IV: 57 Others: 4	US(63) UK(57) IT(1) SP(2) AG(1) AU(2) GM(2)	PC(98) H(17) H/PC(13)	E CE UC I NFI DF

a. Total methodological quality score. See Appendix 5.2 for the reviews' scores on individual items.

b. R=Review, RCT=Randomized Controlled Trial, QE=Quasi-Experiment, CBA=Controlled Before-After study, TS=Time Series with before-after data, UBA=Uncontrolled Before-After study, CA=Controlled-After study, UA/CS=Uncontrolled-After study / Cross-Sectional survey, M=modeling study, Q=Qualitative study.

c. AG=Argentina, AU=Australia, GM=Germany, IT=Italy, NA=Not Applicable, SP=Spain, TW=Taiwan, UK=United Kingdom, US=United States.

d. H=Hospital, HP=Health Plan, IC=Intensive Care, MG=Medical Group, NH=Nursing Home, PC=Primary Care, PH=Pharmacy, QOF=Quality and Outcomes Framework.

e. E=Effectiveness, CE=Cost-Effectiveness, UC=Unintended Consequences, I=Inequalities, NFI=Non-Financial Incentives, DF=Design Features.

5.4.1 To what extent has P4P been effective?

Twenty reviews provide evidence on the effectiveness of P4P. We present the results according to the design of included studies: randomized controlled trials (level I) and non-randomized studies (levels II-IV). *Randomized controlled trials* have largely investigated the impact of P4P on preventive care services such as cancer screening and immunizations. Most reviews rely on the same core set of relatively dated studies conducted in US primary care settings. Dudley et al. (2004) found that among ten dependent variables studied in eight RCTs, six showed a significant relationship with the incentive. For example, one RCT found no difference between intervention and control groups in cancer screening rates after eighteen months, while another found that relatively small payments improved immunization rates by 4 percentage points. Overall, the effect size among the positive studies was moderate at best. Town et al. (2005), focusing on prevention, classified only one of eight outcomes as improved. They classified two studies that found that improved immunization rates were largely due to better documentation as ineffective, whereas Dudley et al. classified them as effective. Nonetheless, all authors (including also Rosenthal & Frank, 2006 and Schatz, 2008) reached the same conclusion, namely that results are mixed and inconclusive and that there is insufficient evidence to support the use of P4P to improve the quality of preventive and chronic care in primary care. Another review, focusing exclusively on nursing home care, identified an RCT (published in 1992) that found small beneficial effects on access and quality (Briesacher et al., 2009).

Most *non-randomized studies* showed improvement in selected quality measures. P4P appears to have had a small positive impact on the quality of care for diabetes and asthma, but not for heart disease (Sorbero et al., 2006; Christianson et al., 2008). Schatz reached a similar conclusion (Schatz, 2008): among fifteen studies (five level II, seven level III, two level IV), ten found positive and four found mixed results. More positive results were found among level III/IV studies than among level II studies. The most comprehensive review was conducted by van Herck et al. (2010), who identified 111 studies. Of these, 30 studies reported an effect size, which ranged from negative to absent to (very) positive. The three studies finding negative effects also found positive results on other measures. Overall, P4P seems to have led to 5 percent improvement in performance, although there is much variation (van Herck et al., 2010). For example, better results have been achieved for immunizations than for cancer screening (Sabatino et al., 2008).

One review focused exclusively on the impact of the Quality and Outcomes Framework (QOF) in the UK (Steel & Willems, 2010), a large national P4P-program that pays bonuses to primary care practices of up to 30 percent of their revenues for reaching targets for about 130 measures. Overall, results from 28 studies (four level II, eight level III, fifteen level IV, one modeling) show that achievement was high in the first year (2004-5) and has increased since. Large improvements were demonstrated in the period 2005-8 especially for diabetes, but also for hypertension, heart disease, and stroke. However, the trend typically showed a

gradual improvement with little change after the QOF was implemented. For diabetes and asthma, a small but significant above-trend increase was found. Another study (level II) found both slightly lower and slightly higher achievement than predicted by the underlying trend. In addition, most studies (all level IV) found no relationship between target achievement and clinical outcomes such as hospital admissions and mortality.

Several reviews discuss studies that assessed the impact of P4P in hospitals (Christianson et al., 2008; Mehrotra et al., 2009; van Herck et al., 2010). Van Herck et al. found that compared to primary care, P4P has more often failed to improve acute inpatient care. Mehrotra et al. provide a detailed analysis of the effects of hospital P4P-programs in the US. The most rigorous evidence (two level II, one level III) comes from a single program, the Hospital Quality Incentive Demonstration (HQID). This program, which ran from 2003 to 2009, incentivized 266 hospitals to perform well on 33 clinical measures (largely processes) pertaining to six conditions. Overall, a 2 to 4 percentage point improvement was found beyond the improvement seen in control hospitals. No impact was found on mortality, despite the fact that for some conditions reductions in 30-day risk-adjusted mortality was explicitly incentivized. Finally, three level IV studies in the US nursing home sector showed small (e.g., improved patient satisfaction) or no effects (Briesacher et al., 2009).

5.4.2 To what extent has P4P been cost-effective?

Twelve reviews provide evidence on P4P cost-effectiveness, although only six explicitly focused on it (Table 5.2). Emmert et al. (2012) made a distinction between full and partial economic evaluations (see chapter 4). Full evaluations consider (program) costs and quality and explicitly link them to each other (e.g., by calculating cost-effectiveness ratio's). Partial evaluations may allow for inferences about cost-effectiveness if the impact is described on both costs and quality. However, results have lower significance than those of full evaluations because the connection between both effects is less clear. Partial evaluations also include simple costs comparisons without analysis of the impact on quality. Emmert et al. (2012) identified three full evaluations (two level I, one level III), which all found improvements in quality against increases in costs. For example, one study calculated a cost per QALY of $12,967 to $30,081 for inpatient heart treatment, while another found an intervention cost of $3 per additional immunization. Van Herck et al. (2010) identified an additional full evaluation (level I) demonstrating cost-effectiveness of a P4P-program for smoking cessation in Germany, but only when combined with training for GPs and free medication for patients.

Regarding partial evaluations, two studies (level I and level II) found quality improvements and cost increases. The level I study, evaluating a program for nursing homes designed to improve access and patient outcomes, found that the program saved $3,000 per stay, but average costs to Medicaid rose by 5 percent, in part due to program costs. Another study (level II) found both cost savings and improved quality, while still another level II study likely demonstrated P4P inefficiency in reducing 30-day mortality for four

conditions in US hospitals. Two cost comparisons (both from the US) showed a positive financial impact. Other reviews (Kane et al., 2004; Petersen et al., 2006; Christianson et al., 2008; Briesacher et al., 2009) discuss studies that were also identified by Emmert et al. and reached similar conclusions. Steel and Willems (2010) found an additional study providing evidence of cost-effectiveness for twelve measures included in the QOF. Although this highlights the potential of P4P to be cost-effective, no economic evaluation of the QOF itself was conducted.

Based on these results, most authors concluded that P4P can potentially be cost-effective, but that convincing evidence is lacking. Although van Herck et al. (2010:8) conclude that "cost-effectiveness (…) is confirmed by the few studies available", the evidence seems not sufficient to draw this conclusion, also because studies typically suffer from methodological limitations (e.g., lack of control group or trend data) and failed to include an appropriate cost and/or effect range.

5.4.3 Which unintended consequences has P4P had?

Nine reviews provide evidence on unintended consequences, including risk selection, spillover effects, gaming behavior, and effects on providers' intrinsic motivation. Three reviews provide weak evidence from three studies that P4P could lead to *risk selection* (Table 5.2). However, two studies (level I and level IV) were conducted in the context of PR (Rosenthal & Frank, 2006). The third study (level II) investigated a performance-based contracting system for providers of substance abuse treatment and found that the likelihood of a patient in the program being in the most severely ill group increased in the control group and decreased in the intervention group.

Spillover effects have been discussed in six reviews (nine studies in total). The findings provide a mixed picture. Four reviews (Christianson et al., 2008; Schatz, 2008; Mehrotra et al., 2009; Scott et al., 2011) discuss results of three evaluations of large P4P-programs for GPs and hospitals. Two studies (level II, US HQID; level II, UK QOF) found no differences in trends in unincentivized and incentivized measures; the third study (level II, US primary care) found no change in unincentivized performance while some incentivized measures improved. Another study from the QOF (level III) found that unincentivized measures improved when they were part of a condition for which there were incentives for other measures (Steel & Willems, 2010). However, performance for two unincentivized conditions was not significantly improved despite that achievement was much lower than for incentivized conditions (which did improve). In addition, qualitative studies found that providers are often concerned about less time for holistic care, deterioration of unincentivized care, and reductions in continuity of care. Finally, van Herck et al. (2010) discuss two additional studies (both level II) from the QOF. The first showed a positive effect on unrewarded aspects of an included condition, a deterioration of unrewarded aspects of two other included conditions, and a reduction in continuity of care. The second study, focusing

on four chronic conditions, found that the effect on recording of incentivized risk factors by GPs was larger for targeted patient groups (i.e., patients with an included condition) than for untargeted groups. It also found evidence of sizable positive spillover effects (an increase of 10.9 percentage points) on unincentivized factors for the targeted groups.

Four reviews discuss findings related to *gaming behavior* (Table 5.2). Most of these include a study (level I) that found that US nursing homes tended to claim they were admitting extremely disabled patients, who then "miraculously" recovered (Petersen et al., 2006). One review discusses "exception reporting" in the QOF (van Herck et al., 2010), which allows primary care practices to exclude (noncompliant) patients from performance calculations but also provides opportunities to increase income by excluding patients for inappropriate reasons. One study (level IV) found low rates of exception reporting in the first year, but it was the strongest predictor of performance; a small number of practices may have achieved high scores by excluding large numbers of patients. A follow-up study (level IV) again found little evidence of widespread gaming; there seemed to be good clinical reasons for the exception reporting rates, which were still low in the second year.

Regarding effects on providers' *intrinsic motivation* and perceived professionalism, results of five qualitative studies are summarized in two reviews (Christianson et al., 2008; Steel & Willems, 2010). Two studies found that P4P did not impair providers' intrinsic motivation and that it had no effect on the quality of professional life, although providers did express more support for targets aligned with professional priorities. However, three other UK studies suggest P4P may result in a loss in autonomy and that it may undermine providers' sense of professionalism. Providers also reported concerns about "a dual agenda in consultations, with less time for holistic care, patients' concerns and non-incentivized care, and a perceived loss in continuity of care" (Steel & Willems, 2010: 120).

5.4.4 To what extent has P4P affected inequalities?

Four reviews provide information on the impact on inequalities (Table 5.2). Most studies addressed the impact on *socioeconomic inequalities*. Alshamsan et al. (2010) identified eighteen studies, most of which examined cross-sectional associations (level IV) between the quality of chronic care and an "area deprivation score" after QOF implementation. Most studies found lower quality in deprived areas compared to affluent areas before or shortly after the QOF, but differences were typically small and appear to have narrowed over time. One study (level IV) investigated a long-term effect of a small P4P-program in the UK in the early 1990s and demonstrated that the initial widening of inequalities in cervical cancer screening rates had almost disappeared after five years. However, two level III studies found that after the QOF, medical records of patients living in affluent areas were more likely to include important risk factors (e.g., smoking status) than those of patients living in deprived areas, a difference that was not evident before. Steel and Willems (2010) found indications of narrowing inequalities between the most and least deprived areas in England, but also

TABLE 5.2 Key findings of included systematic reviews of the literature on effects of P4P

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Alshamsan et al. (2010)				Some evidence that P4P reduced socioeconomic inequalities, but inequalities regarding age, sex and ethnicity persisted. Evidence on long-term effects weak.		Some evidence that low achievement in year t-1 leads to high achievement in year t. Therefore, using measures with low baseline performance and/or adopting a tiered series of targets may yield the largest benefits.
Armour & Pitts (2003)	1 study: financial risk for referrals decreased primary care visits; risk for cost of outpatient tests reduced number of tests; bonuses / withholds for productivity did not change resource use. 1 study: large reductions in number of admissions and visits. Regarding preventive care: increases in immunization rates (1 study), but not in cancer screening rates (1 study).	2 studies: bonuses / withholds for reduced resource use may lead to reduced outpatient expenses and utilization. 1 study: reduced outpatient expenses by 5% as a result of bonuses / withholds.	1 study: physicians at risk for the cost of outpatient tests substituted primary care visits for outpatient tests, which increased the number of visits per enrolee per year by 5%.		1 study: no effect of P4P with semi-annual feedback.	Regarding quality: studies with absolute targets effective, study with relative targets ineffective. Regarding resource use: 1 study found a greater reduction in utilization when directed at individuals whose contracts included a withhold than when directed at groups. 1 study: lack of association between bonuses / withholds and change in resource use may have been result of delayed rewards. 1 study: lack of effect possibly a result of limited physician awareness and narrow time frame of study.
Briesacher et al. (2009)	1 RCT: improved access and clinical outcomes. Modest or no effect found in 3 observational studies.	1 RCT: improved access and outcomes quality against a 5% increase in cost.				
Chaix-Couturier et al. (2000)	1 study: improved immunization rates. 1 study: P4P + feedback did not affect cancer screening rates.				1 study: P4P with feedback had no effect on cancer screening rates.	Study using 2 absolute targets effective in improving immunization rates; study using 2 relative targets ineffective in improving screening rates.

TABLE 5.2 (continued)

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Christianson et al. (2007)	Evidence base for justifying and designing P4P is thin. Few significant impacts reported, and only in selected measures.	1 study: positive ROI of P4P. 1 study: cost per QALY between $13,000-30,000.				
Christianson et al. (2008)	Most studies found improvement in selected quality measures, but the direct effect of P4P is largely unclear due to lack of control groups and concurrent other improvement efforts.	See Christianson et al. 2007	2 studies: no evidence of teaching to the test. 1 study: P4P did not impair GPs' intrinsic motivation. Initial improvements may reflect better documentation.	1 study: better record keeping for oldest patients and patients in most affluent areas; improvement for women larger than for men. Still lower recording for women, older patients, and patients in deprived areas.		2 studies: active provider engagement contributes to better results. 2 studies: lack of impact may have been a result of lack of awareness among providers. 2 programs using explicit performance targets: most money awarded to providers with high baseline quality.
Dudley et al. (2004)	Results are mixed and inconclusive. Among 7 studies focusing on physicians, 5 of 9 variables showed a significant relationship with the incentive while 4 did not.					Individuals: 5 positive and 2 null results; groups: 1 positive and 2 null results. No relation between pay size and response. 2 studies with relative targets: no effect. Among 5 studies with enhanced FFS, 4 were positive and 1 insignificant, while in the 4 studies using bonuses there were only 2 positive results.
Eldridge and Palmer (2009)	Lack of evidence on the effects of any type of P4P in any low-income country setting, mostly due to the absence of control groups.					

TABLE 5.2 (continued)

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Emmert et al. (2012)	Among 7 studies considering both costs and effects, 5 showed improved quality of care can be achieved with higher costs.	P4P can potentially be cost-effective, but results are not convincing.			Feedback and/or PR additional to P4P did not lead to better or worse results.	Weak evidence that larger payments increase (cost-)effectiveness. The 3 studies with a high payment frequency were all relatively successful.
Frølich et al. (2007)	8 RCTs: mixed results. The potential to improve quality through P4P remains unknown.					See Dudley et al.
Giuffrida et al. (2000)	Target payments associated with higher immunization rates, but increase significant in only 1 study.	1 study: additional cost of $3 per extra immunization.				
Kane et al. (2004)	4 studies had positive findings while 5 studies found no effect. Effect size is moderate at best.	1 study: additional cost of $3 per extra immunization.				Effects larger for groups. No dose-response relationship. 2 studies involving relative targets and low awareness found no effects.
Mehrotra et al. (2009)	Of the 8 studies, most lack a control group and the best evidence comes from a single program. Evaluation of this program (3 studies) found a 2 to 4 percentage point improvement beyond the improvement seen in control hospitals.	1 study found an estimated cost per QALY of $12,967-$30,081, a range generally considered cost-effective. Yet this study lacked a control group and trend data.	Difference between intervention and control group on excluded measures was not significant, except for 1 measure (intervention group improved more). No data on spillover effects on other unrewarded aspects.		3 studies: PR may have contributed to observed improvements.	

TABLE 5.2 (continued)

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Petersen et al. (2006)	5 of 6 physician-level and 7 of 9 group-level incentives found partial (5) or positive effects (2). 2 RCTs with group-level incentives found no effect. 1 of the 2 'payment-system level' studies found a positive effect on access.	1 study: combination of incentives to improve access to nursing home care and outcomes saved $3,000 per stay.	4 studies: evidence of unintended effects, including selection and improvements in documentation rather than quality.			5 of 6 physician-level and 7 of 9 group-level incentives found partial (5) or positive effects (2). 2 RCTs with group-level incentives found no effect.
Rosenthal and Frank (2006)	The empirical foundations of P4P are weak. Most studies found no effect with 2 positive findings.		Although not found in the context of P4P, several studies suggest unintended effects are possible, including gaming and selection.		1 study: no effect of feedback only and of P4P + feedback. 1 study: feedback no effect, P4P may lead to better record-keeping. 1 study: P4P improved processes, but no effect of P4P + access to a patient registry + counseling.	
Sabatino et al. (2008)	1 positive result (but no control group) and 2 null results; insufficient evidence.					
Schatz (2008)	RCTs: 3 null, 3 positive (2 better documentation), 1 mixed. Non-randomized studies: 10 positive, 4 mixed, 1 null. Often unclear if effects are due to P4P.		Possible positive spillover effect found in 1 study.		Weak evidence that non-financial incentives contribute to P4P success.	Positive studies typically used larger bonuses and measures more amenable to change.

TABLE 5.2 (continued)

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Scott et al. (2011)	Modest effects for typically only 1 out of several measures. Risk of bias due to methodological limitations.		One study: no evidence of positive/negative spillovers on unincentivized aspects.			
Sorbero et al. (2006)	4 RCTs had mixed results, while 3 reported no effect. 2 quasi-experiments found mixed results, while observational studies tend to report positive results for at least one performance aspect.				Monitoring can boost P4P effect. P4P must be implemented as part of a multi-faceted strategy to performance improvement.	Lack of effects may be due to small payments; weak evidence that 5% of revenues is required. Low awareness contributed to limited effect. Physician engagement, pilot testing, accurate and reliable data, ongoing evaluation, and physician support were reported critical.
Steel and Willems (2010)	Overall, achievement increased since QOF, but post-QOF performance was roughly in line with the trend predicted from pre-QOF years. For some measures there is evidence for performance slightly above the predicted trend.	Evidence of cost-effectiveness for 12 measures with direct therapeutic effect.	No evidence that excluded conditions were neglected more after QOF than before. No signs of reduced intrinsic motivation, but reportedly less attention to patients' concerns, unrewarded care, and continuity of care.	Changes in inequalities were small, variable, and depended on the measure, achievement before QOF, and demographic variable. Differences among age groups attenuated for some conditions; no changes in sex-related inequalities; reduced differences between most and least deprived areas on national level but not necessarily on local levels; mixed findings for ethnicity, with some reductions for some but not all measures.		

TABLE 5.2 (continued)

Reference	Effectiveness	Cost-effectiveness	Unintended consequences	Inequalities	Non-financial incentives	Design features
Town et al. 2005	1 of 8 outcomes showed a significant effect. 1 significant difference found for feedback + bonus compared to control group. 1 study: P4P resulted in improved documentation.	1 study: $3 per extra immunization, which was deemed cost-effective as flu vaccines have been shown to save $117 in direct medical expenses in elderly.			1 study: feedback alone group was not different from control group. No difference between feedback + bonus vs. feedback only.	Neither type of payment nor type of preventive service drives lack of effect. Limited effectiveness may be due to small rewards. Complex rules for rewards are less effective.
van Herck et al. (2010)	5% improvement overall, but much variation. Negative results found in 3 studies, together with positive results on other measures. Positive effects mainly for immunizations, diabetes, asthma, and smoking cessation. P4P most often failed to improve acute care.	4 studies on cost-effectiveness: all positive, though interpretation is difficult.	Mixed evidence of teaching to the test and gaming; very few studies have addressed such effects.	No negative effect on age, ethnic, and socioeconomic inequalities. Evidence from 28 studies suggests reductions in inequalities in the quality of care across groups rather than increases.	3 non-US studies: positive results when P4P is part of a larger quality improvement strategy such as PR. Evidence from US (N=28) more mixed.	More improvement for process measures than for outcomes; larger effects for measures with more room for improvement; involvement of providers, exception reporting, risk adjustment, and extensive communication appear to contribute to positive effects; provider awareness important; relative targets often less effective than absolute ones; no dose-response relation; programs using new money had more positive effects than programs using existing funds; targeting individual physicians or small teams were often more effective than targeting large groups or hospitals.

Note: FFS= Fee for Service, GP=General practitioner, P4P=Pay-for-performance, PR=Public reporting, QALY=Quality-adjusted life year, QOF=Quality and Outcomes Framework, RCT=Randomized controlled trial, ROI=Return on investment, UK=United Kingdom, US=United States.

showed that large differences remained in individual measures and that the worst performing practices remain concentrated in the most deprived areas. Summarizing results from 28 studies (mainly from the UK QOF), van Herck et al. (2010) conclude that the evidence points to a reduction in inequalities across socioeconomic groups rather than an increase.

Alshamsan et al. (2010) identified nine studies (five level III, four level IV) investigating the impact of the QOF on *inequalities related to age, sex, and ethnicity* for stroke, heart disease, and diabetes. Although P4P does not appear to have widened inequalities, existing inequalities persisted; women, older patients, and patients from minority ethnic groups continued to receive lower quality of care after QOF implementation than men, younger patients, and the white British group, although some gaps attenuated. Steel and Willems (2010) had similar findings. For example, both before and after the QOF higher achievement was found for men for nearly all heart disease measures and three of eight diabetes measures. An additional study from Scotland found worse recording of risk factors for women, older patients, and patients in more deprived areas (Christianson et al., 2008).

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

Five level I studies (all conducted in US primary care settings) provide information on the merits of combining P4P with *performance feedback* to providers. One RCT found no effect of combining P4P with feedback on cancer screening rates (Chaix-Couturier et al., 2000; Armour & Pitts, 2003). Another RCT showed that neither "feedback alone" nor "feedback and P4P" improved childhood immunization rates (Rosenthal & Frank, 2006). Town et al. (2005) discuss the results of three additional RCTs. In one study, results from the "P4P and feedback" group were significantly different from those from the control group, but not from results from the "feedback only" group. In the second study, screening rates among the "feedback only" group did not differ significantly from those of the group receiving feedback and a $50 bonus. The third study also could not demonstrate superiority of "P4P and feedback" over "feedback only". Schatz (2008) found some weak evidence that feedback contributes to P4P success, and Sorbero et al. (2006) showed that performance monitoring can have the overall effect of improving performance. Finally, van Herck et al. (2010) found that P4P appears to have had a larger effect when part of an overall quality improvement strategy that also includes structured feedback and PR, although the evidence is not very convincing as studies typically lacked a control group.

Some reviews found evidence that *public reporting* can be more effective when used together with P4P. One level II study found that US hospitals subjected to PR and P4P improved between 2.6 percent and 4.1 percent more in process quality for certain inpatient diagnoses than hospitals subjected only to PR (Mehrotra et al., 2009). Mehrotra et al. also identified two other studies (level II and level III) assessing the impact of the HQID, which combined P4P with PR. Studies indicated a 2 to 4 percentage point improvement beyond the improvement seen in control hospitals. Although the effects of P4P and PR could not be

disentangled, the authors suspect that PR contributed to these findings, perhaps even more than P4P.

5.4.6 Which specific design features have contributed to desired effects?

Several design features seem important in reaching desired effects, although no study has directly investigated their effect. These features relate to the type of measures, targeted entity, type and number of targets, type and size of the incentive, payment frequency, and provider engagement. Regarding *performance measures*, two reviews concluded that P4P will be more effective if desired behaviors are very specific and easy to track, and that complex rules for determining rewards are less effective (Kane et al., 2004; Town et al., 2005). Schatz (2008) adds to this by finding that the use of measures that are amenable to change was associated with positive results in five studies. Larger effects were found for process measures than for outcomes, as well as for measures with more room for improvement (van Herck et al., 2010). Finally, the results suggest that accurate/reliable data and adequate risk adjustment are vital and contribute to positive effects (Sorbero et al., 2006).

Regarding the *targeted entity*, the results suggest that P4P may be more effective when directed at individuals or small teams than when directed at (large) groups. Armour and Pitts (2003) found a study (level IV) in which incentives directed at individual physicians had greater impact on resource use in HMOs than when directed at groups of physicians, which may have been a result of a greater incentive for individuals to use resources prudently since the risk is not shared. In addition, Dudley et al. (2004) (only including RCTs) found five positive and two null results among studies in which the target was individuals and one positive and two null findings among studies in which the target was a group. Furthermore, Petersen et al. (2006), identifying evidence mainly from primary care settings in the US, show that five of the six physician-level studies (two level I, one level II, two level IV) found positive effects while seven of nine group-level studies either found partial (five: one level I, two level II, two level IV) or positive effects (two level I). Two institutional-level studies (both level I) found no effect. Finally, van Herck et al. (2010) found that programs targeting individuals or small teams were often more effective than programs targeting large groups or hospitals.

Regarding *performance targets*, results tend to more positive when absolute targets are used than when relative targets are used. For example, Armour and Pitts (2003) found that two RCTs evaluating programs with absolute targets showed a positive impact while an RCT using relative targets found no effect. Dudley et al. (2004) had a similar result: the two studies with relative targets found no effect, while four of five programs rewarding absolute performance had positive effects. A more recent review (Van Herck et al., 2010) also found programs using absolute targets to be more effective, although the relationship is not straightforward, in part due to the limited number of studies evaluating relative targets. The number of targets also seems to be relevant. Alshamsan et al. (2010) found strong

negative associations between scores in the previous year and improvement under the QOF, suggesting that adopting a series of targets, as in the QOF, may contribute to positive effects. Only using high targets may not motivate low performers, and may result in most rewards being awarded to providers already performing well before P4P (Christianson et al., 2008).

Regarding the *type and size of the incentive*, very little evidence is available on the relative effectiveness of bonuses and penalties. The only evidence is provided by Van Herck et al. (2010), who found that programs based on "new money" seem to have generated more positive effects than programs that relied on reallocation of existing funds. Regarding incentive size, Christianson et al. (2008) only found one study (level II, US Medicaid) finding that health plans that saw the largest improvements in the timeliness of well-baby care paid the largest rewards. Others also did not find a consistent "dose-response" relationship (Dudley et al., 2004; Kane et al., 2004; Van Herck et al., 2010). Three review authors speculate that the limited effectiveness thus far may have been a result of rewards being too small to elicit a response (Town et al., 2005; Sorbero et al., 2006; Schatz, 2008).

Regarding *payment frequency*, Emmert et al. (2012) found that programs in which there was little delay between care delivery and payment were all relatively successful (see chapter 4). In addition, van Herck et al. (2010) cite a level I study comparing the effect of quarterly versus annual payments for individual primary care physicians in a multispecialty group practice in California for nine preventive and chronic care measures. No difference was found between the two trial arms, but this may (also) have been a result of the small rewards (performance did not improve in both arms). Finally, regarding *provider engagement*, better results have been achieved in programs designed collaboratively with providers (e.g., when providers were involved in defining and selecting performance measures) and in which there was direct and extensive communication with providers regarding performance measurement and distribution of rewards (Christianson et al., 2008; van Herck et al., 2010). In several studies that failed to find an effect of P4P (largely level I and II), many providers were actually unaware of the incentives (Sorbero et al., 2006; Christianson et al., 2008).

5.5 DISCUSSION

5.5.1 Summary and comparison with other studies

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

relatively strong designs (level II) were generally less positive than results from studies with weaker designs (levels III and IV). Overall, the impact of physician P4P has been estimated at 5 percent improvement in incentivized measures. The reviews further highlight P4P's potential to be *cost-effective*. Yet most studies use narrow cost and effect ranges. In addition, the evidence largely pertains to relatively small programs. Two recent articles not included in the reviews (level III and level II) provide additional evidence that P4P can potentially be cost-effective. Walker et al. (2010) found that QOF payments were potentially a cost-effective use of resources for most GPs for most of the nine evaluated measures, but QOF administration costs were not taken into account. Cheng et al. (2012) examined the long-term effects of a national program for diabetes in Taiwan and found that P4P patients received more diabetes-specific exams/tests and had fewer hospitalizations than controls. Although total costs were higher in year one, continuously enrolled patients spent less than controls in subsequent years.

Regarding *unintended consequences*, the reviews identified one study finding evidence of risk selection. Other studies provide additional evidence. A qualitative study from California found that the inability to exception report led some physicians to deter noncompliant patients (McDonald & Roland, 2009). In addition, Wang et al. (2011, level II) found that physicians referred severely ill patients to higher-cost facilities under a performance-based incentive system in rural China, and Chen et al. (2011, level III) showed that older patients and patients with greater disease severity and/or comorbidity were more likely to be excluded from the diabetes P4P-program in Taiwan than younger and healthier patients. Chang et al. (2012, level II) had a similar finding. There is some evidence of spillover effects, with some studies finding less improvement for excluded conditions than for included conditions and reductions in continuity of care. Two recent studies (level II and level III) back this finding: Campbell et al. (2010) found a reduction in continuity of care after QOF implementation and Doran et al. (2011) found that although both incentivized and unincentivized aspects improved, improvements associated with financial incentives seem to have been achieved at the expense of small detrimental effects on unincentivized measures. Evidence on gaming behavior and negative effects on intrinsic motivation is absent, although a recent study (level III) revealed that UK GPs probably gamed the system of exception reporting to some extent (Gravelle et al., 2010).

Although many *inequalities* in chronic disease management have not been examined and the long-term effect on inequalities remains unknown, P4P seems to have narrowed socioeconomic inequalities in the UK. No evidence is available for other countries. A study by Doran et al. (2008b, level III) confirms this finding: inequalities in age, sex, and ethnicity have largely persisted, although there were small reductions for some measures. Lee et al. (2011, level II) had a similar result by finding that the QOF was associated with a decrease in inequalities in some measures between ethnic groups, but that clinically important inequalities have persisted. The evidence on the extent to which *non-financial incentives* can enhance

the P4P effect is limited. There is some evidence that feedback alone improves performance and that P4P does not add much when feedback is already provided. Conversely, while PR alone can stimulate quality improvement activity in hospitals (Fung et al., 2008), findings from the HQID in the US indicate that more favorable results can be achieved when P4P is added to PR. However, this only seems to hold for the short-term impact on process quality. A recent study (level II) on the long-term effect of the HQID showed that participation in the program was not associated with larger declines in mortality than those reported for hospitals that were only subjected to PR (Jha et al., 2012). Finally, the results highlight the importance of *program design*. Although the evidence is only suggestive, P4P seems to have been more effective when: measures are used that have much room for improvement instead of measures with low improvement potential; directed at individual physicians or small groups instead of larger groups or institutions; payments are based on providers' absolute performance instead of relative performance; designed collaboratively with providers instead of imposed top-down; larger payments are used. The latter is underscored by a recent level II study from the US that found that an increase in payment triggered an increase in behavioral response (Mullen et al., 2010).

We are aware of one other overview of systematic reviews examining the effects of financial incentives (Flodgren et al., 2011). There are several important differences with our review. First, the authors searched for reviews published until January 2010, while we searched until June 2011. Two additional reviews were published between these two dates (Steel & Willems, 2010; Emmert et al., 2012). Second, Flodgren et al. used other inclusion criteria, resulting in only four included reviews. In addition to a different search period, this large difference with our review can be explained by the fact that the authors required that reviews reported numerical data on outcomes, which was not required in our review. An important consequence of this requirement, however, is that several reviews that included studies investigating other effects besides the impact on incentivized performance (e.g., cost-effectiveness, unintended consequences, impact on inequalities) were excluded. Although these reviews indeed do not consistently report numerical data, they do provide relevant information on other P4P effects for which evidence is scarce already. Another explanation for the difference in the number of included reviews is that, judging from their search strategy, Flodgren et al. did not specifically aim to identify reviews investigating the effect of financial incentives for institutions, leading them to miss the Mehrotra et al. (2009) review on P4P in the hospital setting and the Briesacher et al. (2009) review on P4P for nursing homes (both are not on their list of excluded reviews). Of the four reviews included by Flodgren et al., we excluded three because they did not contain studies on explicit financial incentives or were entirely overlapped by another review that provides more details. Regarding the remaining review that was included in both overviews (Petersen et al., 2006), Flodgren et al. reached a similar conclusion as we did.

5.5.2 Limitations

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

5.5.3 Implications for research and policy

Notwithstanding these limitations, our findings have several implications. First, although many studies found improvements in selected quality measures and suggested that P4P can potentially be effective, at this point the evidence seems insufficient to recommend widespread implementation of P4P. Convincing evidence is still lacking (especially for inpatient care), despite the fact that P4P has widely been applied for many years now. In part, this lack of evidence could be a result from the fact that it is difficult to assess the impact of the financial part of "real-world" P4P-programs. Financial incentives are often introduced simultaneously with other improvement initiatives and thus as only one component of an improvement strategy. In many cases, the objective is solely to improve performance,

not to test the impact of financial incentives per se. However, to facilitate evidence-based policymaking on P4P, it is crucially important that improvement strategies are implemented in the context of rigorous evaluation, using convincing control groups to disentangle the effects of the different components. This would also provide insight in the relative merits of P4P and non-financial incentives; although different types of incentives have shown to be potentially effective when used in isolation, the literature remains almost silent on if and how they should be used together.

Second, thus far P4P evaluations have mainly focused on testing the short-term impact on clinical processes (e.g., screening for cancer, periodically performing eye exams for diabetes patients) and, to a lesser extent, intermediate outcomes (e.g., HbA1c levels of diabetes patients). However, the goal of P4P will typically be to improve health outcomes in the long run. Therefore, future evaluations should also assess the long-term impact on outcomes such as complication rates, hospital readmission rates, mortality, and quality of life. Valuable information will likely become available in the coming years. In the US, the Center for Medicare and Medicaid Services is currently employing a large national P4P-program for hospitals, which will be thoroughly evaluated (Health and Human Services, 2011). In addition, a large hospital P4P-program in England is being evaluated over a five-year period (NHS North West, n.d.). These evaluations also include assessments of health outcomes and costs (including the costs of program administration), which is urgently needed given the limited data that are available on P4P cost-effectiveness.

Third, although evidence is limited, P4P may have several unintended effects, underscoring the importance of ongoing monitoring and more insight in how specific design features may help in mitigating incentives for undesired behavior. We still know very little about the appropriate mix of performance measures that minimizes the risk of providers focusing disproportionately on rewarded performance. Similarly, although risk-adjustment methods for health outcomes have become more sophisticated, there is still a lot to learn about how they can be applied transparently; a specific method may be very effective in leveling the playing field, but incentives for selection will persist if providers perceive it as a black box and therefore reject to support it. Furthermore, undesired effects of P4P will often be a result of diminished intrinsic motivation. It is therefore important that providers are actively involved in designing the program, especially in developing and maintaining the aspects of performance to be measured. This increases the likelihood of provider support and alignment with their professional norms and values (see chapter 2). In this respect, it is also important that program evaluations include qualitative studies to monitor the impact on providers' intrinsic motivation. More generally, insight is required in which design features contribute to desired effects. Our results indicate that program design matters, yet few studies have specifically addressed design features, such as the effect of varying the size of the incentive holding other factors constant. Research is necessary to confirm our findings and to assess their influence in various contexts. In this respect, it is crucially important that

studies consistently report information on the specific setting in which the program was implemented and the study was conducted.

Fourth, although it is reassuring that P4P does not appear to have widened inequalities, most studies relied on cross-sectional data from the UK and many inequalities have persisted. An explanation for the latter may be that, with some notable exceptions (e.g., Blustein et al., 2011; Balicer et al., 2011), most P4P-programs are not designed to address inequalities or lack important features that would enable them to reduce inequalities (Chien et al., 2007). Rewarding improvement in performance and/or directly rewarding reductions in inequalities are good options to enhance current programs. A recent evaluation of the HQID (level II) found that a change in design from rewarding only top performance to also rewarding good performance and improvement resulted in a significant redistribution of funds toward hospitals caring for more disadvantaged populations, although significant gaps remained for incentive payments per discharge (Ryan et al., 2012b).

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

APPENDICES

Appendix 5.1 Literature search history

PubMed (08.07.2011; N=219)

Search "Pay for Performance"[Title/Abstract] OR P4P[Title/Abstract] OR PFP[Title/Abstract] OR "pay for value"[Title/Abstract] OR "payment for quality"[Title/Abstract] OR ("financial incentive" [Title/Abstract] AND effectiveness[Title/Abstract]) OR ("financial incentives" [Title/Abstract] AND effectiveness[Title/Abstract]) OR ("monetary incentive"[Title/Abstract] AND effectiveness [Title/Abstract]) OR ("monetary incentives"[Title/Abstract] AND effectiveness[Title/Abstract]) OR (bonus [Title/Abstract] AND "quality"[Title/Abstract]) OR ("reward"[Title/Abstract] AND "quality" [Title/Abstract]) OR "performance-based payment"[Title/Abstract] OR "performance-based reimburse-ment"[Title/Abstract] OR "performance-based contracting"[Title/Abstract] OR "performance-based pay"[Title/Abstract] OR "output-based payment"[Title/Abstract] OR "incentive reimbursement" [Title/Abstract] OR "incentive program"[Title/Abstract] OR "quality-based purchasing"[Title/Abstract] OR "quality incentive"[Title/Abstract] OR "quality incentives" [Title/Abstract] OR "quality-payment"[Title/Abstract] OR "quality-payments"[Title/Abstract] OR "quality-based payment"[Title/Abstract] OR (quality-based[Title/Abstract] AND payments[Title/Abstract]) Limits: Review, English, German, Spanish

The Cochrane Database of Systematic Reviews (08.07.2011; N=106)

Search "Pay for Performance" OR P4P OR PFP OR "pay for value" OR "pay for quality" OR "payment for quality" OR "payments for quality" OR "value based purchasing" OR ("financial incentive" AND quality) OR ("financial incentives" AND quality) OR ("monetary incentive" AND quality) OR ("monetary incentives" AND quality) OR ("financial incentive" AND effectiveness) OR ("financial incentives" AND effectiveness) OR ("monetary incentive" AND effectiveness) OR ("monetary incentives" AND effectiveness) OR (bonus AND quality) OR (reward AND quality) OR (rewards AND quality) OR "performance-based payment" OR "performance-based reimbursement" OR "performance-based contracting" OR "performance-based pay" OR "output-based payment" OR "incentive reimbursement" OR "incentive program" OR "quality-based purchasing" OR "quality incentive" OR "quality incentives" OR "quality-payment" OR "quality-payments" OR "quality-based payment" OR "quality-based payments" OR "performance-related payment" OR "performance-related payments"; Title, Abstract or Keywords

Embase (08.07.2011; N=236)

(((((((((FT="Pay for Performance" OR FT="P4P") OR FT="PFP") OR FT="pay for value") OR FT=pay for quality") OR FT="payment* for quality") OR FT="value based purchasing") OR FT="financial incentive*" AND quality) OR FT="performance-related payment*") AND (LA=SPANISH OR LA=ENGLISH OR LA=GERMAN)) (((((((((FT="monetary incentive*" AND quality OR FT="financial incentive*" AND effectiveness) OR FT="monetary incentive" AND effectiveness) OR FT=bonus AND quality) OR FT=reward* AND quality) OR FT="performance based payment") OR FT="performance-based reimbursement") OR FT="performance-based contracting") OR FT="performance-based pay") AND (LA=SPANISH OR LA=ENGLISH OR LA=GERMAN)) (((((((FT="output-based payment" OR FT="incentive reimbursement") OR FT="incentive program") OR FT="quality-based purchasing") OR FT="quality incentive") OR FT="quality-payment*") OR FT="quality-based payment*") AND (LA=SPANISH OR LA=ENGLISH OR LA=GERMAN))

Scopus (08.07.2011; N=1,031)

TITLE-ABS-KEY("Pay for Performance") OR TITLE-ABS-KEY("P4P") OR TITLE-ABS-KEY("PFP") OR TITLE-ABS-KEY("pay for value") OR TITLE-ABS-KEY("pay for quality") OR TITLE-ABS-KEY("payment* for quality") OR TITLE-ABS-KEY("value based purchasing") OR TITLE-ABS-KEY("financial incentive*" AND quality) OR TITLE-ABS-KEY("performance related payment*") OR TITLE-ABS-KEY("monetary incentive*" AND quality) OR TITLE-ABS-KEY("financial incentive*" AND effectiveness) OR TITLE-ABS-KEY("monetary incentive" AND effectiveness) OR TITLE-ABS-KEY(bonus AND quality) OR TITLE-ABS-KEY(reward* AND quality) OR TITLE-ABS-KEY("performance-based payment") OR TITLE-ABS-KEY("performance-based reimbursement") OR TITLE-ABS-KEY("performance-based contracting") OR TITLE-ABS-KEY("performance based pay") OR TITLE-ABS-KEY("output-based payment") OR TITLE-ABS-KEY("incentive reimbursement") OR TITLE-ABS-KEY("incentive program") OR TITLE-ABS-KEY("quality-based purchasing") OR TITLE-ABS-KEY("quality incentive") OR TITLE-ABS-KEY("quality-payment*") OR TITLE-ABS-KEY("quality-based payment*")

ISI Web of Knowledge (08.07.2011; N=219)

Topic=("Pay for Performance") OR Topic=("P4P") OR Topic=("PFP") OR Topic=("pay for value") OR Topic=("pay for quality") OR Topic=("payment* for quality") OR Topic=("value based purchasing") OR Topic=("financial incentive*" AND quality) OR Topic= ("performance-related payment*") OR Topic=("monetary incentive*" AND quality) OR Topic=("financial incentive*" AND effectiveness) OR Topic=("monetary incentive" AND effectiveness) OR Topic=(bonus AND quality) OR Topic=(reward* AND quality) OR Topic=("performance-based payment") OR Topic=("performance-based reimbursement") OR Topic=("performance-based contracting") OR Topic=("performance-based pay") OR Topic=("output-based payment") OR Topic=("incentive reimbursement") OR Topic= ("incentive program") OR Topic=("quality-based purchasing") OR Topic=("quality incentive") OR Topic=("quality-payment*") OR Topic=("quality-based payment*")

Appendix 5.2 Methodological quality assessment results

Reference	Total score [a]	Research question appreciable?	Sources documented?	Search strategy documented?	Inclusion criteria defined?	Exclusion criteria defined?	Quality criteria considered?	Independent evaluaItion?	Excluded studies documented?	Replicable data extraction?	Independent data extraction?	Replicable information synthesis?	Evaluation of the evidence?	Answer to the research question?	Integration of evidence in summary?	Discussion of the limitations?
Alshamsam et al. (2010)	93%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Armour and Pitts (2003)	73%	1	1	1	0	0	1	1	0	1	0	1	1	1	1	1
Briesacher et al. (2009)	87%	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1
Chaix-Couturier et al. (2000)	87%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0
Christianson et al. (2007)	87%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0
Christianson et al. (2008)	87%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0

Dudley et al. (2004)	93%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Eldridge and Palmer (2009)	60%	1	1	1	1	0	0	0	0	1	-	1	1	1	1	0
Emmert et al. (2012)	93%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Frølich et al. (2007)	93%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Giuffrida et al. (2000)	100%	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kane et al. (2004)	93%	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Mehrotra et al. (2009)	87%	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
Petersen et al. (2006)	100%	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Rosenthal and Frank (2006)	73%	1	1	1	1	1	1	0	0	1	0	1	1	1	1	0
Sabatino et al. (2008)	80%	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0
Schatz (2008)	67%	1	1	1	0	0	1	0	0	1	0	1	1	1	1	1
Scott et al. (2011)	93%	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
Sorbero et al. (2006)	73%	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1
Steel and Willems (2010)	66%	1	1	0	1	0	1	0	0	1	0	1	1	1	1	1
Town et al. (2005)	73%	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0
van Herck et al. (2010)	100%	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

a. 3 items ('Meta-analysis?', 'Test of heterogeneity?', 'Test of sensitivity?') were inapplicable for all reviews.

Appendix 5.3 Summaries of the main findings of included reviews, by domain

1. To what extent has P4P been effective?

Armour and Pitts (2003) found seven studies examining the effect of explicit financial incentives on resource use and/or quality. Regarding resource use, one study found that placing physicians at financial risk for deficits in referral funds decreased the number of primary care visits, and that financial risk for the cost of outpatient tests reduced the number of outpatient tests. Bonuses and withholds for productivity did not change resource use. Another study found large reductions in the number of hospitals admissions and the mean number of visits as a result of providing bonuses for reduced resource use. Regarding quality, a survey among 766 physicians showed more than half reported feeling pressure to decrease the number of referrals, which many believed negatively impacted the quality of care. Three studies assessed the effect on the quality of care and found mixed results. One RCT found no difference between intervention and control groups in cancer screening rates after eighteen months (semi-annual bonuses of 10 to 20 percent of capitation were paid to the top six practices). Another RCT found that relatively small explicit financial incentives (ten or 20 percent add-on to the current $8 fee for 70 or 80 percent attainment, respectively) can improve immunization rates: the mean immunization rate was 6 percentage points higher than the mean rate in the control group. Also, median change was higher in the intervention group: 10.3 percent versus 3.5 percent. A study from Northern Ireland in the early 1990s found that after bonuses of $1,000 and $3,000 for attaining 70 percent and 80 percent attainment for childhood immunization rates were implemented in primary care. By 1991, 90 percent of GPs had reached the lower target and 77 percent reached the upper target (rates were 12 percent overall in the 1980s). However, it is unclear whether these improvements can be attributed to the financial incentives. *Briesacher et al. (2009)* found little evidence that P4P increases the quality and efficiency in nursing homes. One RCT investigated the impact of a P4P-program in San Diego in the early 1980s. The intervention group received several financial incentives, including bonuses for improved functional status and timely discharges.

The program had beneficial effects on access and quality. Compared to control homes, homes in the intervention group were more likely to admit severely disabled people and to have lower length of stay. In addition, the probability of hospitalization and death was lower in intervention homes. The three observational studies showed a modest or no effect, with one study demonstrating small improvements in patient satisfaction and staffing and employee retention. *Chaix-Couturier et al. (2000)* describe the results of two RCTs that were also included in other reviews, including the one by Armour and Pitts. One found no difference between intervention and control groups in cancer screening rates; the second one found that a potential 10 to 20 percent increase in current fees for reaching absolute performance targets can increase immunization rates: the mean rate was 6 percentage points higher in the intervention group. *Christianson et al. (2007)* found little difference among seven reviews in studies reviewed and thus in the conclusions. All noted the small number of studies on the impact of P4P, and conclude that the evidence base for both justifying and designing P4P-programs to improve quality of care is thin. The authors themselves found that studies evaluating initiatives that aim to reward providers for quality improvement or the attainment of quality benchmarks have mixed results. Few significant impacts are reported, although there was improvement in selected quality measures. None of the RCTs provided unequivocal support for the premise that the use P4P an effective way to improve quality of care. Regarding the non-RCT studies, the impact of P4P was difficult to determine due to concurrent improvement efforts. Even for preventive services, to which relatively much attention has been paid with respect to P4P evaluations, there is limited evidence that targeted interventions with P4P are effective. The few studies in this area with strong designs find small, if any, effects of payments to providers that are intended to improve quality. Finally, the authors found too little evidence on the effect of P4P for institutional providers to draw conclusions. *Christianson et al. (2008)* found that most initiatives showed improvement in selected quality measures, but the incentives were typically implemented together with other improvement efforts and studies often lacked a control group. Regarding the physician-oriented programs, evaluations had similar outcomes, each finding improvement in at least one measure. One study found a modest increase in the improvement rate for asthma and diabetes, but not for heart disease. Two other studies found improvements for diabetes and (small) improvements for asthma. Another physician-oriented study found improvements for one of three evaluated measures. Five studies reported findings of evaluations of hospital P4P-programs. One found sustained improvements in two of the three measures. Three of the other studies were evaluations of the HQID. One study found greater improvement in P4P hospitals for three conditions, the second one reported improvements for two of six incentivized measures but no significant difference in the composite score, and the final study found a modest impact (2.6 to 4.1 percent improvement in the composite score over two years). *Dudley et al. (2004)* found that among eight RCTs, ten dependent variables were analyzed. Among the seven studies focusing on physicians (nine dependent variables), five variables showed a significant relationship to the incentive in the expected direction while four showed no change after the incentive was introduced. The remaining study focused on pharmacists and the result was positive. Thus, overall, the results are mixed and inconclusive. *Eldridge and Palmer (2009)* identified 27 studies found a lack of clear evidence on the effects of any type of performance-based payment in any low-income country setting. This was largely due to the absence of controls in most studies; the only that did include control sites found that they outperformed those with performance-based payments. Nonetheless, most of the papers provided a favorable assessment of performance-based payment. *Emmert et al. (2012)* identified seven studies, six of which showed that improved quality of care can be achieved, often at the cost of higher expenses (see chapter 4). The eight RCTs found by *Frølich et al. (2007)* showed that the evaluated P4P-programs had mixed results, although key aspects such as design elements and contextual factors were not reported. The authors therefore conclude that the potential to improve quality through the use of P4P remains unknown. *Giuffrida et al. (2000)* evaluated the impact of target payments on the professional practice of primary care physicians and healthcare outcomes. They identified two studies, both of which investigated the impact on immunizations. The use of

target payments was associated with improvements in immunizations rates, but the increase was statistically significant in only one study. The authors noted that the evidence was insufficient to draw conclusions about the effectiveness of target payments in improving quality of care. *Kane et al. (2004)* concluded that results are mixed; four studies found positive effects and five studies found no effects. Improvements in documentation may account for the positive effects, and the effect size is moderate at best. One study found that P4P was associated with a 7 percent increase in immunization rates. The authors conclude that the literature is scarce and that there is little evidence that explicit provider financial incentives are effective. *Mehrotra et al. (2009)* found eight studies addressing the effects of three separate P4P-programs. Regarding the first program (the Hawaii Medical Service Association Hospital Quality Service and Recognition Program, two studies), improvements were found over time in complication rates, length of stay, and patient satisfaction. However, there was no control group, no trend data, and no information of whether observed differences between periods were statistically significant. Regarding the second program (BCBS of Michigan's Participating Hospital Agreement, three studies) one study found increases in process measures over time, but statistical results are not reported and the study lacked a control group and trend data. The last program (the Hospital Quality Incentive Demonstration) was analyzed in three studies. The first found that participating hospitals improved more on several process measures than nonparticipating hospitals (overall 9.3 vs. 6.7 percentage points). The second found that participating hospitals experienced a 2.9 percentage point greater improvement than control hospitals on a composite measure constructed from ten measures (the difference was seen consistently for each of the three clinical conditions and for most individual measures). The third found that differences between intervention and control hospitals in their improvement on a composite score for incentivized and non-incentivized measures were not statistically significant. Intervention hospitals had greater improvement on three individual measures (two incentivized, one unincentivized). There were no statistically significant differences in improvements in inpatient mortality between the two groups. Overall, the authors conclude that there remains a substantial gap in the knowledge as the most rigorous evidence (a 2 to 4 percentage point improvement beyond the improvement seen in control hospitals) comes from a single program. However, the effect of P4P without PR remains unknown and the studies only provide evidence on a few clinical conditions. *Petersen et al. (2006)* identified seventeen studies that addressed the question whether explicit financial incentives can improve the quality of care. Five of the six physician-level incentives and seven of the nine group-level incentives found partial (five) or positive effects (two) of the financial incentives on measures of quality. Among the studies on group-level incentives, two RCTs found that the incentives for preventive health services had no effect compared with the control group. One of the two studies investigating incentives at the "payment-system level" found a positive effect on access to care (which was incentivized), while the other found evidence of selection. Overall, the authors conclude that there are few informative studies of explicit financial incentives for quality and that this literature suggests some positive effects of financial incentive programs. *Rosenthal and Frank (2006)* included six studies, four of which were RCTs. The authors conclude that the empirical foundation of P4P in health care is weak. Most studies found no effect with only two positive findings (one study found a 4 percentage point improvement in immunization rates from baseline relative to the control group. Another found a 7.9 and 5.9 percentage point difference between intervention and control groups in identification of smokers and providing quit smoking advice to smokers, respectively, though no significant impact was found on smoking cessation rates). The authors suggest that the limited effectiveness may have been due to small dollar amounts per patient and small shares of eligible patients involved, as well as small sample sizes and the short time period over which the effect of the intervention was studied (in one study the impact of P4P was assessed over an eight month period). *Sabatino et al. (2008)* briefly discuss the results of three studies. The study without a control group found a statistically significant 8 percentage point increase in completed cervical cancer screening within six months of increasing GPs' compensation for performing screening tests. The second study found no significant differences

between intervention (bonus + reminders) and control group (reminders only) in recommended or ordered mammography after one year. The final study also found no statistically significant differences between intervention group (feedback and P4P) and control group (no intervention) in recommended and/or ordered screening tests for breast, cervical, and colorectal cancer. Based on the results of these few studies, the authors conclude that "there is insufficient evidence to determine the effectiveness of provider incentives in increasing screening for breast, cervical, or colorectal cancers." *Schatz (2008)* found that among seven RCTs, all of which focused on preventive services, three found no effect, three found positive effects, and one had mixed results. The three studies that found no effect suffered from small sample sizes, used relatively small bonuses per physician, and focused on Medicaid populations in the US, which generally involves hard-to-reach patients and low reimbursement. The positive results in two trials were primarily due to better documentation. Among the fifteen nonrandomized studies, all but one showed positive (ten) or mixed (positive and null) results (four). Among the seven uncontrolled before-after studies, all but one found positive results (against three positive results among the six controlled before-after studies). Based on these results, the author concludes that P4P can improve quality, but not always. And even for the positive studies, the research designs often do not allow for strong conclusions that observed effects are really a result of the P4P payments. *Scott et al. (2011)* found that six studies showed positive but modest effects on quality of care for some measures, but not all (typically only one out of a range of measures). In the three RCTs, significant effects were found on providers' behavior (e.g., recording of smoking status), but not on measures of smoking cessation. Three other studies that examined testing in diabetes and screening found significant effects only for cervical cancer (typically a 3.5 up to 6 percentage point difference relative to the control group) and eye tests. Another study found increases across all measures immediately after incentive implementation, but there was no statistically significant difference between before- and after-intervention trends. Methodological shortcomings led to substantial risk of bias in most studies, especially selection bias as a result of voluntary participation. The authors conclude that "there is insufficient evidence to support or not support the use of financial incentives to improve the quality of primary health care. Implementation should proceed with caution and incentives schemes should be more carefully designed before implementation." *Sorbero et al. (2006)* found that among the seven RCTs they identified, four had mixed findings and three reported no effect. The two quasi-experiments found mixed results and six before-after studies tended to report positive results for at least one measure. The findings should be interpreted with caution because of small bonuses, short study periods, lack of control groups, and varying contexts. In addition, in at least two positive studies, improvements were largely due to improvements in documentation and charting rather than actual improvements in performance. Furthermore, most of the programs evaluated in these studies do not resemble current programs in terms of size, duration, and magnitude of payments. Thus, the authors conclude that "taken together, the findings (…) suggest that it is still too early to determine the effect of physician-focused P4P-programs. The published literature provides and ambiguous set of results." *Steel and Willems (2010)* found that national data showed high overall achievement in the first QOF year that has increased since. Regarding diabetes, substantial improvements were demonstrated between 2004 and 2008, sometimes even more than 40 percent. Improvements were also demonstrated for other conditions and aspects, including heart disease, stroke, hypertension, and smoking indicators. However, although for some measures the QOF seems to have slightly increased improvement that was already occurring, in nearly all cases the trend showed a gradual improvement over the five years with little change around 2004 when the QOF was implemented. For diabetes and asthma, a small but significant above the trend increase was found, while another study found both slightly lower and slightly higher achievement than that predicted by the underlying trend. Regarding health outcomes, most studies find little relationship between achievement on process measures and clinical outcomes such as admissions and mortality, with one exception: a significant relationship was found between achievement for epilepsy patients and related emergency hospitalizations. The authors highlight the difficulty of drawing

conclusions on the impact of the QOF based on the available observational studies. *Town et al. (2004)* showed that six studies generated eight separate outcomes (four studies on immunizations, two on cancer screening, and one on an assortment of preventive services). In only one of eight outcomes did the financial incentive result in a significant improvement of preventive care performance. One study found a statistically significant difference between the bonus plus feedback group and the control group, but not between the bonus plus feedback group and feedback alone. Another study, classified as ineffective, found that most of the observed increase in immunizations was due to better documentation and not due to physicians providing more immunizations. *Van Herck et al. (2010)* found that 39 studies reported a clinical effect size, which ranged from negative to absent to positive (1 to 10 percent) to very positive (more than 10 percent). Negative results were found in three studies, but these studies also found positive results on other measures. Overall, P4P has led to an estimated 5 percent performance improvement, although there is much variation. For prevention, results were more positive for immunizations than for screening. For chronic care, positive results were reported especially for diabetes but also for asthma and smoking cessation (no effect was found for heart disease). P4P most often failed to improve acute care. Two studies show that P4P may have a positive impact on coordination of care when explicitly rewarded. The authors note that "as the evidence continues to grow, conclusions on the effect of P4P can increasingly be drawn with more certainty, despite that fact that the quality of current evidence is still poor".

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

P4P can be considered cost-effective when improved quality is achieved with equal or lower costs, or alternatively, when the same quality of care is achieved using less financial resources. In the likely case that P4P leads to cost increases, it may still be viewed as cost-effective as long as improvements in quality are large enough. *Armour and Pitts (2003)* found two studies finding that bonuses and withholds for reduced resource use may lead to reduced outpatient expenses and utilization. One additional study, which found that physicians reduced outpatient medical expenditures by an average of 5 percent after having been offered bonuses and withholds (the impact on quality was not investigated). *Briesacher et al. (2009)* identified one RCT showing that while the evaluated P4P-program had a positive impact on access and quality, this came at the cost of a 5 percent increase in the average daily cost to Medicaid due to the bonus payments and an increased administrative burden. *Christianson et al. (2007, 2008)* found two studies that assessed P4P cost-effectiveness. The first study, addressing diabetes care, found a positive return on investment of 1.6 to 1 and 2.5 to 1 in year one and two, respectively. In the second study, QALYs gained for patients hospitalized for heart treatment were compared with the money spent in incentive payments. The authors calculated a cost per QALY of between $12,967 and $30.081. *Emmert et al. (2012)* conclude that P4P has the potential to be cost-effective, although results are not convincing (see chapter 4). Five studies found that the evaluated programs improved quality but also increased expenses. One study showed that savings may be possible even when quality improves. Another study demonstrated cost increases as well as no impact on 30-day mortality. Finally, two studies only investigated the impact on costs and both showed a positive impact. However, narrow cost and/or effect ranges, methodological flaws, and differences across programs in context and design impede strong conclusions about P4P cost-effectiveness. *Giuffrida et al. (2000)* and *Kane et al. (2004)* found one study showing that the additional cost per extra immunization using target payment incentives was $3.02. *Mehrotra et al. (2009)* found one study that performed a cost-utility analysis that found an estimated cost per QALY range of $12,967-30.081, which is generally considered cost-effective. However, the study lacked a control group or trend data. Also, the study did not include costs incurred by hospitals for collecting quality data and for quality improvement activities. *Petersen et al. (2006)* discuss a study showing that using a combination of various types of incentives to improve both access to nursing home care and patient outcomes saved an

estimated $3,000 per stay in a Markov model, despite the administrative and incentive costs of the program. *Steel and Willems (2010)* found a study finding evidence of cost-effectiveness for twelve measures in the QOF with direct therapeutic effect. Cost-effectiveness varied by the measure's baseline achievement, with smaller improvements necessary to be cost-effective at low baseline achievement than at higher baseline achievement. *Town et al. (2004)* discuss one study finding that the 7 percentage point increase in the immunization rate resulted in a cost of $3 per additional immunization. As flu vaccines have been shown to save $117 in direct medical expenditures in the elderly, the authors classify this intervention as cost-effective. *Van Herck et al. (2010)* found one study reporting a 2.5-fold ROI per dollar spent, which seems to have resulted from cost savings. According to the authors, the four studies reporting on P4P cost-effectiveness all find positive results, although interpretation is difficult due to flaws in study design.

3. Which unintended consequences has P4P had?

Armour and Pitts (2003) discuss a study that found that physicians at risk for cost of outpatient tests substituted primary care visits for outpatient tests, increasing the number of visits per enrolee per year by 5 percent. *Christianson et al. (2008)* found two studies that evaluated the impact on unincentivized performance measures. The first found no differences in trends in seventeen non-incentivized measures and the incentivized measures. The other study found no significant change in a composite of eight non-P4P measures (the composite of the six P4P measures also showed no significant change). A qualitative study conducted in the context of the UK QOF found that the program did not impair GPs' intrinsic motivation to provide high quality care. GPs also did not question the performance targets or their implications. The authors further emphasize that the results show that initial improvements may reflect better documentation by providers of care they are already delivering. *Mehrotra et al. (2009)* found a study that assessed whether P4P led to worse performance on quality measures for acute myocardial infarction (AMI) that were not used as a basis for incentive payments. On the composite score for these measures, the difference in improvement between intervention and control hospitals was not statistically significant, but for one individual measure intervention hospitals improved more. However, these other measures were for the same condition (AMI) that was used in the P4P-program; therefore, this study provides no insight in negative or positive spillovers on other conditions not used as a basis for payment. *Petersen et al. (2006)* found four studies finding that P4P had unintended effects, including selection (one study) and improvements in documentation rather than a change in the quality of health care (three studies). The authors therefore note that the findings suggest that adequate design and ongoing monitoring of incentive programs is critical to prevent unintended effects. *Rosenthal and Frank (2006)* cite several studies that suggest unintended consequences (i.e., upcoding under prospective payments, selection under performance-based contracting and public reporting, and gaming), although these studies do not typically focus on P4P per se. The authors conclude that "findings related to selection, gaming, and other forms of unintended consequences are a reminder that even in health care, agents behave strategically, and P4P-programs need to be designed carefully to be welfare improving." *Schatz (2008)* identified one study that found no difference between measures were linked to rewards and measures from the same disease not linked to rewards, which could suggest a positive spillover effect. *Scott et al. (2011)* found one study that did not find evidence of positive or negative spillovers on unincentivized aspects of care, although some rewarded measures improved. *Steel and Willems (2010)* identified several studies that examined unrewarded conditions. Achievement in these conditions has typically been (much) lower than that for incentivized conditions, which did not change after QOF implementation. The authors suggest that unrewarded conditions may have received less policy attention. Qualitative studies found no evidence that providers' intrinsic motivation has been crowded out, although GPs were more supportive about targets aligned with professional norms. In addition, GPs and nurses are concerned about a dual agenda, less time for holistic care, unincentivized care, and reductions in the continuity of care. *Van Herck et al. (2010)* found

that effects on unincentivized aspects varied from null to positive. One study found a reduction in the improvement rate for unincentivized measures for asthma and heart disease after performance reached a plateau. Another study found that P4P had a positive impact on included quality measures for heart disease, COPD, hypertension, and stroke when applied to patient groups not included in the program. One study found no effect on unincentivized measures of access and communication, but did find a decrease in timely access to patients' regular doctors. The authors find little evidence of gaming. However, only a few studies have addressed it.

4. To what extent has P4P affected inequalities?

Alshamsan et al. (2010) found 22 studies assessing the impact of P4P on inequalities in the quality of health care in relation to age, sex, ethnicity and socioeconomic status. Most studies investigate the impact of the QOF; only one study reported findings from the US. One study reported about the long-term impact on inequalities. Eighteen studies addressed socioeconomic inequalities, mostly cross-sectional studies examining associations between the quality of care and an "area deprivation score" after the QOF implementation. For example, several studies calculated the difference in achievement between practices in the least and most deprived areas. One study found a positive association between deprivation and higher quality in the first QOF year, whereas the remaining studies found lower quality in deprived areas compared with affluent areas before or shortly after QOF. Though generally significant, the identified differences were relatively small and appear to have narrowed, sometimes considerably, in the second and third QOF years. For example, one study found that the gap in median achievement between most and least deprived practices narrowed from four to 0.8 percent during the third year. However, two before-after studies found that patients living in deprived areas were less likely to have their medical data (e.g., smoking status, blood pressure) recorded than patients living in affluent areas after the P4P implementation (this difference was not evident before). Only one study focused on the long-term effect of P4P on inequalities, demonstrating that the initial widening of inequalities between affluent and deprived areas in cervical cancer screening coverage had almost disappeared after five years. In sum, the introduction of QOF was associated with reductions in socioeconomic inequalities in chronic disease management, although the extent to which the QOF has contributed to this finding remains largely unclear. In addition, important inequalities in chronic disease management have not been addressed thus far. Nine studies assessed the impact of P4P on age, sex and ethnic inequalities with respect to stroke, coronary heart disease, and diabetes patients. Existing inequalities in quality of care appear to have persisted; especially women, older patients, and those from some minority ethnic groups continued to receive lower quality of care after QOF compared with men, younger patients, and the white British group, respectively. *Christianson et al. (2008)* found one study evaluating the impact of the QOF on record keeping of Scottish GPs for stroke patients. They found a large increase in record keeping in the most affluent areas. In addition, women had a larger increase in documentation than men. However, "inequitable recording still persists, with lower recording for women, older patients, and more deprived patients." *Steel and Willems (2010)* found ten studies that assessed the impact of the QOF on inequalities. All groups benefited from observed improvements in achievement, but the relative rate of improvement differed between groups. As noted by the authors, changes in inequalities were small, variable, dependent on the measure, achievement before QOF, and the demographic variable (age, sex, socioeconomic status, and ethnicity). Regarding age, the gaps in care between age groups for CHD, diabetes, and CVD were attenuated after the QOF. For measures with lower achievement for older people, larger improvement was observed among older people. Regarding sex, both before and after QOF higher achievement was found for men for nearly all CHD and CVD measures and three of eight diabetes measures (thus, inequalities seem to have persisted). Inequalities between the most and least deprived areas have almost completely disappeared in England. However, large differences remain in individual measures and the poorest performing practices remain concentrated in the

most deprived areas. Finally, both before and after the QOF results have been variable regarding ethnicity. Gaps in CHD performance between black and white people reduced after QOF in some measures. For other conditions, variations among ethnic groups were not reduced after the QOF was introduced. *Van Herck et al. (2010)* found some evidence, mainly from the UK, that P4P affects inequalities. In general, P4P did not appear to have a negative effect on age, ethnic, and socioeconomic inequalities. This is backed by evidence from 28 studies, which seems to point to a reduction in inequalities in quality across groups rather than an increase.

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

Armour and Pitts (2003) found that an RCT in which P4P was combined with semi-annual performance feedback in the intervention group showed no impact on cancer screening rates. *Emmert et al. (2012)* concluded that programs that also provided providers with performance feedback and/or publicly reported performance scores did not have more or less favorable results than programs that did not include such non-financial incentives (see chapter 4). *Mehrotra et al. (2009)* showed that the three studies that assessed the impact of the Hospital Quality Incentive Demonstration in the US found a two to 4 percentage point improvement beyond the improvement seen in control hospitals. The authors emphasize that PR may have contributed to these findings (performance scores were simultaneously publicly reported), perhaps more than P4P. However, the studies failed to disentangle the effects of P4P and PR. *Rosenthal and Frank (2006)* found one RCT in which neither performance feedback alone not feedback together with P4P improved childhood immunization rates. Conversely, another RCT found that while neither feedback alone nor enhanced fees improved the likelihood of childhood immunization, providing a rather sizeable bonus did improve immunization rates, although this was primarily achieved through better documentation. Finally, a third RCT showed that while the group with only a financial incentive improved significantly more than the control group, the group with a financial incentive and access to the patient registry and telephonic counseling system showed no improvement relative to the control group. *Schatz (2008)* tentatively concluded that combining P4P with other strategies such as information system enhancements, guidelines, feedback, and public reporting may contribute to P4P success. *Sorbero et al. (2006)* noted that performance monitoring can have the overall effect of improving performance (whether tied to financial incentives or not). Also, interviews with sponsors of P4P-programs and physicians revealed that P4P needs to be implemented as part of a multifaceted strategy to performance improvement. *Town et al. (2004)* found that in one study, the group that only received formal performance feedback failed to increase their mammography referrals more than the group receiving feedback and a $50 bonus. Another study found that the feedback only group was not statistically significantly different from the feedback plus financial incentive group or the control group. *Van Herck et al. (2010)* discuss three non-randomized studies from the UK, Spain, and Argentina that found that P4P may leads to positive results (five to 30 percent effect size) when part of a larger quality improvement strategy also including elements of PR, performance feedback, and provider education. Evidence from the US (28 studies) is more mixed.

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

Alshamsan et al. (2010) found a significant positive association was found between scores in the previous year and improvement, suggesting that using measures with low baseline performance and/or adopting a series of targets (i.e., also using targets that are attainable for low performers) may yield the largest benefits. In addition, the authors referred to both process measures (e.g., recording of smoking status and blood pressure) and outcome measures (e.g., achieving desired blood pressure levels). The observed impact of P4P on inequalities was mainly found for process quality but hardly for outcomes. *Armour and Pitts (2003)* found that

regarding quality of care, the two studies with absolute targets both found a positive impact on performance while the study using relative targets was ineffective. In addition, one study found that incentives directed at individual physicians had a greater impact on resource use than incentives directed at the group level. As noted by the authors, "an individual physician who bears all the risk has a greater incentive to use resources more parsimoniously than physicians who share their risk with a group." The authors of another study finding no effect on cancer screening rates relate their findings to a lack of physician awareness, small incentive size, and limited time frame of the intervention. However, bonuses were 10 to 20 percent of capitation payments for practices in the top quartile, which is larger than in many other P4P-programs. *Chaix-Couturier et al. (2000)* found that regarding negative incentives, previously penalized physicians tended to comply more readily. In addition, the authors found that physicians are more likely to respond when informed about the thresholds that trigger sanctions and on the actual financial risk. *Christianson et al. (2008)* found that one program that was designed collaboratively with participating providers showed sustained improvements for two of the three evaluated measures for three years. Another study reported better results for well-baby care where there was better communication with physicians regarding performance measurement and reward distribution. In addition, two early studies that failed to demonstrate improvements in measures of preventive care found that only about 50 percent of participating practices were aware of the incentives. Christianson et al. found no quantitative evidence with respect to optimal payment size, although one study found that the plans that showed the largest improvements paid the largest rewards. Finally, one studies (both evaluating programs that paid bonuses for achieving preset performance targets) found that most of the payments were awarded to provider already meeting these targets at baseline, although providers at all performance levels showed improvement. *Dudley et al. (2004)* found that among the studies targeting individual providers, there were five positive and two null results, while among the studies in which the target was always or could be a group of physicians, there were one positive and two insignificant findings. The authors found no consistent relationship between the magnitude of the incentive and the response (in fact, the largest single incentive was shown ineffective). In two studies providers' performance was measured in a relative way, and both yielded negative results. In addition, among the five studies adopting enhanced fee-for-service, four positive and one insignificant result were found, whereas when bonuses were used (four studies), the authors found two positive and one null result. Finally, in general the authors found that incentives to achieve performance were more effective when the indicator to be followed required less patient cooperation, highlighting the importance of careful consideration of which performance indicators to include. *Emmert et al. (2012)* found weak evidence that larger payments increase P4P (cost-)effectiveness (see chapter 4). In addition, the one program using only relative targets (combined with a small bonuses and a small penalty for low performers) was likely resulted in cost increases without reducing 30-day mortality. Yet, there is no evidence that the use of penalties affects P4P (cost-)effectiveness. Furthermore, programs targeting a specific type of provider (hospitals, primary care groups, individual physicians, etc.) were not demonstratively more effective than programs targeting another type of provider. Finally, although the authors found no clear effect of payment frequency, the three programs in which there was little delay between care delivery and payment were all relatively successful. *Frølich et al. (2007)* showed that among the five studies (with seven dependent variables) targeting individual providers, five were positive and two negative; among the three studies in which the recipient was a group, one was positive and two were negative. In addition, although the studies did not report much information on incentive size, the authors detected no clear dose-response relationship. Furthermore, five studies assessed the impact of enhanced fee-for-service and four were positive; among the four studies (with five dependent variables) that examined bonuses, two were positive and three negative. One study examined the difference in effect between bonus payments and enhanced fee-for-service and found no difference. *Kane et al. (2004)* conclude that P4P does not work easily and that design matters. While there was some evidence that effects were larger for group practices than solo practices, there is not enough informa-

tion to sort out the causes. The desired behaviors must be very specific and easy to track; complex rules are less effective. In addition, the incentive must be of sufficient size, although the literature is not clear about a dose-response relationship. Both studies evaluating programs with relative performance target(s) found no effect, while of the seven studies with absolute targets, four found modest positive effects. However, these two also examined incentive salience and both found low provider awareness of the program. *Petersen et al. (2006)* showed that five of the six physician-level and seven of the nine group-level incentives found partial (five) or positive effects (two). Two RCTs examining group-level incentives found no effect. *Schatz (2008)* suggests that among the studies finding positive effects, the size of the incentives and the use of measures that are more amenable to change apparently were associated with their success. Studies reporting null effects typically used relatively small bonus payments. However, one positive study achieved the results with a relatively small bonus, so a clear dose-response relationship was not demonstrated. *Sorbero et al. (2006)* noted that the peer-reviewed literature does not provide information about the various design features that may have played a role in an intervention's success or failure. Nonetheless, the evidence suggests that the lack of effects may have been due to the payments being of too small a magnitude to elicit behavior response; there is some weak evidence that a minimum of 5 percent of practice revenues is necessary to capture physicians' attention. In addition, the authors from at least three studies believed that low awareness among physicians of the intervention contributed to the lack of effect. Furthermore, interviews with program leaders revealed that (1) physician involvement and engagement is critical, (2) it is essential to pilot tests the various implementation processes, (3) accuracy and reliability of data and an equitable process for appeals are vital, (4) ongoing evaluation is needed, and (5) physicians need support such as patient registries and education. *Town et al. (2004)* found that studies that failed to find a positive relationship between P4P and the level of preventive care provision are roughly evenly split between bonuses and enhanced fee-for-service. The authors therefore suggest that neither the type of payment nor the type of preventive service drives the lack of findings, though they may ultimately be related to the efficacy of a financial incentive. In addition, the authors note that the rewards were consistently relatively small and that the evidence suggests that small rewards will not be effective in changing physician behavior with respect to preventive care services. Finally, based on the evidence the authors recommend that desired behaviors must be specific and easy to track and that complex rules for determining rewards are less effective. *Van Herck et al. (2010)* found that in general, process measures showed more improvement than intermediate outcomes. In turn, intermediate outcomes showed more improvement than final outcomes. In addition, larger effect sizes were found for measures in which there was more room for improvement. Also, programs in which providers were involved in the selection/definition of measures and targets and in which there was direct and extensive communication towards providers appear to have been more successful (effect sizes beyond 10 percent) than programs that did not; provider awareness has been furthered as important success factors. Furthermore, risk adjustment and exception reporting seem to contribute to positive findings, as reported in several studies from the UK. Programs that adopted relative performance targets were generally less effective (eight studies) than programs using absolute targets or a sliding scale, although the relationship is not straightforward. No clear relationship was found between payment size and P4P effectiveness, but programs based on new money seem to have generated more positive effects than programs that relied on a reallocation of existing funds (e.g., withholding a certain portion of providers' base payments). One RCT (not included in the review but cited in the discussion section) found no difference in effect between quarterly payments and annual payments. Finally, programs that targeted individual physicians or small teams were often (but not always) more effective than programs targeting larger provider groups or hospitals.

CHAPTER 6

Performance profiling in primary care: does the choice of statistical model matter?

With René van Vliet

Medical Decision Making, 2013, DOI: 10.1177/0272989X13498825.

ABSTRACT

Background: Performance profiling is increasingly being used to generate input for improvement efforts in health care. For these to be successful, profiles must reflect true (differences in) provider performance, requiring an appropriate statistical model. Sophisticated models are available to account for the specific features of performance data, but they may be difficult to use, maintain, and explain to providers.
Objective: To assess the influence of the choice of statistical model on the performance profiles of primary care providers for various measures of resource use and quality of care.
Data source: Administrative data (2006-2008) on 2.8 million members of a Dutch health insurer registered with one of 4,396 general practitioners.
Methods: Profiles are constructed for six quality measures and five resource use measures, controlling for differences in casemix. Models include ordinary least squares, generalized linear models, and multilevel models. Separately for each model, providers are ranked on z-scores and classified as outlier if belonging to the 10 percent with the worst or best performance. Impact of statistical model is evaluated using the weighted κ for rankings overall, percentage agreement on outlier designation, and changes in rankings over time.
Results: Agreement among models was typically high overall ($\kappa>0.85$). Agreement on outlier designation was more variable and often below 80 percent, especially for high outliers. Rankings were more similar for process measures than for outcomes and expenses. Agreement among annual rankings per model was low for all models.
Conclusions: Although differences among models were relatively small, in each year the choice of statistical model did affect the rankings. In addition, judging from the fluctuations in model-specific rankings over time, most measures appear to be driven largely by random chance, regardless of the model that is used. Profilers should pay careful attention to both the choice of the statistical model and the performance measures.

6.1 INTRODUCTION

Purchasers and other actors in health care are increasingly interested in comparative information on the performance of healthcare providers. Variation in resource use and quality of care is well-documented, and in many countries, purchasers increasingly use specific measurement approaches to gain insight in providers' relative performance. The data derived from these measurements are often summarized in performance profiles, which may contain information on various aspects of providers' performance and can be used in various ways to spur improvement. For example, they may be used to provide feedback to providers (Van der Veer et al., 2010), to allocate P4P payments, and to steer consumers to high-performing providers via public reporting (Fung et al., 2008) and/or creating selective and tiered provider networks (Brennan et al., 2008).

Evidently, profiling is only useful for these purposes if profiles reflect true provider performance. Random chance and differences in casemix may explain large portions of observed performance variation and can obscure the "signal" of providers' true performance (Adams et al., 2010a; Friedberg & Damberg, 2012). Therefore, if they are to produce useful input for improvement efforts, profiles must take these factors into account. This is true especially for resource use and (clinical) outcome measures (e.g., blood sugar levels of diabetes patients, hospital readmissions) because they are particularly sensitive to random chance and relevant patient characteristics like age and disease severity. To mitigate the role of random variation, measures should only be used when there is sufficient between-provider variation and when a sufficient number of patients can be sampled. To mitigate incentives for risk selection and to ensure fair comparisons, adequate risk adjustment must be applied (Tucker et al., 1996; Rosen et al., 2001; Pope & Kautter, 2007; Ash & Ellis, 2012; Chang et al., 2012; Chen et al., 2012).

In profiling, comparing providers' observed performance to their expected performance (based on their casemix) has become standard (Ash & Ellis, 2012). In practice, purchasers typically calculate expected performance using model-derived (patient-level) predictions. Therefore, in addition to accurate data on patient characteristics, risk adjustment requires an appropriate statistical model, the choice of which will depend on the features of the data (Iezzoni, 2003) such as the type (binary, count, continuous) and the shape of the distribution (e.g., roughly normal or highly skewed). In practice, however, other considerations will likely play a role in this choice as well. Instead of relying on expensive external expertise, purchasers often perform these analyses themselves (typically on an annual basis) and will therefore prefer models that are easy to use and maintain. In addition, for risk adjustment to fulfill its purpose, it is important that providers whose performance is being profiled understand and support the method. If not, even when differences in casemix are appropriately taken into account, providers may still view the risk-adjustment method as a "black box" and be suspicious of its validity (Christianson et al., 2008), which could undermine the entire profiling system. There-

fore, where possible, purchasers would opt for keeping the risk adjustment method simple. An often used method that can easily be applied to many types of performance data is ordinary least squares (OLS). However, performance data often have specific features rendering OLS unsuitable. More sophisticated models, though more difficult to explain and maintain, will usually fit these data better. Nonetheless, despite often being the less suitable method, OLS (at the patient-level) could generate similar *profiling* results (at the provider-level).

In this paper, we use administrative data from a large Dutch health insurer to compare statistical models that can be used for analyzing and risk-adjusting the performance of Dutch general practitioners (GPs) and health centers (HCs) on several measures of quality and resource use. The insurer has been implementing several performance profiling programs in the Dutch primary care sector and, for the reasons mentioned above, wanted insight in the extent to which simple methods (OLS) yield similar profiling results compared to more appropriate sophisticated methods. Previous studies have looked at the impact on profiling results of varying the risk-adjustment methodology (Mukamel et al., 2008; Mukamel & Brower, 1998; Huang et al., 2005b; Thomas et al., 2004; Rosen et al., 2003; Kang & Hong, 2011; Delong et al., 1997; Iezzoni et al. 1996), treatment of patients with extreme values (Thomas & Ward, 2006), definition of performance index (Kang & Hong, 2011; Thomas, 2006; Rosen et al., 2002; Metfessel & Greene, 2012), and method for categorizing providers in different performance categories (Austin et al., 2004; Adams et al., 2010b). This study focuses on the impact of the *statistical model*, holding constant the set of risk-adjusters and other factors. Although there have been some other studies that assessed the influence of the statistical model on performance profiling results, these studies only included a few model types in their comparisons (e.g., two or three). In addition, each of these studies evaluated the impact for only one performance measure: satisfaction with asthma care (Huang et al., 2005b), managed care pharmacy expenses (Cowen & Strawderman, 2002), or in-hospital mortality for patients undergoing coronary artery bypass grafting surgery (Delong et al., 1997; Glance et al., 2006). Our study compares more statistical models and assesses their impact for eleven performance measures applicable to three different populations. In addition, by comparing annual provider rankings over three adjacent years, we also provide insight in the influence of the statistical model on the stability of profiling results over time, which has not been done in previous work. Large fluctuations would indicate that the risk-adjusted measures are mainly driven by random chance instead of true provider performance.

6.2 METHODS

6.2.1 Study setting and data

In the Dutch healthcare sector for curative care, private, risk-bearing insurers are expected to act as prudent purchasers of care on behalf of their members. To adequately fulfill this

role, insurers can use several managed-care instruments, including selective contracting, financial incentives, and performance feedback to providers. Each of these instruments requires an adequate profiling system. In this study, performance profiles are constructed for GPs and HCs using administrative data for the years 2006-2008 obtained from a Dutch insurer. For each year, data on about 2.8 million members are available, including sociodemographic characteristics and proxies for health status. In the Netherlands, these data are routinely available in health insurers' files at no additional cost. For each member, it is known with which GP he/she was registered in a particular year. In the Netherlands, GPs have fixed patient panels and act as gatekeeper to hospital care. Thus, GPs can influence the amount and type of hospital care their patients use. A small but increasing number of GPs hold practice in an HC, which is an entity in which multiple GPs (typically four or five) and other primary care providers (e.g., physiotherapists, dietitians) provide and coordinate care, usually from the same building. In our data, a GP may or may not be affiliated to an HC. Thus, for each member our data provides a link to his/her own GP, and, if this GP is affiliated to an HC, also a link to this HC. Approximately 10 percent of the GPs in our dataset were affiliated to a HC, so the vast majority of members did not receive primary care from GPs working in an HC.

6.2.2 Dependent variables (performance measures)

Using the administrative data, we constructed three types of performance measures: expenses (three measures), utilization of hospital care (two), and clinical quality (six). The expenses measures are GP expenses (generated through visits and diagnostic tests/examinations), prescription medication expenses, and total expenses (the sum of GP, medication, and hospital expenses). Regarding *utilization*, the number of inpatient admissions and outpatient visits are available. Both are indicated by "diagnosis treatment combinations" (DTCs), which were implemented to facilitate contracting for hospital services (Van de Ven & Schut, 2009). A DTC is a predefined "care product", selected by the medical specialist based on the patient's condition and representing all hospital procedures/services related to treating a patient with a specific diagnosis within a fixed period. It is similar to a DRG used by Medicare in the US, except that DTCs are more broadly defined and also include the payment for medical specialists. Finally, providers are compared on clinical process and outcome *quality* for patients with diabetes mellitus (DM) and patients with chronic obstructive pulmonary disease (COPD). For DM, the percentage of patients on statins and the number of DM-related hospital admissions were available. For COPD, three process measures were defined: the percentage of patients using bronchodilators, the percentage of patients using corticosteroids, and the percentage of patients receiving physiotherapy. The number of COPD-related hospital admissions was used as outcome. The result is three types of dependent variables: continuous (expenses, lower is better), count (utilization, lower is better), and binary (clinical processes, higher is better).

A small number of extreme outlier members (109) were excluded to minimize distorting effects on coefficients and profiling results and to increase the chance of algorithm convergence for the more complex statistical models. In addition, because the distributions of medication and total expenses are highly skewed and we could not rule out the possibility that several extremely high values (between 50-100 patients per year) were erroneous (e.g., miscoded), based on a visual inspection these variables were top-coded at €25,000 and €125,000, respectively. Dependent variables for members enrolled for less than a year were annualized and weighted based on months of enrollment. Providers were only included if they had at least 100 patients in each year for the non-disease-specific variables, while for the disease-specific variables they had to have at least 30 patients (we chose these thresholds because they are common in practice and in the literature). After applying these restrictions, 4396, 628, and 517 GPs were included for the non-disease-specific, DM, and COPD measures, respectively. For HCs, these numbers are 120, 45, and 35.

6.2.3 Independent variables (risk adjusters)

The models adjust for various patient characteristics, all derived from the administrative data (Table 6.1). In addition to age and sex, we included five indicators of socioeconomic status, three of which were measured at the member's ZIP-code level. For example, the three categories of educational level (low, medium, high) relate to the average educational level of people living in the member's ZIP-code area. The variable ethnicity is based on the percentage of persons in the ZIP-code area of whom at least one parent was born in Turkey, Africa, Latin-America, or Asia (excluding Japan). This variable was included because different ethnic groups may exhibit different patterns of utilization (Van der Lucht & Verweij, 2010) and may not be equally compliant with recommended treatment (Peeters et al., 2011; Bailey & Kodack, 2011). The variable urbanization is based on the number of adjacent addresses per square kilometer for 2006, and on the number of inhabitants in the member's town/city of residence for 2007-8. We also included two proxies for health status: pharmacy-based cost groups (PCGs) and diagnosis cost groups (DCGs). Both proxies have been developed in the context of the Dutch risk-equalization scheme (which is used to calculate risk-adjusted capitation payments for health insurers, Van de Ven et al., 2004; Prinsze & Van Vliet, 2007)

TABLE 6.1 Included risk adjusters

- Age-sex interactions (38 categories)
- Yes/no living in a deprived area
- Monthly income (ZIP-code level, 10 categories)
- Educational level (ZIP-code level, 3 categories)
- Ethnicity (ZIP-code level, 6 categories in 2006, 5 in 2007-2008)
- Degree of urbanization (5 categories in 2006, 8 in 2007-2008)
- Yes/no died in year of interest
- Pharmacy-based cost groups (20 categories/comorbidies)
- Diagnosis cost groups (13 categories)

and are designed to identify patients with chronic conditions. PCGs are based on prior (outpatient) use of medication. A member is assigned to a certain PCG if prescribed at least 181 defined daily doses of a particular disease-specific medication in the prior year. For example, if a member was prescribed at least 181 defined daily doses of insulin in year t, he/she will be classified in the PCG for diabetes type 1 in year t+1. Our data distinguishes twenty PCGs, all of which relate to a certain chronic condition (e.g., diabetes, heart disease, rheumatoid arthritis, cancer, epilepsy). Members were identified as having DM if classified in the PCG for DM. COPD patients were defined in a similar way among members 45 years of age or older. DCGs are based on the diagnoses of hospitalizations in the prior year. About 500 DTCs for which high future expenses are likely were clustered on homogeneity of expenses, resulting in thirteen DCGs. If a member was admitted to the hospital and classified in one of these DTCs in year t, this member will be classified in the associated DCG in t+1.

All risk-adjusters were carefully developed for the purpose of explaining cost variation at the individual member-level and are therefore appropriate for expenses measures (Van Kleef & Van Vliet, 2011). Because the utilization measures are closely related to expenses, the risk-adjusters are probably also relevant for these measures. This was confirmed when we ran the models; all risk-adjusters were typically significantly associated with the dependent variable. For the process measures, however, this was not always the case, especially regarding the DCGs. But because the pattern of (lack of) significant associations with the dependent variables was not consistent across models and over time, we chose to include all variables in all models to ensure comparability. As a result, all models use the same risk-adjusters.

6.2.4 Model selection

Expenses and utilization data often have specific features that complicate modeling of these data, including a large fraction of people without any consumption (i.e., a large zero mass), heteroskedasticity (i.e., non-constant error variance), and skewed distributions. As modeling by OLS may lead to imprecise estimates, more robust methods have been proposed that recognize the distribution of the data and are less sensitive to the right tail. Another issue is that many methods assume independent observations. Yet it is likely that in our case, the data are not generated independently but in groups because patients with specific characteristics tend to choose and remain with physicians with specific characteristics (Greenfield et al., 2002). Our procedure of selecting models that can accommodate these features comprised two steps. First, we consulted key references in the field of health econometrics and profiling (Jones, 2000; Manning & Mullahy, 2001; Iezzoni, 2003; Deb et al., 2011; Mihaylova et al., 2011) to create a list of relevant *types* of models:

- OLS was applied to all performance measures, including the binary variables. A linear probability model is justified here because the individual expected probabilities are aggregated to the provider level typically yielding an expected probability between 0 and 1.

- Generalized linear models (GLM) can take into account heteroskedasticity while retaining the original scale, thus making retransformation methods superfluous (McCullagh & Nelder, 1989; Iezzoni, 2003). They accommodate skewness via variance-weighting and require specification of a distribution and a nonlinear link-function of the dependent variable that can be modeled (by maximum-likelihood) as a linear function of independent variables. Using the GENMOD procedure in SAS 9.2, we tested several distributions: normal and gamma for expenses; normal, gamma, Poisson and negative binomial (negbin) for counts; and binomial for binary variables.
- Two-part models deal with dependent variables with many zeroes by splitting consumption in two parts: the probability of any consumption and the level of consumption conditional on having any (Jones, 2000). Two-part models are estimated for medication expenses (30 percent zeroes), admissions (92-95 percent), and outpatient visits (75 percent). Parameters are estimated separately for each part (using the same covariates), and the prediction is obtained by multiplying the estimated probability from a probit or logit model by the conditional outcome.
- Multilevel models (MLM, also known as random-effects models) explicitly model the the hierarchical structure of the data, thereby recognizing that nested observations may be correlated. When this is the case, MLMs produce estimates that are more robust to small sample size and more precise as predictions (Goldstein & Spiegelhalter, 1996; Delong et al., 1997; Iezzoni, 2003; Huang et al., 2005b). Intervals around provider-specific performance estimates will also be wider, reflecting the uncertainty arising from both variation of patients within providers and variation between providers (Rice & Jones, 1997; Iezzoni, 2003). Using the GLIMMIX procedure in SAS, we employed two-level models with a random provider intercept with mean zero and constant variance, adjusting for the fixed effects of patients' risk characteristics. We also considered the NLMIXED procedure, but chose GLIMMIX because NLMIXED tends to have problems in achieving an accurate integral approximation in the log-likelihood in models with a relatively large number of random effects (Zhang et al., 2011). All MLMs were estimated by maximum pseudo-likelihood. We also tried Laplace approximation and adaptive quadrature, but as these techniques often resulted in computational (convergence) problems, we decided not to use them further.

We did not include models in which providers are modeled as fixed effects. The reason is that this would often result in unworkable models given the large number of providers. Yet we acknowledge the controversy between fixed- and random-effects models and the fact that both types of models compute provider effects in different ways (DeLong et al., 1997; Cowen & Strawderman, 2002; Racz & Sedransk, 2010; Jones & Spiegelhalter, 2011). We ran OLS models with provider fixed effects for all measures for HCs and for the disease-specific measures for GPs. Results were nearly identical to models without these effects, as also found by others (Cowen & Strawderman, 2002; Glance et al., 2006). In the second step

of our selection procedure, for each of the model types we created a final set of model *specifications* with a comparable fit. Appropriate specifications (i.e., well-fitting links and distributions) were determined using the following criteria and tests:

- Percent explained variance (R^2): 1-[var(residuals)/var(dependent variable)];
- Mean absolute deviation (MAD): the average of the absolute value of the residuals;
- Bayesian Information Criterion: (-2*ln[likelihood])+(number of parameters*ln[n]);
- Pregibon's (1980) link test and the modified Hosmer-Lemeshow test;
- Calibration: extent to which the mean expected value approximates the mean observed value. If the mean expected value differs from the mean observed value, the model requires recalibration, which is achieved by multiplying each member's expected value by a factor obtained from dividing the overall mean observed value by the overall mean expected value. Calibration was also assessed using an OLS regression with the observed outcome as depended variable and the expected outcome as independent variable. If this yields an intercept of 0 and a slope of 1, recalibration is not necessary;
- Adequate convergence of algorithm in all years.

We only included converging models with a satisfactory fit in all years. As a result, models with a good fit in a particular year may still have been excluded. We followed this approach for two reasons: (1) although excluded model specifications sometimes performed better than some included specifications, differences were small; (2) having the same models in all years enables calculations on the stability of profiling results over time.

6.2.5 Model comparison

We calculated agreement among models on provider rankings based on z-scores. The z-score has widely been used in profiling and is preferred over other metrics (Berlowitz et al., 1998; Rosen et al., 2001; Iezzoni, 2003; Thomas et al., 2004). Using the measure-specific patient-level observed and expected values, we calculated the mean observed and mean expected performance level for each provider in each year by summing the observed and expected patient-level values and dividing by the number of patients per provider. The provider-specific z-score is then obtained by dividing the difference between these two means by the standard error of this difference.

Agreement is measured separately for each measure using the weighted κ statistic, which measures agreement between rankings beyond agreement due to chance (Landis & Koch, 1977). For each model, we ranked providers on z-scores and recoded the ranking into twenty equally-sized groups. Next, for each pair of models, we calculated the weighted κ by comparing both rankings. Finally, for each model we calculated the average agreement with all the other models using the weighted κ obtained from the pairwise comparisons with the other models. Models are also compared on the extent to which they agree on outlier designation. A provider is considered an outlier if belonging to the 10 percent providers with the worst performance or best performance. Average percentage agreement was calculated

for each model for both low and high outliers. Finally, models are compared on stability of results over time using the average of the agreement between the rankings of 2006 and 2007, of 2006 and 2008, and of 2007 and 2008.

Agreement statistics are calculated separately for HCs, GPs in an HC, and GPs not in an HC. During 2006-2008, HCs participated in a P4P-program in which most of the measures used in this paper were included. Variation in profiling results over time for (GPs in) HCs could be a reflection of this program having an effect. In that case, results will be more stable for GPs not in an HC.

6.3 RESULTS

Table 6.2 provides descriptive statistics for members and providers. Among the 2.8 million members, the fraction in a PCG increased from 16.2 percent in 2006 to 17.5 percent in 2008, indicating an increase in the prevalence of chronic conditions. Seven percent had at least one hospital admission and 27 percent at least one outpatient visit. GP expenses, medication expenses, and total expenses average to approximately €125, €300, and €1,500 per year, respectively. Process measures for DM and COPD patients (3 percent and 2.5 percent of all members, respectively) remained rather stable, although there were small increases for physiotherapy and statins. About 5 percent of these patients were admitted to a hospital for condition-related reasons. Three percent of the members were not registered with a GP, which are mainly people residing in nursing homes; members whose GP held practice in a HC increased from 8.6 percent in 2006 to 12.7 percent in 2008 (data not shown).

Table 6.3 shows the included models for each performance measure as well as some fit statistics for 2008 (for results for other years, see Appendix 6.1; the magnitude of values sometimes differ across years, but patterns are similar). OLS is often outperformed by several other models, though differences are generally quite small. Regarding the binary measures, OLS yields the lowest R^2 and highest MAD while the MLMs yield the highest R^2 and lowest MAD. A similar pattern can be observed for GP expenses, while for other expenses, alternatives to OLS do not add much. Regarding the count variables, several models yield lower R^2 and higher MAD-values than OLS, but there is always at least one model performing better on both statistics. Two-part models are among the models with the lowest R^2, and for admissions and visits also have the highest MAD. Finally, several models needed to be recalibrated. Models for which this was most necessary typically had the worst fit (e.g., lognormal for COPD-related admissions and gamma-power for medication expenses).

The R^2-values also provide insight in the importance of risk-adjustment. As expected, the models explained a relatively large fraction (22-38 percent) of total member-level variation in expenses. This is also true for outpatient visits (36 percent), whereas for hospital admissions models explain only about 7-12 percent of the variation. As risk-adjustment is

TABLE 6.2 Descriptive statistics of the study sample, by year

	2006	2007	2008
All members – independent variables	*N=2,809,250*	*N=2,802,632*	*N=2,808,838*
Age (mean [SD])	40.1 [23.2]	40.2 [23.2]	40.4 [23.3]
Male (%)	50.5	50.4	50.3
Living in a deprived area (%)	6.7	6.5	6.5
Monthly income (mean [SD]) [a]	5.3 [2.9]	5.3 [2.9]	5.3 [2.8]
Educational level (mean [SD]) [b]	2.0 [0.8]	2.0 [0.8]	2.0 [0.8]
Ethnicity (mean [SD]) [c]	2.1 [1.5]	1.3 [0.8]	1.3 [0.8]
Urbanization (mean [SD]) [d]	2.9 [1.4]	4.7 [2.2]	4.7 [2.1]
Died (%)	0.9	0.8	0.8
In a PCG (%)	16.2	16.9	17.5
In ≥2 PCGs (%)	4.5	4.8	5.1
In ≥3 PCGs (%)	1.2	1.3	1.4
In a DCG (%)	2.6	2.3	2.4
All members – dependent variables	*N=2,809,250*	*N=2,802,632*	*N=2,808,838*
Inpatient admissions (mean [SD])	0.10 [0.46]	0.11 [0.49]	0.10 [0.46]
No inpatient admission (%)	92.7	92.5	92.8
Outpatient visits (mean [SD])	0.52 [1.43]	0.53 [1.52]	0.53 [1.21]
No outpatient visit (%)	72.4	73.0	74.1
GP expenses (mean [SD])	119 [112]	128 [122]	127 [119]
No GP expenses (%)	2.1	2.1	2.2
Medication expenses (mean [SD])	275 [908]	310 [1024]	302 [1048]
No medication expenses (%)	31.9	31.1	29.2
Total expenses (mean [SD])	1,476 [5306]	1,531 [5359]	1,485 [4879]
No expenses (%)	1.6	1.6	1.6
Members with diabetes – dependent variables	*N=86,208*	*N=88,536*	*N=89,320*
On statins (%)	59.4	63.0	63.6
Inpatient admissions (mean [SD])	0.08 [0.38]	0.08 [0.39]	0.07 [0.36]
No inpatient admission (%)	94.0	94.1	94.9
Members with COPD – dependent variables	*N=65,315*	*N=68,927*	*N=69,892*
Receiving physiotherapy (%)	4.17	4.76	5.55
On corticosteroids (%)	33.4	32.6	32.6
On bronchodilators (%)	81.5	80.5	80.2
Inpatient admissions (mean [SD])	0.07 [0.36]	0.08 [0.37]	0.07 [0.36]
No inpatient admission (%)	94.6	94.3	94.6
General practitioners	*N=7,471*	*N=5,447*	*N=5,538*
≥100 patients, all years (n [mean sample size])	4,396 [529]	4,396 [533]	4,396 [529]
≥100 patients, all years + in a HC (n [mean sample size])	355 [688]	355 [692]	355 [653]
≥30 DM patients, all years (n [mean sample size])	628 [70]	628 [71]	628 [72]
≥30 DM patients, all years + in a HC (n [mean sample size])	79 [68]	79 [70]	79 [72]
≥30 COPD patients, all years (n [mean sample size])	517 [58]	517 [60]	517 [61]
≥30 COPD patients, all years + in a HC (n [mean sample size])	66 [55]	66 [59]	66 [59]
Health centers	*N=142*	*N=179*	*N=186*
≥100 patients in all years (n [mean sample size])	120 [1,791]	120 [1,874]	120 [1,841]
≥30 DM patients in all years (n [mean sample size])	45 [131]	45 [141]	45 [141]
≥30 COPD patients in all years (n [mean sample size])	35 [117]	35 [130]	35 [130]

a. Ten categories (1 = lowest income decile, 10 = highest income decile).

b. Three categories (1 = low, 3 = high).

c. Six categories in 2006 (1-6) and 5 categories in 2007-8 (0-4). A high score corresponds to a high percentage of non-Western immigrants living in the member's ZIP-code area.

d. Five categories in 2006 (1-5), 8 categories in 2007-8 (1-8). A high score corresponds to a low level of urbanization of the member's municipality.

TABLE 6.3 Selected fit statistics of included models, by performance measure, 2008

Measure (population)	Model	R^2 [d]	MAD	Calibration [e]
Yes/no physiotherapy (COPD)	OLS	.037	.101	$Y=\hat{Y}$
	GLM binomial-probit	.039	.100	$Y=1.000\hat{Y}$
	MLM normal-id (HCs)	.042	.099	$Y=\hat{Y}$
	MLM normal-id (GPs)	.063	.099	$Y=\hat{Y}$
Yes/no corticosteroids (COPD)	OLS	.061	.411	$Y=\hat{Y}$
	GLM binomial-logit	.061	.411	$Y=\hat{Y}$
	MLM normal-id (GPs)	.085	.411	$Y=\hat{Y}$
Yes/no bronchodilators (COPD)	OLS	.026	.312	$Y=\hat{Y}$
	GLM binomial-logit	.028	.310	$Y=\hat{Y}$
	MLM normal-id (GPs)	.048	.293	$Y=\hat{Y}$
Yes/no statins (diabetes)	OLS	.701	.136	$Y=\hat{Y}$
	GLM binomial-logit	.707	.136	$Y=\hat{Y}$
	MLM normal-id (GPs)	.713	.134	$Y=\hat{Y}$
No. of inpatient admissions (COPD)	OLS	.114	.124	$Y=\hat{Y}$
	GLM normal-log	.109	.126	$Y=0.946\hat{Y}$
	GLM Poisson-power	.113	.123	$Y=\hat{Y}$
	GLM negbin-power	.111	.123	$Y=0.960\hat{Y}$
	GLM gamma-log	.115	.124	$Y=1.000\hat{Y}$
	MLM normal-power (HCs)	.118	.123	$Y=1.024\hat{Y}$
	2-part logit-OLS	.105	.123	$Y=\hat{Y}$
	2-part logit-normal power	.105	.123	$Y=1.000\hat{Y}$
	2-part logit-Poisson power	.105	.123	$Y=1.000\hat{Y}$
No. of inpatient admissions (diabetes)	OLS	.070	.125	$Y=\hat{Y}$
	GLM normal-log	.077	.125	$Y=0.992\hat{Y}$
	GLM Poisson-power	.072	.124	$Y=\hat{Y}$
	GLM negbin-power	.071	.124	$Y=0.971\hat{Y}$
	GLM gamma-log	.071	.124	$Y=1.001\hat{Y}$
	MLM normal-power (HCs)	.073	.124	$Y=1.013\hat{Y}$
	2-part logit-OLS	.069	.124	$Y=\hat{Y}$
	2-part logit-normal log	.069	.124	$Y=1.001\hat{Y}$
	2-part logit-Poisson power	.068	.124	$Y=1.000\hat{Y}$
No. of inpatient admissions (all members)	OLS	.109	.166	$Y=\hat{Y}$
	GLM Poisson-power	.104	.166	$Y=\hat{Y}$
	GLM negbin-power	.101	.166	$Y=0.965\hat{Y}$
	GLM gamma-power	.108	.166	$Y=1.000\hat{Y}$
	MLM normal-id (GPs)	.109	.166	$Y=\hat{Y}$
	2-part logit-OLS	.103	.167	$Y=\hat{Y}$
	2-part logit-normal power	.103	.167	$Y=1.000\hat{Y}$
	2-part logit-Poisson power	.102	.167	$Y=1.000\hat{Y}$
No. of outpatient visits (all members)	OLS	.366	.477	$Y=\hat{Y}$
	GLM Poisson-power	.365	.477	$Y=\hat{Y}$
	GLM negbin-id	.363	.476	$Y=0.998\hat{Y}$
	GLM gamma-id	.363	.476	$Y=1.000\hat{Y}$
	MLM normal-id (GPs)	.369	.474	$Y=\hat{Y}$
	2-part logit-OLS	.365	.478	$Y=\hat{Y}$
	2-part logit-normal power	.367	.479	$Y=1.002\hat{Y}$
	2-part logit-Poisson power	.366	.478	$Y=1.000\hat{Y}$

TABLE 6.3 (continued)

Measure (population)	Model	R^2 [d]	MAD	Calibration [e]
GP expenses (all members) [a]	OLS	.288	52.326	$Y=\hat{Y}$
	GLM normal-power	.288	52.439	$Y=0.999\hat{Y}$
	MLM normal-id (HCs)	.290	52.121	$Y=\hat{Y}$
	MLM normal-id (GPs)	.318	50.701	$Y=\hat{Y}$
	MLM gamma-power (GPs)	.319	50.442	$Y=0.993\hat{Y}$
Medication expenses (all members) [b]	OLS	.375	222.676	$Y=\hat{Y}$
	GLM gamma-power	.335	224.940	$Y=0.929\hat{Y}$
	MLM normal-id (GPs)	.376	222.966	$Y=\hat{Y}$
	2-part probit-OLS	.375	222.165	$Y=1.000\hat{Y}$
	2-part probit-normal power	.362	237.793	$Y=0.957\hat{Y}$
	2-part probit-gamma power	.350	228.838	$Y=0.958\hat{Y}$
Total expenses (all members) [c]	OLS	.226	1452.19	$Y=\hat{Y}$
	MLM normal-id (HCs)	.226	1449.62	$Y=\hat{Y}$
	MLM normal-id (GPs)	.226	1450.54	$Y=\hat{Y}$

Note: COPD = chronic obstructive pulmonary disease, GPs = general practitioners, HCs = health centers, id = identity, MAD = mean absolute deviation, MLM = multilevel model, OLS = ordinary least squares.

a. Expenses generated by GPs through office visits, home visits, and (diagnostic) tests.

b. Expenses related to the use of prescription medication, regardless of prescriber.

c. Sum of GP expenses, medication expenses, and inpatient expenses generated by medical specialists.

d. Percent explained variation at the individual member level.

e. Extent to which the mean of expected values ($\hat{Y}$) approximates the mean of observed values (Y).

no doubt important for these measures, the low R^2-values are probably a result of a combination of inadequate risk-adjustment and the fact that hospital admissions are relatively rare. Even less variation is explained in three of the four clinical process measures. The very high R^2 for statins can be explained by very strong associations with some PCGs (e.g., heart disease).

6.3.1 Agreement among models per year

Table 6.4 presents average levels of agreement for 2008 (figures for 2006-7 are similar for most measures, see Appendices 6.2-6.4; exceptions are higher agreement for HCs for physiotherapy, lower agreement overall but higher agreement on outliers for disease-related admissions, and lower agreement for HCs for outpatient visits in 2006-7 compared to 2008). Agreement on overall rankings is high with κ often above .90, typically above .85, and never below .74. Agreement on outlier designation is more variable but still quite high, and tends to be higher for processes than for outcomes and expenses, for which agreement is often below 80 percent. Overall, models tend to agree better on designation of low outliers than of high outliers, although there are exceptions (e.g., GP expenses for HCs). Models agree somewhat better for GPs than for HCs, especially for disease-related admissions and expenses. Finally, models with similar fit statistics may agree poorly on profiling results. For GP expenses, for example, the normal-power model agrees worse with the other model(s) than OLS.

TABLE 6.4 Average agreement among models (2008) and per model among years (2006-2008), by selected performance measure

		PER YEAR AMONG MODELS (2008)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]
Yes/no bronchodilators	OLS	.96	1.0	1.0	.97	.98	.97	.96	.86	.93	.52	.50	.42	.49	.54	.34	.49	.67	.48
	GLM binomial-logit	.96	1.0	1.0	.96	.98	.96	.95	.86	.86	.52	.50	.58	.49	.55	.35	.49	.57	.57
	MLM normal-id	-	-	-	.96	.98	.97	.96	.86	.93	-	-	-	.49	.54	.34	.49	.57	.52
Yes/no statins	OLS	.97	1.0	1.0	.98	.97	.96	.98	1.0	1.0	.20	.33	.40	.31	.36	.30	.27	.25	.29
	GLM binomial-logit	.97	1.0	1.0	.97	.95	.96	.97	1.0	1.0	.20	.33	.40	.31	.35	.29	.27	.25	.29
	MLM normal-id	-	-	-	.97	.97	.96	.97	1.0	1.0	-	-	-	.31	.34	.27	.27	.25	.29
No. of inpatient admissions (COPD)	OLS	.87	.85	.50	.92	.90	.85	.92	.93	.79	.24	.67	.10	.04	.18	.11	.15	.33	.29
	GLM normal-log	.84	.80	.60	.86	.89	.77	.86	.82	.61	.37	.42	.17	.04	.18	.12	.16	.24	.14
	GLM Poisson-power	.87	.85	.70	.92	.92	.85	.93	.93	.75	.26	.67	.17	.04	.17	.11	.16	.33	.19
	GLM negbin-power	.86	.75	.70	.90	.83	.77	.90	.82	.71	.27	.58	.10	.04	.16	.10	.16	.10	.19
	GLM gamma-log	.87	.85	.70	.92	.92	.86	.93	.93	.79	.25	.67	.17	.04	.18	.09	.16	.33	.29
	MLM normal-id	.82	.80	.60	-	-	-	-	-	-	.20	.42	.17	-	-	-	-	-	-
No. of inpatient admissions (diabetes)	OLS	.90	.84	.80	.92	.95	.88	.91	.97	.84	-.03	.07	.20	.04	.12	.12	.03	.08	.21
	GLM normal-log	.84	.84	.68	.83	.88	.78	.82	.88	.66	-.01	.07	.20	.04	.13	.12	.01	.08	.21
	GLM Poisson-power	.89	.80	.80	.92	.95	.87	.91	.97	.84	-.03	.07	.13	.03	.12	.10	.02	.08	.21
	GLM negbin-power	.87	.84	.72	.91	.90	.83	.89	.97	.75	-.05	.07	.07	.04	.14	.13	.00	.08	.13
	GLM gamma-log	.90	.84	.80	.93	.95	.88	.92	.97	.84	-.03	.07	.20	.04	.12	.11	.02	.08	.21
	MLM normal-id	.84	.80	.68	-	-	-	-	-	-	-.00	.07	.13	-	-	-	-	-	-
GP expenses (all members) [b]	OLS	.85	.75	.83	.92	.93	.88	.93	.92	.90	.51	.64	.36	.54	.64	.54	.51	.57	.44
	GLM normal-power	.74	.58	.88	.88	.88	.81	.89	.83	.84	.47	.44	.42	.53	.60	.51	.50	.54	.43
	MLM normal-id	.85	.75	.88	.92	.92	.87	.93	.92	.88	.52	.64	.42	.53	.63	.53	.51	.58	.42
	MLM gamma-power	-	-	-	.90	.91	.84	.92	.85	.84	-	-	-	.53	.64	.50	.51	.63	.46

TABLE 6.4 (continued)

		PER YEAR AMONG MODELS (2008)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]
Medication expenses (all members) [c]	OLS	.77	1.0	.67	.90	.92	.87	.91	.89	.86	.47	.78	.33	.47	.52	.49	.47	.57	.38
	GLM gamma-power	.77	1.0	.67	.82	.86	.76	.83	.81	.76	.55	.72	.36	.46	.50	.52	.47	.52	.39
	MLM normal-id	-	-	-	.90	.93	.87	.90	.89	.88	-	-	-	.48	.53	.50	.48	.58	.38

Note: '-' indicates that the relevant model did not converge or that the covariance parameter for the random provider intercept was not significant. COPD = chronic obstructive pulmonary disease, GP = general practitioner, id = identity, MLM = multilevel model, negbin = negative binomial, OLS = ordinary least squares.

a. Two-part models were excluded from these comparisons, for three reasons: expected values are very similar to those from the one-part models (the correlation coefficient was typically >.99), they do not appear to fit the data better, and calculation of standard errors is relatively complex.

b. Expenses generated by general practitioners through office visits, home visits, and (diagnostic) tests.

c. Expenses related to the use of prescription medication, regardless of prescriber.

d. Average pairwise agreement (weighted κ) on overall rankings on z-scores.

e. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the worst performance based on z-scores.

f. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the best performance based on z-scores.

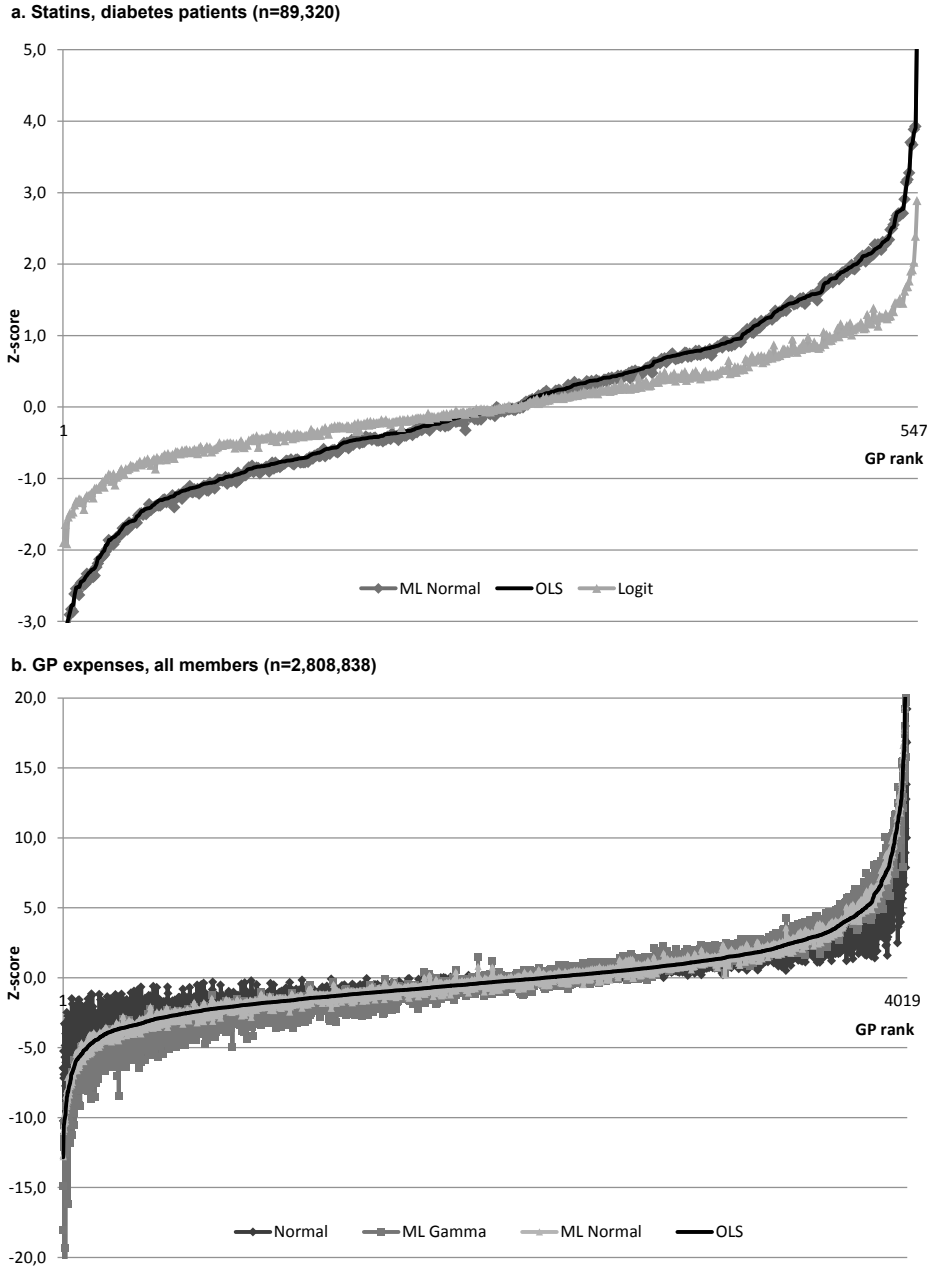


FIGURE 6.1 Distributions of z-scores for GPs (not in a health center) for two measures, 2008

Note: ML = Multilevel, OLS = ordinary least squares. The Figure displays GPs' z-scores produced by the different models for two different measures: the percentage of diabetes patients on statins (a process quality measure, three models, relevant for 547 GPs) and total GP expenses (a resource use measure, four models, relevant for 4,019 GPs). GPs are ranked on their (OLS-derived) z-scores from highest performance (rank 1) to lowest performance (rank 547 or 4,019, depending on the measure).

Figure 6.1 shows the distribution of z-scores for GPs for two measures: statins and GP expenses. Despite high agreement among models, differences may be large for individual providers. In addition, highly similar rankings do not preclude large differences, which become visible when an absolute threshold is used to discern providers. For example, for statins (panel a) a threshold of (-)2 results in lower agreement on outlier designation between OLS and logit than presented in Table 6.4. Plots like Figure 6.1 also visualize differences between measures. For example, assuming an absolute threshold, much more GPs will be classified as outlier for GP expenses (panel b) than for other measures for which z-scores have a much smaller range.

6.3.2 Agreement among years per model

Table 6.4 also shows limited agreement among annual rankings per model, ranging from absent (DM-related admissions) to fair (statins, COPD-related admissions) to moderate (all other measures) (see Appendix 6.2-6.4 for results for the other measures). Agreement on outlier designation is higher than agreement overall, but still fairly low. No model consistently produces more or less stable results than other models. Our hypothesis that results would be less stable for (GPs in) HCs than for GPs not in a HC is not confirmed: there appears to be no relationship between type of provider and the stability of profiling results. Limiting the analysis to providers with at least 100 patients for disease-specific variables and 1,000 patients for non-disease-specific variables did not change these results, although for some measures agreement increased by up to 15 percentage points.

6.4 DISCUSSION

This study has investigated the influence of the statistical model on performance profiling results for primary care providers. Our main goal was to determine whether different statistical methods selected based on statistical as well as relevant practical criteria (from a purchaser's perspective) generate different profiling results. Our results showed that profiling results are sensitive to the statistical model that is used and that the choice of model does indeed seem to matter, especially for clinical outcome measures and expenses.

However, differences were relatively small and the choice of model may not be as important as other choices such as the set of risk-adjusters, definition of performance index, and method for categorizing provider performance (Adams et al., 2010b; Austin et al., 2004; Delong et al., 1997; Huang et al., 2005b; Kang & Hong, 2011; Mukamel & Brower, 1998; Mukamel et al., 2008; Rosen et al., 2002; Rosen et al., 2003; Thomas et al., 2004; Thomas, 2006; Thomas & Ward, 2006). In addition, simple methods have important practical advantages. For example, OLS can be applied to all measures and data, a feature not shared by many other models that may work well for one year and fail to converge in the next. For purchas-

ers, these might be sufficient reasons to choose OLS (or a logit model for binary variables). Nonetheless, caution is clearly warranted. Agreement of 75-95 percent among rankings suggests that the models still relatively often classify providers in different performance categories, which, depending on the purpose for which the rankings are used (e.g., performance feedback, pay-for-performance, public reporting), may have far-reaching (financial) consequences for providers. In addition, compared to agreement overall, agreement on outlier designation was lower and more variable. For example, for non-disease specific measures and using 10 percent cutoffs for both tails to determine outliers status, even 5 percent disagreement means that the choice of model alone determines for 44 GPs (4,396 GPs * 20 percent * 5 percent) whether they will be classified as outlier or not, which may be hard to justify. Thus, for each individual measure selected for profiling, decision-makers are faced with a difficult tradeoff between identifying the best-fitting model each year (a cumbersome task), and simply using a well-known method that is easy to apply, maintain, and explain, but may also result in somewhat different provider classifications.

The first option can be time-consuming and expensive, especially if providers are profiled on many measures and if the modeling is outsourced to an external (commercial) party. In addition, it may result in mixed signals toward providers. For example, it may be confusing for providers if the purchaser tries to convince them about a new sophisticated method for a specific (type of) measure ("this specification is best suited to account for your specific patient mix for this measure!"), while in the year before they had just been convinced about the merits of another method for the same measure. A practical solution may be to use a simple and easy to apply, maintain, and explain model (e.g., OLS), and to compare the results with the results of a relatively simple 2-level MLM (e.g., assuming a normal distribution and identity link). In our data, such MLM specifications had little convergence problems and were often (but not always) among the models with the best fit statistics. In addition, as noted above, these models have advantages that may be appealing to providers, although these advantages may be difficult to explain (Friedberg & Damberg, 2012).

There were several other notable findings. First, models agreed more for processes than for outcomes and expenses. Yet we did observe differences for processes as well, and since even small differences can be important, the conclusion that the choice of model does not matter for processes cannot be justified. Second, models tended to agree more on designation of low outliers than of high outliers, especially for utilization and expenses. The explanation is that for high outliers, expected utilization and expenses are high compared to what is observed. Because generalized linear (mixed) models can better predict high expenses and utilization than OLS, agreement on high-outlier designation will be lower than on low-outlier designation. This is an important finding as most P4P-programs only reward high performance (see chapter 3). Third, agreement was higher for GPs than for HCs, especially for disease-related admissions and expenses. It thus seems that the choice of model matters

more for HCs than for GPs. Fourth, profiling results varied substantially over time. As this is unlikely to be a result of a specific intervention, it probably reflects random variation and low reliability (Berlowitz et al., 1998; Friedberg & Damberg, 2012). Results were most unstable for hospital admissions and total expenses, which is not surprising because these are more difficult to influence by providers than other types of measures such as processes (Hofer et al., 1999; Krein et al., 2002). Measures will be more reliable when sample size and the intraclass correlation (i.e., the proportion of total variation that can be attributed to between-provider variation) are large (Nyweide et al., 2009; Scholle et al., 2008; Adams et al., 2010a). Limiting the analysis to providers with more patients increased agreement, but much variation remained, implying relatively low between-provider variances. Results were most stable for GP expenses. Given the large range in z-scores (Figure 6.1b), this may be a particularly useful measure for profiling. However, in view of GPs' gatekeeping role, purchasers should then be cautious that GPs are not penalized for successfully keeping patients out of the hospital and/or not rewarded for unwarranted referrals.

This study has several limitations. First, we identified COPD patients using the PCG for chronic nonspecific respiratory conditions and the patients' age (≥45 years). As a result, we probably overestimated the number of COPD patients in our data. Second, outlier providers were arbitrarily defined. We chose a relative threshold for determining outlier status because this is common in practice. An absolute threshold, however, for example based on conventional levels for statistical testing, may yield different results. A provider will then be an outlier when the absolute value of the z-score is larger than 1.96 (P=0.05) or 2.58 (P=0.01). Using a large p-value mitigates the risk of incorrectly classifying outliers as average, but also increases the risk of classifying average providers as outliers. Absolute thresholds have the advantage that they are transparent and that they may provide stronger incentives for providers to improve (see chapter 2). Conversely, with a relative threshold, purchasers know exactly how many providers will be designated as outlier, which, when the profiling results are used for allocating incentive payments, provides budgetary certainty for the purchaser. The sensitivity of our results to the method of categorizing provider performance merits further study. Third, we analyzed all inpatient admissions and outpatient visits (i.e., grouped together), and not, for example, only ambulatory-care sensitive ones (AHRQ, 2001), which might have been the preferred approach in profiling primary care providers. Although separate analysis of different types of admissions/visits could have been valuable, admission/visit type could not be derived from our data. But even if ambulatory-care sensitive admissions could have been identified, they may well have been too rare to reliably model (Ash & Ellis, 2012).

A fourth limitation is that our set of risk-adjusters was based on administrative information that is routinely available in insurers' files. This information was not generated for profiling purposes but for explaining variation in costs for the purpose of calculating risk-adjusted capitation payments for insurers. Ideally, risk-adjustment for performance

profiling would use detailed (clinical) information from medical records and patient surveys, especially regarding clinical quality. However, collecting such data on a routine basis is expensive and we expect that in practice, insurers will often mainly use data already available in their files. In addition, although we would have preferred to use individual-level information on socioeconomic status, ZIP-code-level variables have been shown to discern broadly similar patterns compared to the corresponding individual-level variables (Zaslavsky & Epstein, 2005; Krieger, 2003). Finally, our results may not generalize to other settings and measures. We looked at a specific group of providers (Dutch GPs with fixed patient panels acting as gatekeepers) using administrative data from one insurer. Given the widespread use of performance profiling, future research should investigate whether our results are confirmed in other settings for reliable and commonly used performance measures.

In summary, although simple methods like OLS have advantages from a practical viewpoint, they may produce different profiling results compared to more suitable methods. Therefore, the choice of statistical model for performance profiling should be made with care, especially when results are used as input for 'high-stakes' improvement efforts. In addition, regardless of the model, performance comparisons should preferably be conducted over multiple time periods to gain insight in the extent to which the measures are driven by chance and thus if they are potentially suitable for profiling. Even for process measures, over which providers supposedly have much control, random chance may determine providers' relative positions to a large extent, which, depending on how and by whom the profiles are used, can have far-reaching (financial) consequences for providers.

APPENDICES

Appendix 6.1 Fit statistics of included models, by performance measure, 2007-2008

Measure	Model	R^2 [d]		MAD		Calibration [e]	
		2007	*2008*	*2007*	*2008*	*2007*	*2008*
Yes/no physiotherapy (COPD)	OLS	.026	.037	.089	.101	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM binomial probit	.027	.039	.089	.100	$Y=1.000\hat{Y}$	$Y=1.000\hat{Y}$
	MLM normal id (HCs)	.033	.042	.089	.099	$Y=\hat{Y}$	$Y=\hat{Y}$
	MLM normal id (GPs)	.059	.063	.089	.099	$Y=\hat{Y}$	$Y=\hat{Y}$
Yes/no corticosteroids (COPD)	OLS	.054	.061	.414	.411	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM binomial logit	.054	.061	.414	.411	$Y=\hat{Y}$	$Y=\hat{Y}$
	MLM normal id (GPs)	.081	.085	.415	.411	$Y=\hat{Y}$	$Y=\hat{Y}$
Yes/no bronchodilators (COPD)	OLS	.024	.026	.309	.312	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM binomial logit	.025	.028	.308	.310	$Y=\hat{Y}$	$Y=\hat{Y}$
	MLM normal id (GPs)	.051	.048	.292	.293	$Y=\hat{Y}$	$Y=\hat{Y}$
Yes/no statins (diabetes)	OLS	.615	.701	.183	.136	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM binomial logit	.618	.707	.178	.136	$Y=\hat{Y}$	$Y=\hat{Y}$
	MLM normal id (GPs)	.634	.713	.179	.134	$Y=\hat{Y}$	$Y=\hat{Y}$
No. of inpatient admissions (COPD)	OLS	.103	.114	.136	.124	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM normal log	.104	.109	.138	.126	$Y=0.951\hat{Y}$	$Y=0.946\hat{Y}$
	GLM Poisson power	.104	.113	.135	.123	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM negbin power	.103	.111	.135	.123	$Y=0.970\hat{Y}$	$Y=0.960\hat{Y}$
	GLM gamma log	.105	.115	.136	.124	$Y=1.001\hat{Y}$	$Y=1.000\hat{Y}$
	MLM normal power (HCs)	.109	.118	.136	.123	$Y=1.026\hat{Y}$	$Y=1.024\hat{Y}$
	2-part logit-OLS	.100	.105	.135	.123	$Y=\hat{Y}$	$Y=\hat{Y}$
	2-part logit-normal power	.100	.105	.135	.123	$Y=1.000\hat{Y}$	$Y=1.000\hat{Y}$
	2-part logit-Poisson power	.100	.105	.135	.123	$Y=1.000\hat{Y}$	$Y=1.000\hat{Y}$
No. of inpatient admissions (diabetes)	OLS	.070	.070	.141	.125	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM normal log	.070	.077	.143	.125	$Y=0.989\hat{Y}$	$Y=0.992\hat{Y}$
	GLM Poisson power	.069	.072	.141	.124	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM negbin power	.067	.071	.141	.124	$Y=0.969\hat{Y}$	$Y=0.971\hat{Y}$
	GLM gamma log	.070	.071	.141	.124	$Y=1.000\hat{Y}$	$Y=1.001\hat{Y}$
	MLM normal power (HCs)	.072	.073	.141	.124	$Y=1.017\hat{Y}$	$Y=1.013\hat{Y}$
	2-part logit-OLS	.065	.069	.140	.124	$Y=\hat{Y}$	$Y=\hat{Y}$
	2-part logit-normal log	.065	.069	.140	.124	$Y=1.003\hat{Y}$	$Y=1.001\hat{Y}$
	2-part logit-Poisson power	.065	.068	.140	.124	$Y=0.999\hat{Y}$	$Y=1.000\hat{Y}$
No. of inpatient admissions (all members)	OLS	.100	.109	.181	.166	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM Poisson power	.096	.104	.181	.166	$Y=\hat{Y}$	$Y=\hat{Y}$
	GLM negbin power	.093	.101	.181	.166	$Y=0.945\hat{Y}$	$Y=0.965\hat{Y}$
	GLM gamma power	.100	.108	.181	.166	$Y=0.999\hat{Y}$	$Y=1.000\hat{Y}$
	MLM normal id (GPs)	.101	.109	.181	.166	$Y=\hat{Y}$	$Y=\hat{Y}$
	2-part logit-OLS	.095	.103	.181	.167	$Y=\hat{Y}$	$Y=\hat{Y}$
	2-part logit-normal power	.095	.103	.182	.167	$Y=1.000\hat{Y}$	$Y=1.000\hat{Y}$
	2-part logit-Poisson power	.094	.102	.181	.167	$Y=1.000\hat{Y}$	$Y=1.000\hat{Y}$

Measure	Model	R^2 [d]		MAD		Calibration [e]	
		2007	*2008*	*2007*	*2008*	*2007*	*2008*
No. of outpatient visits (all members)	OLS	.333	.366	.669	.477	Y=Ŷ	Y=Ŷ
	GLM Poisson power	.324	.365	.670	.477	Y=Ŷ	Y=Ŷ
	GLM negbin id	.332	.363	.669	.476	Y=0.981Ŷ	Y=0.998Ŷ
	GLM gamma id	.332	.363	.668	.476	Y=1.000Ŷ	Y=1.000Ŷ
	MLM normal id (GPs)	.337	.369	.664	.474	Y=Ŷ	Y=Ŷ
	2-part logit-OLS	.332	.365	.670	.478	Y=Ŷ	Y=Ŷ
	2-part logit-normal power	.337	.367	.671	.479	Y=1.004Ŷ	Y=1.002Ŷ
	2-part logit-Poisson power	.331	.366	.670	.478	Y=1.000Ŷ	Y=1.000Ŷ
GP expenses (all members) [a]	OLS	.250	.288	55.604	52.326	Y=Ŷ	Y=Ŷ
	GLM normal power	.251	.288	55.710	52.439	Y=0.999Ŷ	Y=0.999Ŷ
	MLM normal id (HCs)	.251	.290	55.591	52.121	Y=Ŷ	Y=Ŷ
	MLM normal id (GPs)	.278	.318	54.139	50.701	Y=Ŷ	Y=Ŷ
	MLM gamma power (GPs)	.279	.319	54.005	50.442	Y=0.994Ŷ	Y=0.993Ŷ
Medication expenses (all members) [b]	OLS	.421	.375	221.376	222.676	Y=Ŷ	Y=Ŷ
	GLM gamma power	.367	.335	225.222	224.940	Y=0.916Ŷ	Y=0.929Ŷ
	MLM normal id (GPs)	.423	.376	221.786	222.966	Y=Ŷ	Y=Ŷ
	2-part probit-OLS	.421	.375	220.667	222.165	Y=1.002Ŷ	Y=1.000Ŷ
	2-part probit-normal power	.404	.362	238.875	237.793	Y=0.952Ŷ	Y=0.957Ŷ
	2-part probit-gamma power	.381	.350	229.854	228.838	Y=0.949Ŷ	Y=0.958Ŷ
Total expenses (all members) [c]	OLS	.271	.226	1501.37	1452.19	Y=Ŷ	Y=Ŷ
	MLM normal id (HCs)	.271	.226	1500.03	1449.62	Y=Ŷ	Y=Ŷ
	MLM normal id (GPs)	.272	.226	1499.20	1450.54	Y=Ŷ	Y=Ŷ

Note: COPD = chronic obstructive pulmonary disease, GPs = general practitioners, HCs = health centers, id = identity, MAD = mean absolute deviation, MLM = multilevel model, OLS = ordinary least squares.

a. Expenses generated by general practitioners through office visits, home visits, and (diagnostic) tests.

b. Expenses related to the use of prescription medication, regardless of prescriber.

c. Sum of GP expenses, medication expenses, and expenses generated in-hospital by medical specialists.

d. Percent explained variation at the individual member level.

e. Extent to which the mean of the observed values (Y) equals the mean of the expected values (Ŷ).

Appendix 6.2 Average agreement among models (2006) and per model among years (2006-2008), by performance measure

		PER YEAR AMONG MODELS (2006)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]
Yes/no physiotherapy (COPD)	OLS	.97	1.0	1.0	.96	.92	.98	.97	.93	1.0	.46	.42	.75	.44	.34	.57	.30	.19	.33
	GLM binomial probit	.97	1.0	1.0	.95	.89	.97	.95	.86	1.0	.44	.42	.58	.44	.31	.56	.31	.19	.33
	MLM normal id	.94	1.0	1.0	.95	.92	.97	.95	.93	1.0	.42	.33	.75	.44	.32	.57	.29	.19	.33
Yes/no corticosteroids (COPD)	OLS	.99	1.0	1.0	.97	.98	.99	.96	1.0	1.0	.43	.50	.58	.41	.41	.54	.43	.71	.48
	GLM binomial logit	.99	1.0	1.0	.97	.98	.99	.96	1.0	1.0	.43	.50	.50	.41	.39	.54	.43	.71	.48
	MLM normal id	-	-	-	.95	.98	.98	.94	1.0	1.0	-	-	-	.41	.39	.54	.42	.71	.48
Yes/no bronchodilators (COPD)	OLS	.97	1.0	1.0	.97	.97	.91	.97	.86	1.0	.52	.50	.42	.49	.54	.34	.49	.67	.48
	GLM binomial logit	.97	1.0	1.0	.96	.98	.87	.97	.93	1.0	.52	.50	.58	.49	.55	.35	.49	.57	.57
	MLM normal id	-	-	-	.96	.97	.89	.97	.93	1.0	-	-	-	.49	.54	.34	.49	.57	.52
Yes/no statins (diabetes)	OLS	.91	1.0	1.0	.97	.96	.96	.97	.94	1.0	.20	.33	.40	.31	.36	.30	.27	.25	.29
	GLM binomial logit	.91	1.0	1.0	.95	.95	.96	.96	.88	1.0	.20	.33	.40	.31	.35	.29	.27	.25	.29
	MLM normal id	-	-	-	.96	.96	.97	.97	.94	1.0	-	-	-	.31	.34	.27	.27	.25	.29
No. of inpatient admissions (COPD)	OLS	.83	.90	.75	.87	.86	.72	.88	.93	.68	.24	.67	.10	.04	.18	.11	.15	.33	.29
	GLM normal log	.68	.80	.50	.81	.82	.67	.79	.93	.57	.37	.42	.17	.04	.18	.12	.16	.24	.14
	GLM Poisson power	.84	.90	.75	.88	.88	.74	.89	.93	.68	.26	.67	.17	.04	.17	.11	.16	.33	.19
	GLM negbin power	.80	.90	.50	.81	.79	.49	.83	.71	.57	.27	.58	.10	.04	.16	.10	.16	.10	.19
	GLM gamma log	.84	.90	.75	.89	.89	.75	.89	.93	.71	.25	.67	.17	.04	.18	.09	.16	.33	.29
	MLM normal id	.71	.80	.65	-	-	-	-	-	-	.20	.42	.17	-	-	-	-	-	-
No. of inpatient admissions (diabetes)	OLS	.88	.84	.64	.90	.90	.88	.90	.88	.91	-.03	.07	.20	.04	.12	.12	.03	.08	.21
	GLM normal log	.79	.80	.56	.86	.88	.83	.85	.78	.84	-.01	.07	.20	.04	.13	.12	.01	.08	.21
	GLM Poisson power	.89	.84	.72	.91	.88	.87	.90	.88	.91	-.03	.07	.13	.03	.12	.10	.02	.08	.21
	GLM negbin power	.84	.80	.60	.87	.83	.76	.87	.84	.75	-.05	.07	.07	.04	.14	.13	.00	.08	.13
	GLM gamma log	.89	.84	.76	.91	.89	.89	.91	.88	.91	-.03	.07	.20	.04	.12	.11	.02	.08	.21
	MLM normal id	.83	.84	.72	-	-	-	-	-	-	-.00	.07	.13	-	-	-	-	-	-

		PER YEAR AMONG MODELS (2006)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]	All[d]	Low[e]	High[f]
No. of inpatient admissions (all members)	OLS	.93	.97	.94	.95	.95	.91	.95	.96	.92	.28	.33	.50	.13	.25	.18	.18	.30	.17
	GLM Poisson power	.93	.97	.94	.93	.94	.88	.94	.94	.88	.28	.33	.44	.13	.24	.18	.18	.31	.20
	GLM negbin power	.89	.92	.83	.90	.91	.80	.91	.90	.81	.26	.31	.44	.13	.22	.18	.18	.30	.19
	GLM gamma power	.93	.97	.94	.95	.95	.92	.96	.95	.92	.28	.33	.47	.13	.25	.17	.18	.32	.20
	MLM normal id	-	-	-	.95	.95	.91	.95	.95	.92	-	-	-	.13	.25	.17	.19	.32	.18
No. of outpatient visits (all members)	OLS	.94	.97	.97	.93	.93	.91	.92	.96	.93	.40	.39	.50	.40	.51	.46	.36	.45	.43
	GLM Poisson power	.93	.97	.97	.92	.93	.91	.91	.96	.92	.40	.39	.47	.40	.50	.45	.36	.46	.42
	GLM negbin id	.93	.92	.92	.93	.89	.87	.92	.96	.92	.41	.42	.50	.40	.48	.46	.39	.47	.49
	GLM gamma id	.95	.97	.97	.94	.94	.92	.93	.95	.93	.41	.36	.50	.40	.50	.46	.37	.46	.44
	MLM normal id	-	-	-	.88	.88	.87	.86	.87	.90	-	-	-	.41	.52	.46	.39	.46	.45
GP expenses (all members) [a]	OLS	.90	.83	.92	.92	.91	.86	.93	.90	.87	.51	.64	.36	.54	.64	.54	.51	.57	.44
	GLM normal power	.83	.71	.83	.88	.86	.80	.89	.82	.83	.47	.44	.42	.53	.60	.51	.50	.54	.43
	MLM normal id	.89	.79	.92	.91	.91	.84	.93	.90	.88	.52	.64	.42	.53	.63	.53	.51	.58	.42
	MLM gamma power	-	-	-	.90	.90	.83	.92	.87	.86	-	-	-	.53	.64	.50	.51	.63	.46
Medication expenses (all members) [b]	OLS	.74	.92	.67	.87	.91	.84	.86	.86	.82	.47	.78	.33	.47	.52	.49	.47	.57	.38
	GLM gamma power	.74	.92	.67	.79	.84	.75	.77	.81	.67	.55	.72	.36	.46	.50	.52	.47	.52	.39
	MLM normal id	-	-	-	.87	.91	.84	.86	.89	.82	-	-	-	.48	.53	.50	.48	.58	.38
Total expenses (all members) [c]	OLS	.99	1.0	.92	.99	1.0	.99	.99	1.0	1.0	.20	.28	.25	.18	.25	.24	.20	.33	.26
	MLM normal id	.99	1.0	.92	.99	1.0	.99	.99	1.0	1.0	.20	.28	.25	.18	.25	.24	.20	.32	.27

Note: '-' indicates that the relevant model did not converge or that the covariance parameter for the random provider intercept was not significant. COPD = chronic obstructive pulmonary disease, GP = general practitioner, id = identity, MLM = multilevel model, negbin = negative binomial, OLS = ordinary least squares.

a. Two-part models were excluded from these comparisons, for three reasons: expected values are very similar to those from the one-part models (the correlation coefficient was typically >.99), they do not appear to fit the data better, and calculation of standard errors is relatively complex.

b. Expenses generated by general practitioners through office visits, home visits, and (diagnostic) tests.

c. Expenses related to the use of prescription medication, regardless of prescriber.

d. Average pairwise agreement (weighted κ) on overall rankings on z-scores.

e. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the worst performance based on z-scores.

f. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the best performance based on z-scores.

Appendix 6.3 Average agreement among models (2007) and per model among years (2006-2008), by performance measure

		PER YEAR AMONG MODELS (2007)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]
Yes/no physiotherapy (COPD)	OLS	.94	1.0	1.0	.95	.91	.98	.94	.86	1.0	.46	.42	.75	.44	.34	.57	.30	.19	.33
	GLM binomial probit	.93	1.0	1.0	.94	.88	.99	.93	.71	1.0	.44	.42	.58	.44	.31	.56	.31	.19	.33
	MLM normal id	.93	1.0	1.0	.93	.90	.99	.92	.86	1.0	.42	.33	.75	.44	.32	.57	.29	.19	.33
Yes/no corticosteroids (COPD)	OLS	1.0	1.0	1.0	.97	.95	.99	.96	1.0	1.0	.43	.50	.58	.41	.41	.54	.43	.71	.48
	GLM binomial logit	1.0	1.0	1.0	.97	.96	.99	.96	1.0	1.0	.43	.50	.50	.41	.39	.54	.43	.71	.48
	MLM normal id	-	-	-	.95	.92	.98	.92	1.0	1.0	-	-	-	.41	.39	.54	.42	.71	.48
Yes/no bronchodilators (COPD)	OLS	.97	1.0	.75	.96	.95	.92	.97	1.0	.93	.52	.50	.42	.49	.54	.34	.49	.67	.48
	GLM binomial logit	.97	1.0	.75	.96	.95	.90	.98	1.0	.86	.52	.50	.58	.49	.55	.35	.49	.57	.57
	MLM normal id	-	-	-	.95	.94	.91	.97	1.0	.93	-	-	-	.49	.54	.34	.49	.57	.52
Yes/no statins (diabetes)	OLS	.95	1.0	1.0	.96	.96	.95	.96	1.0	.88	.20	.33	.40	.31	.36	.30	.27	.25	.29
	GLM binomial logit	.95	1.0	1.0	.95	.94	.91	.94	1.0	.94	.20	.33	.40	.31	.35	.29	.27	.25	.29
	MLM normal id	-	-	-	.96	.95	.93	.96	1.0	.94	-	-	-	.31	.34	.27	.27	.25	.29
No. of inpatient admissions (COPD)	OLS	.91	1.0	.90	.93	.94	.90	.93	.79	.96	.24	.67	.10	.04	.18	.11	.15	.33	.29
	GLM normal log	.80	1.0	.90	.87	.91	.83	.87	.82	.96	.37	.42	.17	.04	.18	.12	.16	.24	.14
	GLM Poisson power	.91	1.0	.90	.93	.94	.86	.93	.86	.96	.26	.67	.17	.04	.17	.11	.16	.33	.19
	GLM negbin power	.87	1.0	.50	.90	.84	.81	.90	.75	.86	.27	.58	.10	.04	.16	.10	.16	.10	.19
	GLM gamma log	.92	1.0	.90	.93	.94	.90	.94	.86	.96	.25	.67	.17	.04	.18	.09	.16	.33	.29
	MLM normal id	.89	1.0	.90	-	-	-	-	-	-	.20	.42	.17	-	-	-	-	-	-
No. of inpatient admissions (diabetes)	OLS	.90	.96	96	.92	.95	.90	.93	1.0	.88	-.03	.07	.20	.04	.12	.12	.03	.08	.21
	GLM normal log	.83	.80	.80	.85	.93	.85	.88	1.0	.75	-.01	.07	.20	.04	.13	.12	.01	.08	.21
	GLM Poisson power	.91	.96	.96	.93	.96	.89	.94	1.0	.88	-.03	.07	.13	.03	.12	.10	.02	.08	.21
	GLM negbin power	.88	.96	.96	.91	.93	.85	.92	1.0	.75	-.05	.07	.07	.04	.14	.13	.00	.08	.13
	GLM gamma log	.91	.96	.96	.93	.96	.90	.94	1.0	.88	-.03	.07	.20	.04	.12	.11	.02	.08	.21
	MLM normal id	.84	.96	.96	-	-	-	-	-	-	-.00	.07	.13	-	-	-	-	-	-

		PER YEAR AMONG MODELS (2007)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]
No. of inpatient admissions (all members)	OLS	.94	.97	.94	.96	.94	.93	.96	.95	.92	.28	.33	.50	.13	.25	.18	.18	.30	.17
	GLM Poisson power	.95	.92	.89	.95	.94	.91	.95	.94	.90	.28	.33	.44	.13	.24	.18	.18	.31	.20
	GLM negbin power	.93	.97	.89	.93	.89	.85	.94	.92	.88	.26	.31	.44	.13	.22	.18	.18	.30	.19
	GLM gamma power	.96	.97	.94	.96	.95	.94	.96	.97	.92	.28	.33	.47	.13	.25	.17	.18	.32	.20
	MLM normal id	-	-	-	.96	.94	.93	.96	.95	.92	-	-	-	.13	.25	.17	.19	.32	.18
No. of outpatient visits (all members)	OLS	.93	.86	.94	.93	.93	.92	.93	.93	.88	.40	.39	.50	.40	.51	.46	.36	.45	.43
	GLM Poisson power	.92	.89	.83	.93	.92	.91	.92	.92	.86	.40	.39	.47	.40	.50	.45	.36	.46	.42
	GLM negbin id	.93	.81	.94	.93	.87	.89	.93	.91	.88	.41	.42	.50	.40	.48	.46	.39	.47	.49
	GLM gamma id	.94	.89	.94	.95	.93	.93	.94	.92	.92	.41	.36	.50	.40	.50	.46	.37	.46	.44
	MLM normal id	-	-	-	.92	.89	.91	.90	.88	.91	-	-	-	.41	.52	.46	.39	.46	.45
GP expenses (all members) [a]	OLS	.88	.79	.92	.94	.92	.86	.93	.93	.93	.51	.64	.36	.54	.64	.54	.51	.57	.44
	GLM normal power	.79	.58	.96	.91	.85	.80	.89	.83	.92	.47	.44	.42	.53	.60	.51	.50	.54	.43
	MLM normal id	.87	.79	.96	.95	.92	.88	.93	.93	.93	.52	.64	.42	.53	.63	.53	.51	.58	.42
	MLM gamma power	-	-	-	.93	.91	.84	.92	.91	.92	-	-	-	.53	.64	.50	.51	.63	.46
Medication expenses (all members) sions ([b]	OLS	.74	.92	.67	.88	.90	.84	.88	.92	.82	.47	.78	.33	.47	.52	.49	.47	.57	.38
	GLM gamma power	.74	.92	.67	.80	.82	.74	.80	.83	.71	.55	.72	.36	.46	.50	.52	.47	.52	.39
	MLM normal id	-	-	-	.88	.90	.84	.89	.92	.83	-	-	-	.48	.53	.50	.48	.58	.38
Total expenses (all members) [c]	OLS	.98	1.0	1.0	.99	.99	.99	.99	.97	.97	.20	.28	.25	.18	.25	.24	.20	.33	.26
	MLM normal id	.98	1.0	1.0	.99	.99	.99	.99	.97	.97	.20	.28	.25	.18	.25	.24	.20	.32	.27

Note: '-' indicates that the relevant model did not converge or that the covariance parameter for the random provider intercept was not significant. COPD = chronic obstructive pulmonary disease, GP = general practitioner, id = identity, MLM = multilevel model, negbin = negative binomial, OLS = ordinary least squares.

a. Two-part models were excluded from these comparisons, for three reasons: expected values are very similar to those from the one-part models (the correlation coefficient was typically >.99), they do not appear to fit the data better, and calculation of standard errors is relatively complex.

b. Expenses generated by general practitioners through office visits, home visits, and (diagnostic) tests.

c. Expenses related to the use of prescription medication, regardless of prescriber.

d. Average pairwise agreement (weighted κ) on overall rankings on z-scores.

e. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the worst performance based on z-scores.

f. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the best performance based on z-scores.

Appendix 6.4 Average agreement among models (2008) and per model among years (2006-2008), by performance measure

		PER YEAR AMONG MODELS (2008)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]
Yes/no physiotherapy (COPD)	OLS	.89	.75	.88	.95	.89	.97	.95	.93	.86	.46	.42	.75	.44	.34	.57	.30	.19	.33
	GLM binomial probit	.85	.75	.75	.92	.84	.94	.92	.86	.86	.44	.42	.58	.44	.31	.56	.31	.19	.33
	MLM normal id	.88	.50	.88	.94	.88	.97	.95	.93	.86	.42	.33	.75	.44	.32	.57	.29	.19	.33
Yes/no corticosteroids (COPD)	OLS	.98	1.0	1.0	.98	.96	.99	.96	1.0	1.0	.43	.50	.58	.41	.41	.54	.43	.71	.48
	GLM binomial logit	.98	1.0	1.0	.98	.97	.99	.97	1.0	1.0	.43	.50	.50	.41	.39	.54	.43	.71	.48
	MLM normal id	-	-	-	.97	.95	.98	.95	1.0	1.0	-	-	-	.41	.39	.54	.42	.71	.48
Yes/no bronchodilators (COPD)	OLS	.96	1.0	1.0	.97	.98	.97	.96	.86	.93	.52	.50	.42	.49	.54	.34	.49	.67	.48
	GLM binomial logit	.96	1.0	1.0	.96	.98	.96	.95	.86	.86	.52	.50	.58	.49	.55	.35	.49	.57	.57
	MLM normal id	-	-	-	.96	.98	.97	.96	.86	.93	-	-	-	.49	.54	.34	.49	.57	.52
Yes/no statins (diabetes)	OLS	.97	1.0	1.0	.98	.97	.96	.98	1.0	1.0	.20	.33	.40	.31	.36	.30	.27	.25	.29
	GLM binomial logit	.97	1.0	1.0	.97	.95	.96	.97	1.0	1.0	.20	.33	.40	.31	.35	.29	.27	.25	.29
	MLM normal id	-	-	-	.97	.97	.96	.97	1.0	1.0	-	-	-	.31	.34	.27	.27	.25	.29
No. of inpatient admissions (COPD)	OLS	.87	.85	.50	.92	.90	.85	.92	.93	.79	.24	.67	.10	.04	.18	.11	.15	.33	.29
	GLM normal log	.84	.80	.60	.86	.89	.77	.86	.82	.61	.37	.42	.17	.04	.18	.12	.16	.24	.14
	GLM Poisson power	.87	.85	.70	.92	.92	.85	.93	.93	.75	.26	.67	.17	.04	.17	.11	.16	.33	.19
	GLM negbin power	.86	.75	.70	.90	.83	.77	.90	.82	.71	.27	.58	.10	.04	.16	.10	.16	.10	.19
	GLM gamma log	.87	.85	.70	.92	.92	.86	.93	.93	.79	.25	.67	.17	.04	.18	.09	.16	.33	.29
	MLM normal id	.82	.80	.60	-	-	-	-	-	-	.20	.42	.17	-	-	-	-	-	-
No. of inpatient admissions (diabetes)	OLS	.90	.84	.80	.92	.95	.88	.91	.97	.84	-.03	.07	.20	.04	.12	.12	.03	.08	.21
	GLM normal log	.84	.84	.68	.83	.88	.78	.82	.88	.66	-.01	.07	.20	.04	.13	.12	.01	.08	.21
	GLM Poisson power	.89	.80	.80	.92	.95	.87	.91	.97	.84	-.03	.07	.13	.03	.12	.10	.02	.08	.21
	GLM negbin power	.87	.84	.72	.91	.90	.83	.89	.97	.75	-.05	.07	.07	.04	.14	.13	.00	.08	.13
	GLM gamma log	.90	.84	.80	.93	.95	.88	.92	.97	.84	-.03	.07	.20	.04	.12	.11	.02	.08	.21
	MLM normal id	.84	.80	.68	-	-	-	-	-	-	-.00	.07	.13	-	-	-	-	-	-

		PER YEAR AMONG MODELS (2008)									PER MODEL AMONG YEARS (2006-2008)								
		Health centers			GPs, no health center			GPs, yes health center			Health centers			GPs, no health center			GPs, yes health center		
Measure	**Model** [a]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]	All [d]	Low [e]	High [f]
No. of inpatient admissions (all members)	OLS	.93	.97	.86	.95	.94	.97	.95	.96	.97	.28	.33	.50	.13	.25	.18	.18	.30	.17
	GLM Poisson power	.95	.97	.86	.95	.94	.93	.95	.95	.97	.28	.33	.44	.13	.24	.18	.18	.31	.20
	GLM negbin power	.93	.92	.83	.93	.90	.89	.93	.92	.92	.26	.31	.44	.13	.22	.18	.18	.30	.19
	GLM gamma power	.95	.97	.89	.96	.95	.95	.96	.96	.97	.28	.33	.47	.13	.25	.17	.18	.32	.20
	MLM normal id	-	-	-	.96	.95	.94	.96	.96	.97	-	-	-	.13	.25	.17	.19	.32	.18
No. of outpatient visits (all members)	OLS	.94	.94	1.0	.95	.94	.94	.94	.95	.94	.40	.39	.50	.40	.51	.46	.36	.45	.43
	GLM Poisson power	.95	.94	1.0	.95	.94	.93	.95	.96	.96	.40	.39	.47	.40	.50	.45	.36	.46	.42
	GLM negbin id	.94	.94	1.0	.94	.89	.89	.94	.94	.90	.41	.42	.50	.40	.48	.46	.39	.47	.49
	GLM gamma id	.95	.94	1.0	.96	.95	.95	.95	.97	.96	.41	.36	.50	.40	.50	.46	.37	.46	.44
	MLM normal id	-	-	-	.95	.94	.94	.94	.95	.92	-	-	-	.41	.52	.46	.39	.46	.45
GP expenses (all members) [a]	OLS	.85	.75	.83	.92	.93	.88	.93	.92	.90	.51	.64	.36	.54	.64	.54	.51	.57	.44
	GLM normal power	.74	.58	.88	.88	.88	.81	.89	.83	.84	.47	.44	.42	.53	.60	.51	.50	.54	.43
	MLM normal id	.85	.75	.88	.92	.92	.87	.93	.92	.88	.52	.64	.42	.53	.63	.53	.51	.58	.42
	MLM gamma power	-	-	-	.90	.91	.84	.92	.85	.84	-	-	-	.53	.64	.50	.51	.63	.46
Medication expenses (all members) sions ([b]	OLS	.77	1.0	.67	.90	.92	.87	.91	.89	.86	.47	.78	.33	.47	.52	.49	.47	.57	.38
	GLM gamma power	.77	1.0	.67	.82	.86	.76	.83	.81	.76	.55	.72	.36	.46	.50	.52	.47	.52	.39
	MLM normal id	-	-	-	.90	.93	.87	.90	.89	.88	-	-	-	.48	.53	.50	.48	.58	.38
Total expenses (all members) [c]	OLS	1.0	1.0	1.0	.99	1.0	1.0	.99	.97	1.0	.20	.28	.25	.18	.25	.24	.20	.33	.26
	MLM normal id	1.0	1.0	1.0	.99	1.0	1.0	.99	.97	1.0	.20	.28	.25	.18	.25	.24	.20	.32	.27

Note: '-' indicates that the relevant model did not converge or that the covariance parameter for the random provider intercept was not significant. COPD = chronic obstructive pulmonary disease, GP = general practitioner, id = identity, MLM = multilevel model, negbin = negative binomial, OLS = ordinary least squares.

a. Two-part models were excluded from these comparisons, for three reasons: expected values are very similar to those from the one-part models (the correlation coefficient was typically >.99), they do not appear to fit the data better, and calculation of standard errors is relatively complex.

b. Expenses generated by general practitioners through office visits, home visits, and (diagnostic) tests.

c. Expenses related to the use of prescription medication, regardless of prescriber.

d. Average pairwise agreement (weighted κ) on overall rankings on z-scores.

e. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the worst performance based on z-scores.

f. Average pairwise percentage agreement on classification of providers to the 10 percent providers with the best performance based on z-scores.

CHAPTER 7

Profiling individual physicians using administrative data from a single insurer: variance components, reliability, and implications for performance improvement efforts

With René van Vliet

Medical Care, 2013, 51(8): 731-739.

ABSTRACT

Background: Individual physicians are increasingly being subjected to comparative performance assessments. When single-insurer data are used to profile individual physicians' performance, reliable measurements and comparisons are uncertain because of small sample sizes.

Methods: Administrative data (2006-2008) from a Dutch health insurer are used to examine variation in general practitioners' (GPs) performance on expenses (five measures), utilization of hospital care (two measures), and clinical quality for diabetes and chronic obstructive pulmonary disease (COPD, six measures). Unadjusted and adjusted multilevel models are used to separate total variance in a measure in a between-GP and a within-GP component. The components are used to calculate intraclass correlation coefficients (ICCs), reliability, and sample size requirements at common reliability thresholds.

Results: Average ICCs varied between 0.07 percent (hospital admissions) and 8.34 percent (physiotherapy for COPD patients). Risk adjustment often greatly changed the relative size of variance components and often led to lower ICCs. In addition, ICCs and thus reliability generally decreased over time. Eight measures had reliabilities above 0.70, and three of these (all GP-related expenses) above 0.90. Measures related to utilization of hospital care had reliabilities below 0.60 or even 0.50. For five measures, the vast majority of GPs had sufficient patients to reach 0.70 reliability. At a reliability of 0.90, however, there were no measures for which all GPs met the sample size requirements.

Conclusions: Reliable measurement of individual physicians' performance using single-purchaser data is challenging. For most measures reliability was insufficient to allow for high-stakes applications or even any application of profiling. Future research should continue to explore methods for enhancing the reliability of individual physicians' performance profiles.

7.1 INTRODUCTION

In many countries significant practice variation and escalating costs have led to increased emphasis on finding effective ways to increase the efficiency of healthcare delivery. Healthcare providers, including individual physicians, physician groups, and institutions, are increasingly being held accountable for the quality and costs of care and subjected to comparative performance assessment or "profiling" (Hanchak & Schlackman, 1995; Landon et al., 2003; Institute of Medicine, 2007). The results of these assessments are used by purchasers as input for various improvement efforts, including feedback to providers, P4P, public reporting, and selective contracting (Fung et al., 2008; Brennan et al., 2008). Yet if done in an uninformed way, profiling can produce meaningless results that may undermine any possible positive effect on performance (Christiansen & Morris, 1997; Shahian et al., 2001; Adams et al., 2010a) and can even provide incentives for undesired behavior (Hofer et al., 1999; Krein et al., 2002; Dranove et al., 2003; Chen et al., 2011).

Performance profiles require at least two essential features to be useful for improvement interventions: adequate risk adjustment to prevent systematic misclassification of providers due to differences in casemix (Tucker et al., 1996; Iezzoni, 2003; Pope & Kautter, 2007; Ash et al., 2012), and adequate reliability to prevent random misclassification of providers due to chance (Safran et al., 2006; Scholle et al., 2008; Adams et al., 2010a). In this paper, the focus is primarily on the latter. When profiles have low reliability, they are driven by random chance instead of true performance, and interventions based on them may arbitrarily and unfairly penalize or reward providers and patients may be misled. For profiles to provide a reliable picture of performance differences among providers, a sufficient number of patients per provider must be sampled. In addition, variation *between* providers must be sufficiently large relative to patient-to-patient variation *within* providers. When providers' performance scores are concentrated, great precision is needed to reliably distinguish them, which can be achieved by sampling large numbers of patients. When scores are dispersed, rankings may be reproducible even when scores are estimated with limited precision.

Previous research examining performance variation and reliability has predominantly focused on physician groups or hospitals (Rodriquez et al., 2012; Huang et al., 2005; van Diskhoeck et al., 2011) and/or used data from large public purchasers (Lyratzopoulos et al., 2011; Salisbury et al., 2010; Roland et al., 2009) or pooled, mostly cross-sectional data from multiple purchasers or government agencies (Sequist et al., 2011; van Dishoeck et al., 2011; Mehrotra et al., 2010; Adams et al., 2010a; Safran et al., 2006). However, although multipayer initiatives can be very helpful in increasing sample size and in reducing the burden of data collection for providers, they are still rare (Rodriguez et al., 2012). Consequently, profiling is predominantly being applied by individual purchasers. In addition, as the majority of physicians in the US (and throughout Europe) still work in solo or small-group practices (Landon & Normand, 2008) and individual physicians make important decisions that affect

performance, assessing individual physicians' performance continues to be the predominate approach to profiling (Selby et al., 2010; Mehrotra et al., 2010; Kaplan et al., 2009; Scholle et al., 2008; Safran et al., 2006; Krein et al., 2002; Katon et al., 2000; Hofer et al., 1999).

When single-purchaser data are used to profile individual physicians' performance, adequate reliability is questionable because of small sample sizes. Indeed, the few studies that have analyzed variation among individual physicians using single-purchaser data have generally found low reliability (Selby et al., 2010; Scholle et al., 2008; Katon et al., 2000; Hofer et al., 1999). These studies were all conducted in the US, analyzed specific sets of measures (e.g., only clinical process measures or only measures related to one specific condition), and did not provide insight in changes in variability and reliability over time. By analyzing performance variation among Dutch general practitioners (GPs) for various types of measures across multiple years, the goal of this paper is to further examine the feasibility of comparative performance assessments for individual physicians using single-purchaser data. Specifically, we address three questions: (1) to what extent can GPs be reliably compared on measures of resource use and quality of care derived from the administrative data from a single purchaser? (2) What is the influence of adjustment for sociodemographic and clinical patient characteristics on estimated variance components and reliability? (3) To what extent do variance components and reliability vary over time?

7.2 IMPORTANCE OF RELIABILITY IN THE CONTEXT OF PROFILING

In general, statistical reliability indicates the reproducibility or consistency of a measure across repeated measurements (Nunnally & Bernstein, 1994). In this paper, reliability indicates the proportion of variation in GP-level scores attributable to true variation between GPs, and therefore the extent to which observed scores adequately discriminate GPs' performance (Lyratzopoulos et al., 2010). The less random variation or "noise" there is in the performance estimates, the more "signal" they contain and the higher the reliability will be (Adams et al., 2010a). When reliability is low, there is greater risk that providers will be misclassified, for example as average, when they are actually above or below average (Adams et al., 2010a; Safran et al., 2006). A reliability coefficient below 0.7 is usually considered undesirable for profiling; values above 0.9 are required for "high-stakes" applications like payment incentives and public reporting (Hofer et al., 1999; Safran et al., 2006; Scholle et al., 2008; Sequist et al., 2011; Lyratzopoulos et al., 2011).

Reliability is a function of sample size per provider (e.g., a GP) and the intraclass correlation coefficient (ICC). The ICC represents the proportion of total variation attributable to variation between GPs, and is calculated as the between-GP variance divided by the between-GP variance plus the within-GP variance (Snijders & Bosker, 2011). A high

ICC thus indicates greater variation between GPs relative to residual variation within GPs. The ICC also provides insight in the extent to which GPs can influence the measure being analyzed. GPs have more control over process measures than over clinical outcomes and resource use, so ICCs can be expected to be higher for the former than for the latter.

A recent literature review (Fung et al., 2010) found that across the 22 identified studies physician-level ICCs varied between 0 and 19 percent and were typically below 10 percent. This suggests that the proportion of variation in measures explained by physicians will almost always be "low", and that the majority of the variation will be found at the patient level. Nonetheless, it may often be more efficient and effective to intervene through physicians, especially if they are more directly accessible than patients (Fung et al., 2010).

7.3 METHODS

7.3.1 Study setting and data

In the Netherlands, risk-bearing health insurers are expected to act as prudent purchasers of care on behalf of their members. To fulfill this role, insurers have various managed-care tools at their disposal, many of which require an adequate profiling system. Because the primary care sector is often patients' first contact point with the healthcare system and the place where most preventive and chronic care services are provided and coordinated from, primary care is an important sector for insurers. In this respect, insurers have increasingly been confronted with the question if GPs can be reliably compared on measures derived from data routinely available in insurers' administrative files.

In this paper, we use administrative data over the years 2006-2008 from a large Dutch health insurer. For each year, data on approximately 2.8 million members are available, including sociodemographic characteristics, proxies for health status, and a link to the GP with whom members were registered. In the Netherlands, virtually everyone is registered with a GP, so GPs have fixed patient panels. Because they act as gatekeepers, GPs can exert influence on the amount and type of specialist and hospital care their patients use.

7.3.2 Performance measures (dependent variables)

We constructed three types of performance measures: expenses, utilization of hospital care, and clinical quality. The *expenses* measures, all top-coded at the 99th percentile to prevent a few extreme observations from contributing a disproportionate share to the variance, are GP expenses (excluding medications), GP medication expenses, total medication expenses (all prescribers), total GP expenses (sum of GP expenses and GP medication expenses), and total expenses (sum of GP expenses, total medication expenses, and hospital expenses). For *utilization of hospital care*, we used the number of inpatient admissions and outpatient visits. Both are indicated by "diagnosis treatment combinations" (DTC), which were imple-

mented in the Dutch healthcare system to facilitate contracting for hospital services (Van de Ven & Schut, 2009). A DTC is a predefined care product, selected by the medical specialist based on the patient's condition and representing all hospital procedures/services related to treating a patient with a specific diagnosis within a fixed period. It is similar to a DRG used by Medicare in the US, except that a DTC is more broadly defined and also include payment for medical specialists. Finally, we constructed *clinical quality* measures for diabetes and chronic obstructive pulmonary disease (COPD). For diabetes, the percentage of patients on a statin and the number of diabetes-related admissions are available. For COPD, we defined three process measures: the percentage of patients using corticosteroids (long-term control medication), the percentage of patients using bronchodilators (rescue medication), and the percentage of patients receiving physiotherapy. The number of COPD-related admissions is used as outcome measure. We thus have three types of dependent variables: continuous (expenses, lower is better), count (utilization, lower is better), and binary (clinical processes, higher is better). For members enrolled for less than a year, dependent variables were annualized and weighted based on months of enrollment in the analyses. For example, for a person enrolled for three months with medication expenses €300 the annualization would result in a value of €1,200. This observation would then get a weight of 0.25 in the analyses.

7.3.3 Risk adjustment (independent variables)

Table 7.1 lists the patient characteristics that are included in the models that adjust for casemix (described below). In addition to age and sex, we included five indicators of socioeconomic status, three of which were measured at the member's ZIP-code level. For example, the three categories of "educational level" refer to the average educational level of people living in the member's ZIP-code area. "Ethnicity" is based on the percentage of persons living in the member's ZIP-code area with at least one parent born in Turkey, Africa, Latin-America, or Asia (excluding Japan). This variable was included because different ethnic groups may exhibit different patterns of utilization (Van der Lucht & Verweij, 2010) and may not be equally compliant with recommended treatment (Peeters et al., 2011; Bailey & Kodack, 2011). "Urbanization" is based on the number of adjacent addresses and on the number of inhabitants in the member's town or city of residence.

We also included two proxies for health: pharmacy-based cost groups (PCG) and diagnosis cost groups (DCG). Both have been developed for the Dutch risk-equalization scheme for health insurers and designed to identify patients with chronic conditions (Van de Ven et al., 2004; Prinsze & Van Vliet, 2007). PCGs are based on prior outpatient use of medication. A member is assigned to a certain PCG if prescribed at least 181 defined daily doses of the relevant disease-specific drug in the prior year. For example, if a member was prescribed at least 181 defined daily doses of dopaminergic agents in year t, he/she will be classified in the PCG for Parkinson in year t+1. Our data distinguishes twenty PCGs, all of which relate to a chronic condition. Members were identified as having diabetes if classified in a PCG

TABLE 7.1. Risk adjusters included in the full models

Variable	Coding
Age-sex interactions [a]	38 (19x2) categories
Living in a deprived area	Yes/no
Monthly income	ZIP-code, 10 categories (1 = low, 10 = high)
Educational level	ZIP-code, 3 categories (1 = low, 3 = high)
Ethnicity	ZIP-code, 5 categories (1 = less than 5% non-Western immigrants, 5 = more than 40% non-Western immigrants)
Urbanization	5 categories (1 = low, 5 = high)
Deceased	Yes/no
Pharmacy-based cost groups [b]	20 comorbidities: chronic nonspecific respiratory conditions, cystic fibrosis, high cholesterol, cancer, Crohn's disease, diabetes type 1, diabetes type 2 (2 forms), epilepsy, glaucoma, grow hormones, heart diseases, HIV/AIDS, kidney diseases, mental disorders, neurological disorders, Parkinson's disease, rheumatoid arthritis, thyroid disorders, transplantations
Diagnosis cost groups	13 categories

a. For the COPD measures, fewer age-sex groups were included as we defined COPD patients as being 45 years of age or older.
b. Not all comorbidities are included for all patient groups and measures: for diabetes patients, the PCGs for diabetes are excluded; for the statins measure, the PCG "high cholesterol" is excluded; for COPD patients, the PCG "chronic nonspecific respiratory conditions" is excluded.

for diabetes. COPD patients were defined in a similar way among members 45 years of age or older. DCGs are based on the diagnoses of prior hospitalizations. About 500 DTCs for which high future expenses are likely were clustered based on homogeneity of expenses, resulting in thirteen DCGs. If a member was admitted to the hospital and classified in one of these DTCs in year t, this member will be classified in the associated DCG in year t+1.

All of these variables have been developed for explaining cost variation at the individual-member level (for the purpose of calculating risk-adjusted capitation payments for health insurers) and are therefore appropriate for the expenses and utilization measures (Van Kleef & Van Vliet 2010). This was confirmed when we ran the models; virtually all independent variables were significantly associated with the dependent variables. For process measures, however, this was not always the case, especially regarding the DCGs. But because the pattern of (lack of) significant associations with the dependent variables was not consistent over time, we chose to include all independent variables in the models to ensure comparability.

7.3.4 Statistical analysis

Using the GLIMMIX procedure in SAS 9.2, we employed generalized linear multilevel models with a random GP intercept with mean zero and constant variance to separate total variance in between- and within-GP components. Appropriate distributions and link-functions were identified by monitoring algorithm convergence and goodness-of-fit statistics. As the expenses measures were skewed and the residuals showed a significant

departure from normality, we chose a lognormal distribution for these measures. For disease-specific hospital admissions, the negative binomial distribution with a log-link was used, whereas for the total number of hospital admissions and outpatient visits we used the Poisson distribution with a log-link. Finally, for binary variables a binomial distribution and logit-link were assumed. Models for the disease-specific measures were estimated by maximum likelihood with Laplace approximation, whereas for all other measures residual pseudo-likelihood was used.

For each measure, both an "empty model" and a "full model" were estimated. The former is equivalent to a one-way analysis of variance with random effects and provides insight in the basic partitioning of variance (Krein et al., 2002; Snijders & Bosker, 2011). The latter adjusts for the fixed effects of patient characteristics assuming these effects are constant across GPs. Separate ICCs are calculated using the variance components generated by both models. Using the adjusted ICCs and mean sample sizes per GP, reliability was calculated using the Spearman-Brown prophecy: (n x ICC) / (1 + [n – 1] x ICC), with n denoting sample size (Nunnally & Bernstein, 1994; Snijders & Bosker, 2011). This formula was also used to calculate sample size requirements at two common reliability thresholds (0.7 and 0.9).

The full models were also used to generate GPs-specific performance scores (residuals). These scores are based on pooling of information across GPs and "shrunken" towards the grand mean with the degree of shrinkage dependent on the amount of information per GP (i.e., the sample size). The estimates account for regression-to-the-mean bias, deal with uncertainty because of small sample size, and appropriately deal with dependent observations through explicit modeling of the hierarchical structure of the data (Christiansen & Morris, 1997; DeLong et al., 1997; Setodji & Shwartz, 2013). In essence, a GP's observed performance score is adjusted by using information from all GPs to reduce the likelihood of misclassifying this GP. When the within-GP error is high (e.g., because of a small sample size), the performance estimate is shrunken more toward the mean than when the within-GP error is low. In other words, when overall reliability is high, the models "borrow" less from the mean performance of all providers to generate the GP-specific performance scores than when overall reliability is low (Friedberg & Damberg, 2011).

7.3.5 GP sample

Variation is analyzed only for GPs who did not work in a health center, which is an entity in which multiple GPs and other primary care providers (e.g., physiotherapists, practice nurses, dietitians) provide and coordinate care, usually from the same building. During 2006-2008, health centers participated in a P4P-program in which many of the measures analyzed here were included. The effect of this program is a potential confounder that we can easily eliminate without much loss of data (about 200,000 records). In addition, we only analyze variation for GPs for whom we have data for the entire 3-year period so that changes in variances would not reflect removal or introduction of GPs over time. Furthermore, we

only included GPs whose sample size did not vary by more than 250 patients from one year to the next to ensure that variation is analyzed only for patients for whom GPs can reasonably be held accountable. Finally, for the generic measures GPs were only included if they had at least 100 patients in all years. For disease-specific measures, GPs had to have at least 30 patients in all years. We chose these thresholds as they are common in practice and literature (e.g., Scholle et al. 2009; Kaplan et al. 2009). The final sample consists of 4,019 GPs for the generic measures, 537 GPs for the diabetes measures, and 447 GPs for the COPD measures.

7.4 RESULTS

7.4.1 Descriptive statistics

Table 7.2 provides some descriptive statistics. The average age among all members is 41 years; for COPD and diabetes patients this is 68 and 66, respectively. Compared to the average member, COPD and diabetes patients more often live in areas with a lower average income and educational level. The proportion of members with an (additional) chronic condition increased over time: from 16.6 to 18.0 percent for all members, from 46.9 to 49.8 percent for COPD patients, and from 65.7 to 73.4 percent for diabetes patients. Table 7.3 shows the unadjusted means and ranges for the thirteen measures. Variation among GPs appears to be largest for physiotherapy and corticosteroids, with coefficients of variation (CV; the standard deviation divided by the mean) averaging to 0.83 and 0.29, respectively. Regarding utilization, GP-level variation is observed particularly for disease-related admissions. For example, the average number of COPD-related admissions varies between 0.00 and 0.43 per GP in 2008 (CV=0.78). For expenses, variation appears to be largest for medication expenses.

7.4.2 Variance component analysis

Table 7.4 presents the estimated between-GP variance components and ICCs. Measures for which a relatively large proportion of the variance is attributable to GPs are physiotherapy (average adjusted ICC=8.34 percent), statins (4.43 percent), and GP medication expenses (4.27 percent). Measures with the smallest adjusted ICCs are all hospital admissions (0.07 percent), diabetes-related hospital admissions (0.61 percent), outpatient visits (0.72 percent), and total expenses (0.84 percent). ICCs tend to be highest for measures over which GPs have much control and lowest for measures particularly sensitive to chance and (prescribing) behavior of other providers. An exception is the number of COPD-related admissions.

Risk adjustment had a varying impact on the relative magnitude of between- and within-GP variances. Measures for which ICCs were largely unaffected are physiotherapy, corticosteroids, diabetes-related admissions, and the three types of GP expenses. For

TABLE 7.2. Descriptive statistics of the study sample, by year

	All members (4,019 GPs)			Members with COPD (447 GPs)			Members with diabetes (537 GPs)		
	2006	*2007*	*2008*	*2006*	*2007*	*2008*	*2006*	*2007*	*2008*
n	2,070,078	2,036,490	2,056,025	26,016	26,684	26,994	37,832	38,045	38,166
Age [mean (SD)]	40 (23)	41 (23)	41 (23)	68 (12)	68 (12)	68 (12)	66 (15)	66 (14)	66 (14)
Male (%)	50.7	50.6	50.5	52.7	53.1	53.7	53.2	54.8	53.6
Living in a deprived area (%)	6.0	5.9	5.9	7.8	7.6	7.4	11.9	11.8	11.9
Monthly income [mean (SD)]	5.4 (2.9)	5.4 (2.9)	5.4 (2.9)	3.5 (2.5)	3.5 (2.5)	3.5 (2.5)	3.5 (2.5)	3.5 (2.5)	3.5 (2.5)
Educational level [mean (SD)]	2.0 (0.8)	2.0 (0.8)	2.0 (0.8)	1.6 (0.8)	1.6 (0.8)	1.7 (0.8)	1.6 (0.8)	1.6 (0.8)	1.6 (0.8)
Ethnicity [mean (SD)]	1.3 (0.8)	1.3 (0.8)	1.3 (0.7)	1.3 (0.8)	1.3 (0.8)	1.3 (0.8)	1.4 (0.9)	1.4 (0.9)	1.4 (0.9)
Urbanization [mean (SD)]	3.0 (1.4)	3.0 (1.3)	3.0 (1.4)	2.9 (1.5)	2.9 (1.5)	2.9 (1.5)	2.8 (1.5)	2.8 (1.5)	2.8 (1.5)
Deceased (%)	0.7	0.8	0.8	3.0	3.2	3.0	1.9	2.1	1.9
In a pharmacy-based cost group (%)	16.6	17.4	18.0	46.9	47.9	49.8	65.7	68.8	73.4
In ≥2 pharmacy-based cost groups (%)	4.6	5.0	5.3	18.1	19.1	20.8	22.9	24.2	26.3
In ≥3 pharmacy-based cost groups (%)	1.2	1.3	1.4	5.4	5.9	6.6	5.1	5.7	6.2
In a diagnosis cost group (%)	2.7	2.4	2.5	15.1	12.7	13.3	12.3	10.9	10.7

TABLE 7.3. Unadjusted means and ranges for thirteen performance measures, 2006-2008

Measure	Year	Mean, member level (SD)	Mean, GP level (SD)	Range across GPs
Fraction of COPD patients who received physiotherapy	2006	0.043 (0.202)	0.044 (0.040)	0.000 - 0.222
	2007	0.048 (0.213)	0.049 (0.040)	0.000 - 0.211
	2008	0.056 (0.228)	0.055 (0.041)	0.000 - 0.211
Fraction of COPD patients who used bronchodilators	2006	0.818 (0.386)	0.820 (0.081)	0.541 - 0.983
	2007	0.810 (0.392)	0.812 (0.079)	0.585 - 1.000
	2008	0.811 (0.392)	0.812 (0.076)	0.539 - 0.979
Fraction of COPD patients who used corticosteroids	2006	0.369 (0.482)	0.361 (0.103)	0.103 - 0.709
	2007	0.360 (0.480)	0.355 (0.102)	0.107 - 0.681
	2008	0.356 (0.479)	0.348 (0.097)	0.114 - 0.669
Fraction of diabetes patients who used statins	2006	0.588 (0.492)	0.588 (0.123)	0.265 - 0.916
	2007	0.629 (0.483)	0.628 (0.113)	0.289 - 0.933
	2008	0.639 (0.481)	0.637 (0.111)	0.276 - 0.916
No. of COPD-related hospital admissions, COPD patients	2006	0.081 (0.393)	0.085 (0.062)	0.000 - 0.383
	2007	0.091 (0.430)	0.094 (0.067)	0.000 - 0.391
	2008	0.081 (0.390)	0.086 (0.066)	0.000 - 0.425
No. of diabetes-related hospital admissions, diabetes patients	2006	0.087 (0.406)	0.086 (0.054)	0.000 - 0.421
	2007	0.087 (0.447)	0.086 (0.057)	0.000 - 0.579
	2008	0.077 (0.394)	0.076 (0.048)	0.000 - 0.291
No. of hospital admissions, all members	2006	0.101 (0.444)	0.091 (0.031)	0.007 - 0.241
	2007	0.108 (0.475)	0.098 (0.033)	0.013 - 0.321
	2008	0.099 (0.442)	0.090 (0.031)	0.005 - 0.262
No. of outpatient visits, all members	2006	0.526 (1.432)	0.495 (0.144)	0.178 - 1.377
	2007	0.530 (1.538)	0.505 (0.149)	0.187 - 1.244
	2008	0.358 (1.235)	0.343 (0.100)	0.074 - 0.900

TABLE 7.3. (continued)

Measure	Year	Mean, member level (SD)	Mean, GP level (SD)	Range across GPs
GP expenses, all members	2006	€121 (€111)	€114 (€16)	€61 - €228
	2007	€130 (€123)	€124 (€17)	€72 - €252
	2008	€128 (€120)	€122 (€18)	€55 - €278
GP medication expenses, all members	2006	€182 (€517)	€159 (€65)	€2 - €509
	2007	€206 (€572)	€180 (€73)	€1 - €640
	2008	€204 (€584)	€180 (€66)	€4 - €576
Total medication expenses, all members	2006	€281 (€1,021)	€252 (€91)	€34 - €925
	2007	€319 (€1,178)	€289 (€104)	€39 - €1,195
	2008	€313 (€1,232)	€286 (€101)	€53 - €1,370
Total GP expenses, all members	2006	€303 (€569)	€273 (€75)	€93 - €730
	2007	€336 (€640)	€304 (€84)	€102 - €883
	2008	€331 (€643)	€302 (€77)	€118 - €834
Total expenses, all members	2006	€1,494 (€5,558)	€1,349 (€419)	€278 - €3,638
	2007	€1,562 (€6,146)	€1,432 (€444)	€335 - €4,858
	2008	€1,501 (€5,055)	€1,389 (€384)	€396 - €5,180

TABLE 7.4. Estimated variance components, intraclass correlations, reliabilities, and sample size requirements for thirteen performance measures, 2006-2008

Measure	Year	Between-GP variance, unadjusted [a]	Between-GP variance, adjusted [b]	Intraclass correlation, unadjusted	Intraclass correlation, adjusted	GP-level reliability, adjusted	Mean sample size	Required sample size (r=0.70)	GPs with sufficient size (r=0.70)	Required sample size (r=0.90)	GPs with sufficient size (r=0.90)
Physiotherapy (COPD)	2006	0.3923	0.3923	10.65%	10.65%	0.87	58	20	100%	75	15.7%
	2007	0.2837	0.2874	7.94%	8.03%	0.84	60	27	100%	103	7.4%
	2008	0.2080	0.2230	5.95%	6.35%	0.80	60	34	94.9%	133	4.0%
Bronchodilators (COPD)	2006	0.1700	0.1364	4.91%	3.98%	0.71	58	56	37.8%	217	0.7%
	2007	0.1417	0.1223	4.13%	3.58%	0.69	60	63	31.1%	242	0.2%
	2008	0.1240	0.1100	3.63%	3.24%	0.67	60	70	24.2%	269	0.2%
Corticosteroids (COPD)	2006	0.1133	0.1132	3.33%	3.33%	0.67	58	68	22.1%	262	0.2%
	2007	0.1161	0.1145	3.41%	3.36%	0.68	60	67	26.2%	259	0.2%
	2008	0.1026	0.1035	3.02%	3.05%	0.65	60	74	19.7%	286	0.0%
Statins (diabetes)	2006	0.2105	0.1938	6.01%	5.56%	0.80	71	40	88.5%	153	3.4%
	2007	0.1743	0.1349	5.03%	3.94%	0.74	71	57	60.5%	219	1.1%
	2008	0.1633	0.1297	4.73%	3.79%	0.74	71	59	56.8%	228	0.7%
COPD-related admissions	2006	0.0017	0.0011	0.02%	1.64%	0.49	58	140	2.9%	541	0.0%
	2007	0.1129	0.1082	1.28%	3.25%	0.67	60	69	24.4%	268	0.0%
	2008	0.1149	0.0736	1.34%	2.57%	0.61	60	89	11.4%	342	0.0%
Diabetes-related admissions	2006	0.0495	0.0007	0.65%	0.54%	0.28	71	429	0.0%	1656	0.0%
	2007	0.0612	0.0404	0.71%	0.95%	0.40	71	244	0.4%	942	0.0%
	2008	0.0241	0.0172	0.23%	0.35%	0.20	71	624	0.0%	2,590	0.0%
All hospital admissions	2006	0.0002	0.0000	0.29%	0.08%	0.28	515	3,031	0.9%	11,693	0.0%
	2007	0.0002	0.0000	0.30%	0.07%	0.27	507	3,238	0.7%	12,489	0.0%
	2008	0.0002	0.0000	0.29%	0.07%	0.28	512	3,145	0.8%	12,131	0.0%
Outpatient visits	2006	0.0030	0.0014	1.35%	0.80%	0.80	515	291	57.3%	1,123	11.9%
	2007	0.0029	0.0014	1.32%	0.74%	0.79	507	312	51.7%	1,204	10.5%
	2008	0.0016	0.0008	1.00%	0.63%	0.76	512	370	41.0%	1,427	7.0%

TABLE 7.4. (continued)

Measure	Year	Between-GP variance, unadjusted [a]	Between-GP variance, adjusted [b]	Intraclass correlation, unadjusted	Intraclass correlation, adjusted	GP-level reliability, adjusted	Mean sample size	Required sample size (r=0.70)	GPs with sufficient size (r=0.70)	Required sample size (r=0.90)	GPs with sufficient size (r=0.90)
GP expenses	2006	0.0120	0.0086	4.16%	4.28%	0.96	515	52	100%	201	78.3%
	2007	0.0102	0.0067	3.56%	3.30%	0.95	507	68	100%	263	63.3%
	2008	0.0106	0.0076	3.70%	3.96%	0.95	512	57	100%	218	75.2%
GP medication expenses	2006	0.2667	0.1781	4.37%	4.90%	0.96	515	45	100%	175	85.1%
	2007	0.2722	0.1768	4.46%	4.97%	0.96	507	45	100%	172	86.4%
	2008	0.1777	0.0991	2.98%	2.95%	0.94	512	77	100%	296	56.5%
Total medication expenses	2006	0.1681	0.0544	2.62%	1.56%	0.89	515	147	91.6%	569	21.4%
	2007	0.1679	0.0559	2.65%	1.63%	0.89	507	141	93.7%	544	22.2%
	2008	0.1317	0.0408	2.12%	1.22%	0.86	512	188	82.6%	727	17.4%
Total GP expenses	2006	0.0330	0.0149	3.45%	3.30%	0.95	515	68	100%	263	63.2%
	2007	0.0329	0.0135	3.40%	3.01%	0.94	507	75	100%	290	57.4%
	2008	0.0271	0.0105	2.82%	2.46%	0.93	512	93	100%	358	43.2%
Total expenses	2006	0.0453	0.0116	1.97%	0.87%	0.82	515	265	62.7%	1,022	13.6%
	2007	0.0451	0.0122	2.03%	0.87%	0.82	507	265	62.8%	1,020	13.1%
	2008	0.0392	0.0111	1.73%	0.78%	0.80	512	298	55.9%	1,151	11.3%

a. All unadjusted variance components statistically significant at 1 percent significance level, except for diabetes-related admissions in 2008 (P=0.08).

b. All adjusted variance components statistically significant at 1 percent significance level, except for diabetes-related admissions in 2007 (P=0.01) and 2008 (P=0.14).

physiotherapy and corticosteroids, risk adjustment does not appear to be relevant as both variance components and ICCs did not change. This is not the case for the other four measures (diabetes-related admissions and the three types of GP expenses), for which variances components were significantly reduced. For a second group of measures (bronchodilators, statins, hospital admissions, outpatient visits, total medication expenses, and total expenses), ICCs decreased considerably after adjustment. For COPD-related admissions, an increase in ICC is observed.

7.4.3 Reliability and sample size requirements

Table 7.4 further shows the GP-level reliabilities and sample size requirements. Three measures (physiotherapy, statins, and outpatient visits) consistently have reliabilities between 0.70 and 0.80. For five other measures, reliability is between 0.80 and 0.90 [total (medication) expenses] or above 0.90 (GP-related expenses). All other measures have reliabilities below 0.70 or even 0.50. For the COPD-specific measures, this mainly is a result of small sample size, while for (diabetes-related) hospital admissions it is primarily due to low ICCs. At a reliability of 0.70, most GPs have sufficient patients for physiotherapy, GP-related expenses, and total medication expenses. To a lesser extent, this also holds for statins, outpatient visits, and total expenses. At a reliability of 0.90, there are no measures for which all GPs meet the required sample size. We also calculated the proportion of GPs with sufficient patients to reach 0.80 reliability (see Appendix 7.1). Compared with using 0.90 the proportion of GPs with sufficient sample size is larger especially for physiotherapy and statins, but these proportions decrease over time to 37 percent and 14 percent in 2008, respectively. For GP expenses, virtually all GPs meet the sample size requirements.

Figure 7.1 shows GP-specific scores and 95 percent confidence intervals for four measures for 2008. A low ICC in combination with small sample size (panel b) results in low reliability and GPs being indistinguishable from each other and from the mean. A higher ICC with the same sample size (panel a) leads to better distinguishable scores, although still few GPs can be directly compared. Sampling large numbers of patients and GPs will often result in even better distinguishable scores, even with low ICCs (panel c). However, GPs can best be profiled on measures with high ICCs and large patient samples (panel d).

7.4.4 Variability and reliability over time

Most ICCs decreased over time, especially for process measures and GP-related expenses (Table 7.4). For the process measures, this was primarily a result of reduced between-GP variation, whereas for (total) GP expenses both the within-GP and the between-GP variance decreased (but the latter more than the former). For GP medication expenses, the within-GP variance remained stable, whereas the between-GP variance decreased considerably.

As a result of the decreasing ICCs (which does not appear to be due to performance improvement, see Table 7.3), reliability and the proportion of GPs meeting sample size re-

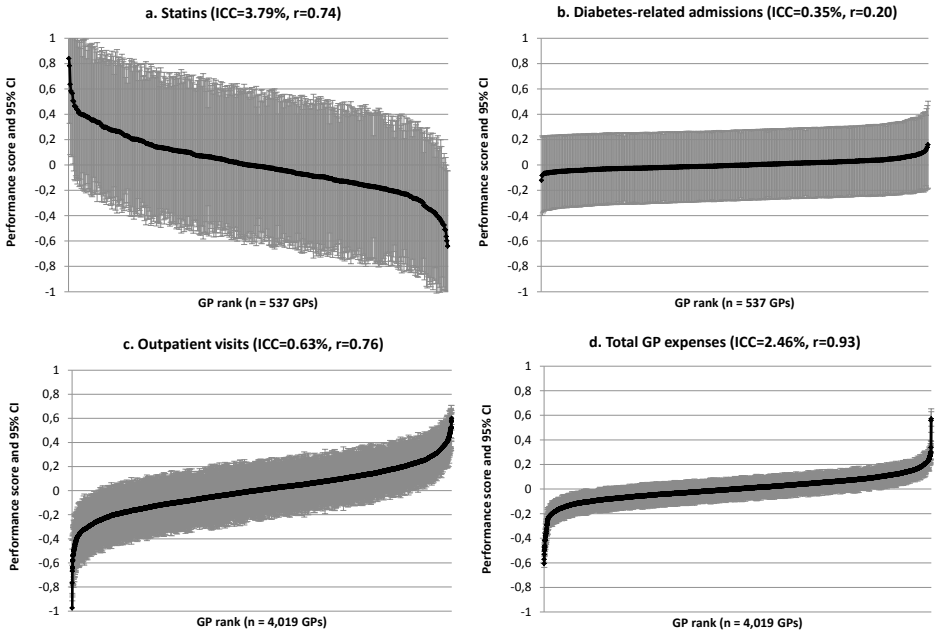


FIGURE 7.1 GP-specific performance scores and 95% confidence intervals for four measures, 2008

Note: ICC = intraclass correlation coefficient, r = GP-level reliability. The scores represent the "shrunken" estimates (GP-specific residuals) derived from the full multilevel models incorporating random GP intercepts. GPs are ranked on their performance from good to poor, resulting in a descending picture for statins (a high positive score = high performance) and an ascending picture for the other three measures (a high positive score = low performance).

quirements also decreased. For example, for GP medication expenses and total GP expenses there was a sharp decline in the proportion of GPs with sufficient sample size to reach a reliability of 0.90, despite the fact that reliability dropped two percentage points and on average remained above 0.90. This suggests that the samples sizes of many GPs were only slightly above the required sample size, and that the decline in ICC pushed them to the lower side of the threshold. For other measures (bronchodilators, statins, and outpatient visits), reliability dropped below minimum thresholds. For statins, the proportion of GPs with a sufficient sample size to reach 0.70 reliability also decreased drastically from 88.5 to 56.8 percent.

7.5 DISCUSSION

Using administrative data from a large Dutch health insurer, we examined the feasibility of conducting comparative assessments of individual GPs' performance. We found that reliable performance measurement can be feasible but is challenging, even using data from an in-

surer with substantial market share and for physicians with fixed patient panels. Adjustment for demographic and health-related patient characteristics often resulted in considerable changes in the relative magnitude of between-GP and within-GP variances, also for process measures. Finally, variability and reliability are not static figures, indicating the importance of longitudinal reliability analyses.

Of the thirteen measures, five did not achieve the minimum reliability threshold for low-stakes applications of profiling. For two of these (bronchodilators and corticosteroids), this was mainly due to the small sample size. However, because ICCs do indicate significant between-GP variation and reliabilities are only just below 0.70, it may still yield value to intervene, for example by discussing the results with GPs. For the three other measures (all related to utilization of hospital care), reliabilities were much lower, primarily due to low ICCs. Thus, variation in these measures appears to be driven largely by chance (or unobserved patient characteristics), and they are therefore unsuitable for profiling. Other measures [physiotherapy, statins, outpatient visits, and total (medication) expenses] had reliabilities between 0.70 and 0.90, making them useful primarily for low-stakes applications of profiling. For physiotherapy and statins, increasing the sample size may make them appropriate for high-stakes applications, but this will probably not be possible using single-insurer data. For outpatient visits and total (medication) expenses, effective interventions may still be worthwhile despite low ICCs (below 1.5 percent) because total variation is large. In these cases a low ICC still represents much utilization and resources in absolute terms, so effective intervention at the GP level (which can be relatively high-stakes in view of the relatively high reliability) may still translate in substantial efficiency gains.

For the three types of GP expenses), reliability was consistently estimated above 0.90. As this was a result not only of large sample sizes but also of relatively large ICCs, profiling GPs on these measures seems very feasible and useful for high-stakes applications. Yet even for these measures caution is warranted. First, many individual GPs did not meet the sample size requirements for reaching 0.90 reliability, especially for total GP expenses (this does not apply when 0.80 is used as a threshold for high-stakes efforts, see Appendix 7.1). Second, reliability may decrease further over time. Third, GPs should not be penalized for keeping patients out of the hospital and not rewarded for unjustified referrals. Finally, reliable measurement is a necessary but not sufficient precondition to prevent misclassifying GPs in performance categories (Lyratzopoulos et al., 2011; Adams et al., 2010a). For example, certain payment schemes may convert highly reliable measures into less reliable pay-out rules (Roland et al., 2009). Regardless of reliability, risk of misclassification will be highest around threshold values defining the categories, and will decrease as scores are farther away from the threshold, and with higher reliability. Therefore, it may be warranted to establish some "zone of uncertainty" around threshold values, denoting scores that cannot be confidently differentiated from the threshold (Safran et al., 2006; Kaplan et al., 2009). Narrower uncertainty zones will be required for measures with high reliability than for measures with

lower reliability. Because using multiple thresholds require multiple uncertainty zones, there is a need to limit the number of performance categories (e.g., two or three) to prevent complexity resulting from overlapping zones (Safran et al., 2006).

Our findings are consistent with previous research. Comparable studies found similar ICCs (3-4 percent) and reliabilities (between 0.65-0.90) for process measures (Selby et al., 2010; Scholle et al., 2008). Hofer et al. (1999) analyzed hospitalizations for diabetes and found an ICC of 1 percent and a reliability of 0.17. We found a slightly lower ICC, but a higher reliability due to larger sample sizes. Huang et al. (2005) analyzed variation in similar measures, but then among physician groups and for asthma. For use of bronchodilators and corticosteroids, they found ICCs of 3.1 percent and 7.6 percent and reliabilities of 0.77 and 0.89, respectively. For asthma-related hospitalizations, they found an ICC of 1.35 percent, about one percentage point lower than what we found for COPD-related hospitalizations. Lower ICCs for clinical outcomes and utilization than for process measures were also found in other studies (Sequist et al., 2011; Krein et al., 2002), although differences are not always as pronounced as we found here. Adams et al. (2010) analyzed physicians' cost profiles and found a median reliability of 0.53, which is considerably lower than what we found for total expenses. This difference will partly be a result of smaller sample sizes, but it may also be caused by different methods of constructing profiles, different definitions of the measure, and different physician samples (e.g., Adams et al. also included specialist physicians). The impact of risk adjustment has also been observed elsewhere (Rodriguez et al., 2012; Lyratzopoulos et al., 2011; Salibury et al., 2010; Turenne et al., 2008; Hofer et al., 1999), underscoring that tailored risk-adjustment is essential for a successful profiling system.

This study has several limitations. First, we defined COPD patients using the PCG for chronic nonspecific respiratory conditions and the patients' age (≥45 years). As a result, we probably overestimated the number of COPD patients in our data. Second, our set of risk-adjusters was not specifically developed for profiling but for calculating risk-adjusted capitation payments for insurers. Reliability estimates may have been different under better risk adjustment, especially for the disease-specific measures. A related concern is that some adjusters provide opportunities for gaming. For example, a GP could increase the number of patients classified in a PCG by increasing the number of defined daily doses, effectively making his patient population look sicker than it is in reality. Incentives for such behavior will be larger when results are used for high-stakes efforts. Third, our results may not generalize to other settings. We looked at a specific group of providers (Dutch GPs with fixed patient panels acting as gatekeepers) using data from one insurer. Future research should investigate whether our results are confirmed in other settings. Finally, different results may have been found with other (types of) measures, such as patient-reported outcomes (Lyratzopoulos et al., 2011; Safran et al., 2006; Solomon et al., 2002).

Despite these limitations, our findings have several policy implications. First, profiling individual GPs' performance using single-purchaser data is challenging but can be feasible

for measures more directly under GPs' control. This excludes (disease-specific) hospital admissions and sometimes also process measures because of small patient samples. Even for reliable measures, however, caution is warranted, especially when scores are used to classify GPs in performance categories. Second, reliability analyses based on cross-sectional data may not suffice as variability may vary over time. Any measure used for profiling should be periodically evaluated to see whether or not it continues to be an appropriate indicator of performance. In the US, for example, the Centers for Medicare and Medicaid Services constantly monitor whether a 12-month period and 100 completed surveys for measuring clinical outcomes and patient experience, respectively, continue to be appropriate for generating reliable hospital-level performance scores to be used for allocating incentive payments (Health and Human Services, 2011).

Third, several measures exhibiting significant between-GP variation were nevertheless deemed unsuitable for profiling because of small sample sizes. One option to increase sample size is to pool data over multiple years. Although this may enhance reliability (Adams et al., 2010; Berlowitz et al., 1998), detecting improvements will be harder and it results in less timely and thus less useful information (Solomon et al., 2002). A second option is to pool data across multiple insurers. Several examples have shown the feasibility and merits of multi-payer initiatives (Rodriguez et al., 2012; Sequist et al., 2011; Scholle et al., 2009; Safran et al., 2006; McDermott et al., 2006; Damberg et al., 2005), but have also indicated potential barriers related to antitrust regulations and technical issues (e.g., pooling data from different information systems). A third option is to create composites, which have been shown to be more reliable than individual measures (Holmboe et al., 2010; Kaplan et al., 2009; Scholle et al., 2008; Greenfield et al., 2002). However, creating composites involves difficult decisions on methods for aggregating measures and on the weights of individual measures (Reeves et al., 2007). In addition, composites may result in less actionable/useful information for providers to improve their performance and for patients to choose a provider. After all, only an overall score is reported, not the scores on the constituent items. The more individual measures are aggregated to higher levels (e.g., separate chronic-disease composites to the "chronic-care" level), the less actionable and useful the information will be for providers and patients.

Although these options may help in increasing effective sample size, it is unclear what the impact will be on between-provider and within-provider variation, which also drive reliability. In addition, increases in sample size may not always be enough to guarantee reliable profiles. Therefore, the focus should always at least be on developing and using measures displaying significant between-provider variation (Adams et al., 2010a).

In conclusion, reliable measurement of individual GPs' performance using administrative data from a single insurer is challenging. For most measures reliability was insufficient for high-stakes applications or even any application of profiling. Before measures are used for profiling, they should first be assessed on their reliability, preferably using longitudinal

data. In view of the prevailing organization and implementation of profiling – individual physicians using single-purchaser data – future research should continue to explore methods for enhancing the reliability of individual physicians' performance profiles.

APPENDIX

Appendix 7.1. Estimated variance components, intraclass correlations, reliabilities, and sample size requirements (r=0.80) for thirteen performance measures, 2006-2008

Measure	Year	Between-GP variance, unadjusted [a]	Between-GP variance, adjusted [b]	Intraclass correlation, unadjusted	Intraclass correlation, adjusted	GP-level reliability, adjusted	Actual sample size (mean)	Required sample size (r=0.80)	GPs with sufficient size (r=0.80)
	2006	0.3923	0.3923	10.65%	10.65%	0.87	58	34	91,3%
Physiotherapy (COPD)	2007	0.2837	0.2874	7.94%	8.03%	0.84	60	46	63,1%
	2008	0.2080	0.2230	5.95%	6.35%	0.80	60	59	36,5%
	2006	0.1700	0.1364	4.91%	3.98%	0.71	58	96	8,1%
Bronchodilators (COPD)	2007	0.1417	0.1223	4.13%	3.58%	0.69	60	108	6,5%
	2008	0.1240	0.1100	3.63%	3.24%	0.67	60	120	5,6%
	2006	0.1133	0.1132	3.33%	3.33%	0.67	58	116	5,6%
Corticosteroids (COPD)	2007	0.1161	0.1145	3.41%	3.36%	0.68	60	115	6,0%
	2008	0.1026	0.1035	3.02%	3.05%	0.65	60	127	4,5%
	2006	0.2105	0.1938	6.01%	5.56%	0.80	71	68	42,5%
Statins (diabetes)	2007	0.1743	0.1349	5.03%	3.94%	0.74	71	98	15,6%
	2008	0.1633	0.1297	4.73%	3.79%	0.74	71	101	14,2%
	2006	0.0017	0.0011	0.02%	1.64%	0.49	58	240	0,2%
COPD-related admissions	2007	0.1129	0.1082	1.28%	3.25%	0.67	60	119	5,4%
	2008	0.1149	0.0736	1.34%	2.57%	0.61	60	152	2,0%
	2006	0.0495	0.0007	0.65%	0.54%	0.28	71	736	0,0%
Diabetes-related admissions	2007	0.0612	0.0404	0.71%	0.95%	0.40	71	419	0,0%
	2008	0.0241	0.0172	0.23%	0.35%	0.20	71	1151	0,0%

Measure	Year	Between-GP variance, unadjusted [a]	Between-GP variance, adjusted [b]	Intraclass correlation, unadjusted	Intraclass correlation, adjusted	GP-level reliability, adjusted	Actual sample size (mean)	Required sample size (r=0.80)	GPs with sufficient size (r=0.80)
All hospital admissions	2006	0.0002	0.0000	0.29%	0.08%	0.28	515	5197	0,1%
	2007	0.0002	0.0000	0.30%	0.07%	0.27	507	5551	0,0%
	2008	0.0002	0.0000	0.29%	0.07%	0.28	512	5392	0,1%
Outpatient visits	2006	0.0030	0.0014	1.35%	0.80%	0.80	515	499	25,3%
	2007	0.0029	0.0014	1.32%	0.74%	0.79	507	535	22,5%
	2008	0.0016	0.0008	1.00%	0.63%	0.76	512	634	19,3%
GP expenses	2006	0.0120	0.0086	4.16%	4.28%	0.96	515	89	100,0%
	2007	0.0102	0.0067	3.56%	3.30%	0.95	507	117	98,0%
	2008	0.0106	0.0076	3.70%	3.96%	0.95	512	97	100,0%
GP medication expenses	2006	0.2667	0.1781	4.37%	4.90%	0.96	515	78	100,0%
	2007	0.2722	0.1768	4.46%	4.97%	0.96	507	77	100,0%
	2008	0.1777	0.0991	2.98%	2.95%	0.94	512	131	95,4%
Total medication expenses	2006	0.1681	0.0544	2.62%	1.56%	0.89	515	253	65,7%
	2007	0.1679	0.0559	2.65%	1.63%	0.89	507	242	68,8%
	2008	0.1317	0.0408	2.12%	1.22%	0.86	512	323	50,4%
Total GP expenses	2006	0.0330	0.0149	3.45%	3.30%	0.95	515	117	97,7%
	2007	0.0329	0.0135	3.40%	3.01%	0.94	507	129	95,9%
	2008	0.0271	0.0105	2.82%	2.46%	0.93	512	159	89,7%
Total expenses	2006	0.0453	0.0116	1.97%	0.87%	0.82	515	454	29,5%
	2007	0.0451	0.0122	2.03%	0.87%	0.82	507	454	29,1%
	2008	0.0392	0.0111	1.73%	0.78%	0.80	512	512	24,4%

a. All unadjusted variance components statistically significant at 1 percent significance level, except for diabetes-related admissions in 2008 (P=0.08).

b. All adjusted variance components statistically significant at 1 percent significance level, except for diabetes-related admissions in 2007 (P=0.01) and 2008 (P=0.14).

CHAPTER 8

Conclusions and discussion

This chapter first summarizes the main findings by answering the research questions formulated in the introduction. Next, the relevance of these findings for the Dutch healthcare system is discussed. Finally, some suggestions for further research are provided.

8.1 BACKGROUND AND ANSWERS TO THE RESEARCH QUESTIONS

Healthcare systems around the world are characterized by a suboptimal delivery of healthcare services. There has been a growing belief among policymakers that many deficiencies (e.g., in the quality of care) stem from flawed provider payment systems creating perverse incentives for healthcare providers. In several countries this has led to reforms based on *pay-for-performance* (P4P), a payment approach in which healthcare providers receive explicit financial incentives to improve the quality and efficiency of care. Over the past decade, P4P has attracted widespread interest, with programs being uncritically implemented in many countries. Because healthcare providers respond to financial incentives and because performance measurements have become increasingly sophisticated, many policymakers view P4P as a promising improvement strategy.

A theoretical basis for applying financial incentives to improve the quality and efficiency in health care can be found in agency theory, which studies contractual relationships characterized by information asymmetry, conflicting interests, and outcome uncertainty. The theory offers several strategies the relatively ill-informed party (the principal) could apply to prevent the relatively well-informed party (the agent) from exploiting his information surplus. One of these is incentives. The choice for a particular incentive scheme depends on the information possessed by the principal (e.g., a purchaser of care) on the outcome and on the agent's (i.e., the healthcare provider) efforts. In health care, *multitasking* implies that payments for providers will always consist of a base component that is unrelated to performance. Yet all base payment systems have shortcomings, which can be mitigated by supplementing the base component with a P4P element.

In contrast to what the widespread interest in P4P suggests, its effectiveness has not been convincingly demonstrated. In part, this may be due to the limited knowledge about crucial aspects of the design and implementation of P4P. In addition, the evidence on effects of P4P has become fragmented and thus difficult to comprehend and use. This goal of this thesis was to address these issues by analyzing key conceptual and practical issues in the design and implementation of P4P, by synthesizing empirical literature on effects of P4P, and by addressing important empirical questions about the complex issue of performance measurement. In addition, based on the findings several recommendations were provided. The thesis consists of three main parts: design of P4P in theory and practice, effects of P4P, and statistical issues in performance measurement.

8.1.1 Design of pay-for-performance in theory and practice

As noted by many commentators, the lack of convincing evidence on the effectiveness of P4P may have partly been a result of flaws in the design of current P4P-programs. Despite over a decade of experimenting with P4P, little is still known about which design features contribute to (un)desired effects. Given that the interest in P4P is unlikely to diminish in the coming years, knowledge of crucial design features is required. In this respect, insight is also necessary in how P4P is currently being designed in practice and in the extent to which this design is adequate given the specific context of implementation. These issues led to the first two research questions:

Q1. What are crucial design features of a successful P4P-program?
Q2. How is P4P currently being designed in practice and to what extent is this design adequate?

Question 1 was addressed in *chapter 2* by identifying and analyzing relevant theoretical work and (empirical) literature on P4P-program design. The analysis revealed several key design features, which were classified in three categories: what to incentivize (definition/measurement of performance, risk adjustment, provider engagement), whom to incentivize (individual physicians vs. groups of physicians), and how to incentivize (rewards vs. penalties, payment size, number and type of performance targets, payment frequency, program duration). Although the idea underlying P4P is simple, designing a fair and effective P4P-program is a complex undertaking requiring consideration of many interrelated aspects and potential pitfalls. Strong conclusions on adequate P4P-program design were not possible because (1) whether or not a specific P4P-program has been designed adequately depends on the context of implementation, (2) given a particular context, appropriate choices regarding certain design elements may conflict, (3) practical difficulties, specifically regarding availability of accurate relevant data, may impede adequate design, (4) there are limitations in the interpretation of the theories used for predicting provider behavior, and (5) empirical evidence on the influence of specific design choices in practice is virtually absent. Nonetheless, several tentative conclusions could be drawn, which are summarized in Table 8.1.

Using this categorization, question 2 was addressed in *chapter 3* by reviewing major P4P-programs implemented throughout the world. Since several overviews of the design of P4P in the United States (US) were already available, the methodology focused on identifying programs implemented outside the US. In total, thirteen programs from nine different countries were identified, eight of which were initiated by a public purchaser and five by private health insurers. All programs incentivize clinical quality and most only use positive incentives, actively involve providers in the design, target (groups of) primary care providers, and pay on an annual basis. However, we also observed considerable heterogeneity, particularly regarding the performance measures, use of risk-mitigating measures, payment size, and number and type of performance targets.

TABLE 8.1 Key elements of adequate P4P-program design and implementation

What to incentivize
- Performance is ideally defined broadly, provided that the set of measures remains comprehensible
- Concerns that P4P may encourage "risk selection" and "teaching to the test" should not be dismissed
- Outcome and resource use measures should only be used with adequate risk adjustment and sufficient sample size
- In addition to risk adjustment, other strategies to mitigate incentives for risk selection may still be necessary
- Measure sets should at least incorporate "high-impact" measures. The less technical / more indeterminate aspects of care such as patient satisfaction and continuity of care are ideally also included or at least regularly monitored
- P4P incentives should be aligned with professional norms and values; it is therefore vital that providers are actively involved in program design and in developing, selecting, and maintaining the performance measures
- Monitoring, feedback, and information technology are important in preventing undesired provider behavior

Whom to incentivize
- Group-level incentives will typically be preferred over individual-level incentives mainly because performance profiles are more likely to be statistically reliable as a result of larger numbers of patients
- Individual-level or small-group incentives as well as using measures with small available samples sizes will become increasingly feasible as methods for constructing composite performance scores continue to evolve
- One should be cautious with applying schemes incorporating both individual- and group-level incentives
- Participation is ideally voluntary provided that broad participation among eligible providers can be realized

How to incentivize
- Whether rewards or penalties should be used is context-dependent. Offering providers a choice among schemes also including penalties may be a viable option
- Although increasing the size of the incentive increases its strength (up to a certain point), relatively low-powered incentives are preferred, provided that providers' opportunity costs of improving performance are covered
- Provider-specific absolute performance targets and/or a uniform series of absolute targets, possibly combined with piece-rates for each appropriately managed patient, are preferred over single targets and relative targets
- The time lag between care delivery and payment should be minimized
- P4P should be a permanent component of provider compensation, but is ideally decoupled from base payments
- Performance measures should be reevaluated periodically and regularly be replaced or updated (as necessary)

The programs share several design features with the typical P4P-program in the US: clinical quality is most commonly incentivized and typically gets most weight; measure sets tend to be quite small; outcome measures are not often used and when they are, they pertain to similar aspects; engagement of providers is considered a critical success factor; most programs target physician groups in primary care; and payments are usually made on an annual basis. There are notable differences as well. Programs in the US rely more on efficiency measures and measures related to adoption of information technology (IT). In addition, negative incentives are more often used in the US. Also, although payment size appears to be similar for physicians, for hospitals non-US programs generally apply more generous payments. Finally, relative targets are more often found in the US.

Among the identified programs there seems to be ample room to enhance incentives for desired behavior and to mitigate incentives for undesired behavior. Shortcomings mainly pertain to the number and type of performance measures, risk adjustment for outcome measures and resource use, measurement reliability, payment frequency, and number of performance targets. Conversely, for some aspects design does seem adequate in most programs, including provider involvement, type of performance targets, and voluntary participation. Overall, however, the conclusion seems justified that many current P4P-

programs have shortcomings, especially regarding design elements related to preventing risk selection and a disproportionate focus on performance aspects that are measured and rewarded ("teaching to the test"). The variation in the use of risk-mitigating measures indicates that purchasers, though clearly concerned about them, are uncertain about how to effectively prevent undesired effects. Therefore, more insight is required in how these can be prevented. In addition, many shortcomings in the design, including low payment frequencies, small sets of measures, limited use of outcomes, and lack of risk adjustment, can be traced back to a lack of accurate relevant data. Therefore, efforts should continue to focus on implementing sophisticated IT for recording, extracting, and exchanging patient-level data. Furthermore, current programs typically focus on one specific sector and/or type of provider. However, improved patient outcomes and efficiency require coordination across sectors and alignment of incentives for all providers in the continuum of care. Customized IT and forms of prospective bundled payment could be useful in attaining this goal. Finally, although the observed heterogeneity in design will partly be a reflection of contextual differences, it also may be the result of practical implementation difficulties and/or the limited knowledge about what works when it comes to implementing P4P in practice. Sound empirical research on the influence of design choices is therefore needed.

8.1.2 Effects of pay-for-performance

Along with the interest in P4P, the literature on effects of P4P has expanded rapidly over the past two decades. However, the evidence has become fragmented. Several reviews have attempted to synthesize the evidence, but they often had different foci and hence different conclusions. Consequently, it remains challenging to comprehend this literature and to extract success factors and pitfalls when it comes to implementing P4P. In addition, literature reviews have typically overlooked a crucial aspect of P4P performance: cost-effectiveness. Although high-quality care is clearly an important goal, resources are scarce and ideally allocated to improvement efforts yielding most value for money. These issues led to two additional research questions:

Q3. What is the current state of evidence on the cost-effectiveness of P4P?
Q4. What is the current state of evidence on effects of P4P?

Chapter 4 addressed question 3 by systematically reviewing the literature on the cost-effectiveness of P4P. The main focus was on identifying full economic evaluations that simultaneously consider costs and effects of the P4P intervention (Type I). Partial evaluations with separate effectiveness and cost assessments were labeled Type II. Simple cost comparisons were also included and divided in studies which also provide information (not derived from the study itself) on how quality has likely developed (Type III) and studies only providing information on the financial impact (Type IV). Nine studies were included:

three Type I, four Type II, and two Type IV. Eight studies were conducted in the US, five of which were initiated by private purchasers. Studies typically adopted the insurer perspective in assessing costs and effects.

The results show that P4P can potentially be cost-effective, although the evidence is not convincing. Type I studies found improvements in quality against increases in costs. However, these studies only examined program costs, not the *impact* on healthcare costs. In addition, two studies evaluated only one process measure, and the third study lacked a convincing control group. Type II studies assessed broader sets of measures and more often analyzed changes in outcomes. Two studies examined program costs and the impact on costs, but details were not provided. Two studies, one of which did not assess program costs, found quality improvements and cost increases. Another study showed that savings may be possible while quality improves, while still another (not assessing program costs) observed neither reduced mortality rates nor cost savings. The two Type IV studies both reported cost savings.

The literature on P4P cost-effectiveness has important limitations. Studies typically failed to assess all relevant types of costs or did not report in detail about them, and often suffer from methodological flaws such as lack of convincing control groups. In addition, studies vary greatly in focus. While some studies evaluated programs with a time frame of less than a year, others analyzed effects over several years. Three programs targeted a chronic disease, one study focused on acute care, and five studies focused on prevention. Furthermore, the design of evaluated P4P-programs varied considerably, although details were often not provided. For these reasons, the evidence does not allow for a definitive conclusion about the cost-effectiveness of P4P. Longitudinal evaluations with broad ranges of costs and effects are needed to expand the evidence base.

Question 4 was dealt with in *chapter 5*. A systematic search in five literature databases identified 22 systematic reviews analyzing evidence on a wide variety of effects and mediating factors, that is, effectiveness, cost-effectiveness, unintended consequences, impact on inequalities in quality of care among specific population subgroups, and the influence of non-financial incentives and specific design features. Some reviews focused on one or several conditions or on one specific sector. Others only included studies with a particular design while still others had no restrictions. Most reviews only included studies from the US and the UK, but studies from other countries have increasingly been identified. Most studies were conducted in primary care, although P4P in other sectors (e.g., inpatient care) has increasingly been evaluated.

Regarding *effectiveness*, most studies focused on prevention and/or chronic care provision in primary care. Findings of the few randomized controlled trials provide a mixed picture, justifying the conclusion that there is insufficient evidence to support or not support the use of P4P to improve performance for these aspects. Non-randomized studies

have typically found improvements in at least one measure, but results from studies with relatively strong designs tend to be less positive than results from studies with weaker designs. A notable finding was that the QOF (the world's largest P4P-program in terms of performance measures and bonus potential, see the description in the introduction) has generally not been able to accelerate improvements in quality that were already occurring. Overall, however, the impact of physician P4P was estimated at 5 percent improvement in incentivized performance. The reviews further highlight P4P's potential to be *cost-effective*, but the evidence is too thin to draw strong conclusions (as already concluded in chapter 4). Regarding *unintended consequences*, several studies have found evidence of risk selection behavior, such as older patients and patients with greater disease severity being more likely to be excluded from a P4P-program than younger/healthier patients. In addition, there is some evidence of negative spillover effects on unrewarded aspects of performance, with several studies finding reductions in continuity of care and less improvement for excluded conditions than for included conditions. Evidence on gaming behavior and negative effects on providers' intrinsic motivation is virtually absent. Regarding *inequalities*, P4P seems to have narrowed socioeconomic inequalities in the UK. Yet inequalities related to age, sex, and ethnicity have largely persisted, although there were small attenuations for some measures. Regarding *non-financial incentives*, there is some weak evidence that feedback alone can result in improvement and that P4P does not add much when feedback is already provided. Conversely, while public reporting alone can stimulate quality improvement activity in hospitals, some findings indicate that more favorable results can be achieved when public reporting is combined with P4P, although these findings only pertain to the short-term impact on process quality. Finally, the results confirm that *program design* matters. Although the evidence is only suggestive, P4P seems to have been more effective when: (1) measures are used with much improvement potential instead of measures with little improvement potential, (2) directed at individual physicians or small groups instead of larger groups or institutions, (3) payments are based on providers' absolute performance instead of relative performance, (4) designed collaboratively with providers instead of imposed top-down, and (5) larger payments are used.

Overall, although many studies have found improvements in selected quality measures and demonstrated that P4P can potentially be effective, at this point the evidence seems insufficient to recommend widespread implementation of P4P. Convincing evidence is still lacking (the majority of studies are observational studies), especially for inpatient care. To facilitate evidence-based policy-making on P4P, it is important that multifaceted improvement strategies are implemented in the context of rigorous evaluation, using convincing control groups to disentangle the effects of the different components. These evaluations should also assess the long-term impact on health outcomes and costs, which thus far has largely been ignored. In addition, although the results indicate that design matters, few if any studies have specifically addressed design features. Furthermore, unintended effects

emphasize the importance of ongoing monitoring and gaining more insight in how specific design features may help in mitigating incentives for undesired behavior. Finally, although it is reassuring that P4P does not seem to have widened inequalities, many inequalities have persisted and most studies only relied on cross-sectional data. Most programs are not designed to address inequalities or lack important features that would enable them to reduce inequalities. Rewarding improvement in performance and/or directly rewarding reductions in inequalities seem good options to improve current programs.

8.1.3 Statistical issues in performance measurement

In health care, provider performance measurements may be particularly sensitive to random chance and certain patient characteristics and behaviors. To account for that, an appropriate statistical model is essential. Various models are available for analyzing and risk-adjusting performance differences among providers. In practice, purchasers prefer relatively simple models that are easy to implement, maintain, and explain to providers. However, performance data in health care have specific features rendering simple models largely unsuitable for modeling these data. Nonetheless, if patient-level modeling results are aggregated to the provider-level, simple models may well yield similar performance assessment results compared to more appropriate sophisticated models.

In addition to adequate risk adjustment to prevent systematic misclassification of providers due to differences in casemix, measurements require adequate reliability to prevent random misclassification of providers due to chance. When measurements have low reliability, P4P incentives based on them may arbitrarily and unfairly penalize or reward providers. Reliable discrimination of provider performance requires a sufficient number of patients per provider as well as sufficient variation between providers relative to variation within providers. In practice, comparative performance assessments are often focused on individual physicians and mainly performed by individual private purchasers. Yet when single-purchaser data are used to compare individual physicians on their performance, adequate reliability is particularly uncertain due to small sample sizes. These issues led to the final two research questions:

Q5: To what extent does the choice of statistical model used for risk adjustment affect the results of comparative provider performance assessments?
Q6. To what extent can individual physicians be reliably compared with respect to their performance on measures derived from the administrative data of a single private care purchaser?

Both questions were addressed using member-level administrative data from a large Dutch health insurer. Regarding question 5, *chapter 6* empirically examined the extent to which a variety of statistical models produce different rankings of Dutch GPs and health centers regarding their performance on several measures of quality and resource use. The analysis

revealed that, holding constant the set of risk adjusters and definition of performance index, the choice of statistical model does seem to matter, especially for outcome measures and expenses. However, differences were small, and possibly small enough for purchasers to opt for a simple method like ordinary least squares. However, caution is warranted because despite relatively high agreement among risk-adjusted rankings, the statistical models still often classified providers in different performance categories. Also, agreement on outlier designation was lower and more variable compared to agreement on rankings overall. This was the case especially for high outliers (i.e., high performers), which is an important finding because most P4P-programs only reward high performance (as found in chapter 3). Another finding is that regardless of the model used, provider rankings varied considerably over time, especially for hospital admissions and total expenses. As this variation is unlikely to be a result of a specific intervention, it probably (partly) reflects random variation.

It should be noted that our set of risk adjusters was not developed for conducting provider performance assessments but for calculating capitation payments for health insurers. Consequently, the risk adjusters were best suited for the expenses measures, while for the other measures (e.g., clinical quality measures) much of the patient-level variation remained unexplained. Also, some risk adjusters can be directly influenced by providers and thus provide opportunities for gaming. Thus, although there is no reason to believe our results would have been different under better risk adjustment, development of measure-specific risk-adjustment models merits high priority.

Regarding question 6, *chapter 7* empirically assessed the reliability of performance measurements for GPs by analyzing GP-level variation in performance across multiple years. The results showed that reliable performance measurement using single-purchaser administrative data is challenging but can be feasible for measures directly under physicians' control. Another important finding was that adjustment for sociodemographic and health-related patient characteristics often resulted in considerable changes in the relative magnitude of between-GP and within-GP variances (even for process measures), again underscoring that risk-adjustment is essential for a successful performance measurement and comparison system. Also, the results clearly showed that variability and reliability are not static figures, indicating the importance of repeated reliability analyses over time.

Of the thirteen analyzed measures, five did not achieve the minimum reliability threshold agreed upon in the literature (i.e., 0.70) to allow for comparative performance assessment. For two of these (both process quality measures), this was mainly due to the small sample size, while for three measures (all related to utilization of hospital care) it was mainly due to limited between-GP variation relative to within-GP variation. Five other measures (two process and three resource use measures) had reliabilities between 0.70 and 0.90, making them useful primarily for "low-stakes" applications of performance assessment such as providing providers with feedback on their performance. Finally, for

three measures (GP expenses excluding medications, GP medication expenses, and total GP expenses) reliability was consistently above 0.90, which was not only a result of large sample sizes but also of significant between-GP variation. However, even for these measures caution is warranted when used for P4P because: (1) many individual GPs did not meet the required sample size for reaching 0.90 reliability, (2) the results showed that reliability may decrease over time, (3) GPs should not be penalized for keeping patients out of the hospital and not rewarded for unwarranted referrals (cost shifting), (4) and reliable measurement is a necessary but not sufficient precondition to prevent misclassifying GPs in performance categories (e.g., low, medium, high).

Several measures exhibiting significant between-GP variation were nevertheless deemed unsuitable for comparative performance assessment because of small sample sizes. Although there are useful options to increase sample size (e.g., pooling data over multiple years and/or across multiple purchasers, computing composite scores), they are not without limitations. In addition, it is unclear beforehand how these strategies will impact on between- and within-provider variation, both of which also drive reliability. Moreover, for many measures increases in sample size will not be sufficient to guarantee reliable comparison. Therefore, the focus should always at least be on developing and using measures exhibiting significant nonrandom between-provider variation.

8.2 RELEVANCE FOR THE NETHERLANDS

Pay-for-performance is also being applied in the Netherlands. Experiences with P4P in the Dutch healthcare system date back to the 1980s. In 1984, two sickness funds assessed whether financial incentives could facilitate efficient substitution from hospital care to primary care (van Tits & Nuyens, 1987; van Tits, 1989). GP practices were incentivized to reduce the number of referrals to medical specialists as well as the number of inpatient days and costs of prescription medication. Participating practices could earn bonuses of maximally 30 percent of generated savings. Since this experiment, various government-appointed committees recommended the use of explicit financial incentives to increase efficiency in health care. In 1994 the Biesheuvel-committee advocated for providing efficiency-markups to GPs succeeding in reducing unnecessary prescription of medication, referrals, and diagnostic tests (committee Modernisering curatieve zorg, 1994; Vermaas, 1995). A subsequent committee also emphasized the need for such incentives, and further argued that the markups should also be based on quality of care, effective coordination of care, and IT adoption (committee Toekomstige financiering huisartsenzorg, 2001). As a result, several health insurers initiated small-scale P4P experiments. In 2001, for example, CZ employed a combined P4P-feedback program for 120 GPs to promote efficient prescribing behavior (Gruisen & Muijrers, 2002).

The reforms toward regulated competition in the Dutch healthcare sector enacted during the previous two decades have provided health insurers with additional incentives (e.g., increased financial risk) and tools (e.g., the possibility to selectively contract with providers, the functional description of covered benefits, the gradual abolition of price regulation) for managing the care. Insurers are expected to use these tools to act as prudent purchasers of care on behalf of their insured clients. Recent initiatives show that insurers are slowly taking up this role, among others by implementing P4P via differentiated provider contracts. For example, VGZ and CZ initiated a P4P experiment in primary care, incentivizing GPs to score well on measures of clinical quality, patient experience, and organizational aspects (Kirschner et al., 2008, 2009; Braspenning et al., 2008). Other insurers provide bonuses to GPs for efficient prescribing behavior (e.g., Habets et al., 2009; IVM, 2012), and programs are being implemented for health centers, pharmacies, physiotherapists (Eijkenaar & Edgar, 2012), and recently also hospitals. Regarding the latter, Achmea has concluded a five-year contract with the Zaans Medisch Centrum (a medium-sized general hospital), incorporating bonuses and financial penalties regarding attainment of targets for cost savings and mortality rates (Het Financieele Dagblad, 2011).

To date, however, P4P in the Dutch healthcare system has been limited to temporary small-scale experiments, usually in primary care and typically with minor financial consequences for providers. In June 2011, the Council for Public Health and Care (RVZ) advocated for a more widespread application of financial incentives to promote improvements in quality, characterizing P4P as an attractive and flexible instrument for insurers to use in their purchasing policy (RVZ, 2011). As argued by the RVZ, such incentives are desirable given the inefficiencies in the Dutch healthcare system and the fact that current payment systems do not contain incentives for an efficient delivery of high-quality care (the more care a provider delivers, the more he gets paid, irrespective of the quality with which the care was provided and the outcome for the patient). The current Minister of Health (E. Schippers) underscores the need for payment reform, and has announced that in the coming years she will strongly facilitate the development and implementation of innovative payment methods designed to stimulate good outcomes in terms of both quality and costs (Tweede Kamer, 2012). As this thesis has demonstrated, however, expectations regarding the returns of P4P should be tempered. Designing a successful P4P-program is highly complex, and empirical studies on the (cost-)effectiveness of P4P have repeatedly shown disappointing results. In addition, little is still known about which specific configuration of P4P works best in a given context. Although there are certainly examples of positive experiences with P4P, the extent to which these results can also be achieved in the Dutch context is unclear (Stevens & Shojania, 2011; Sutton et al., 2012).

For example, a crucial precondition for successfully implementing P4P is transparency in the quality of care. In the Netherlands, this does not yet seem to be sufficiently the case (Algemene Rekenkamer, 2013). An evaluation of a recent P4P-program showed that GPs'

information systems provided insufficient support for generating required performance data (Kirschner et al., 2009). In addition, evaluation of the bundled payment pilots for chronic conditions showed that existing IT systems (which are essential for generating reliable performance data and for facilitating efficient coordination of care) have important limitations, including difficulties with integrating data systems from different sectors. Moreover, transparency in quality within care groups (i.e., organizations that typically act as "main contractor" in the negotiations with health insurers, receive bundled payments, and contract with individual providers that provide the care) is evolving slowly, and interpretation of reported performance remains difficult due to differences in casemix and technical problems with data registration and extraction (EIB, 2012).

This does not mean that there has not been any progress. The Dutch Association of Health Insurers (ZN) is actively working on making quality of care transparent, for example by developing sector-specific "purchasing guides" that can be used by insurers in their contract negotiations with providers. Another initiative is the Routine Outcome Monitoring in the mental health sector. In this program, based on an agreement between ZN and the Dutch Association of Mental Health Care (GGZ Nederland), providers measure outcomes and patient experiences before, during, and after treatment using standardized measurement instruments. Insurers and providers already make binding (financial) agreements about data collection and reporting. It is expected that in 2014 insurers will begin with paying providers based on outcomes and patient experience (RVZ, 2011). Transparency is also being realized via the Zichtbare Zorg (ZiZo) program, which was enacted in 2007 by the government and provides support to providers, insurers, and patient associations in generating valid and reliable information on quality. For hospital care, for example, quality information can now be generated for 40 different conditions. However, in 2012 only 14 percent of the 333 measures included in these sets were outcomes (NZa, 2012). One reason for this is the lack of adequate risk adjustment. In addition, many measures are not reliable enough to enable accurate comparison of hospitals. As part of ZiZo, studies are being conducted to examine how the reliability of quality information can be increased (e.g., Koolman et al., 2011). If these efforts are successful and as clinical data on patient characteristics increasingly become available in electronic format, comparative performance assessments for providers will become increasingly feasible. A final example is the annual "monitor prescribing behavior GPs" published by the Institute for Appropriate use of Medications (IVM). In this monitor, the quality (i.e., adherence to guidelines of the Dutch Society of General Practitioners) and efficiency (i.e., choosing the least expensive drug in case of equal efficacy) of GPs' prescribing behavior are made transparent for 28 validated prescribing measures. Scores on these measures and differences therein are calculated on the national, regional, insurer, and GP practice level. A major advantage of this initiative is that the measures are calculated using pooled claims data for the entire patient populations of all GPs (not only for a fraction of this population, which would be the case if insurers

would use their own claims data to calculate the measures for their own members only). As shown in the most recent monitor, the quality and efficiency of GPs' prescribing behavior can be significantly improved (IVM, 2012).

Notwithstanding this progress, health insurers may remain cautious with investing in performance measures and quality improvement because competitors can often benefit from these investments without incurring the costs (the free-rider problem). It is therefore important that efforts continue to focus on generating standardized measures sets – in a coordinated way with ongoing input from providers, insurers, and patient organizations – that in principle can be used by all insurers and providers. The monitor prescribing behavior is a good example of how this can be realized in practice (IVM, 2012). This also increases the likelihood of provider support because providers are less likely to be confronted with different quality requirements from different insurers. In this respect, significant investments in IT systems will be required for recording and collating the data for measuring and risk-adjusting the measures. The new Quality Institute (part of the Health Care Insurance Board, in 2013 renamed Care Institute the Netherlands) could play an important facilitating role in this by combining the knowledge obtained via ZiZo and other quality initiatives and institutions.

Finally, breakthrough improvements in quality and efficiency require effective coordination of care across sectors, which in turn requires alignment of financial incentives for all providers in the continuum of care. Current P4P-programs tend to be implemented within a specific sector, while different types of providers continue to work and be paid in silos. In the Netherlands, the "fence" between primary care and hospital care has been one of the most important barriers for an efficient delivery of care. This fence not only refers to the different payment systems, but also to the different medical records and protocols, the different organizations, and a general lack of coordination and continuity of care. Prospective bundled payment combined with effective P4P seems a promising strategy to achieve the desired coordination (Rosenthal, 2007a). There has already been some progress in this regard, for example with the introduction of payments to hospitals and medical specialists per diagnosis-treatment-combination (DBC) care-product, and the payments to care groups per "Chain-DBC" for chronic conditions (De Bakker et al., 2012). Nonetheless, payment reform is still in its infancy. Major challenges for the coming years are bridging the gap toward *integral* payments (i.e., payment for care provided by all relevant providers), and making the step from payment per *patient* to payment per *insured* (see Table 1.1). In this respect, attention should be paid to experiences from other countries, such as the US, the UK, and Germany (Eijkenaar et al., 2012). These experiences could be useful in answering important questions regarding implementation of bundled payments and P4P in the Netherlands, such as: who takes the lead? How to reach consensus with and among healthcare providers about performance measurement, program design, and the desired modes of coordination of

care? Who invests in the necessary IT systems and electronic medical records? Who collects the performance data and who pays for that? To what extent do insurers and providers have sufficient data to measure performance in a reliable and valid way? How detailed should the risk adjustment be? And how to deal with specific issues in the context of regulated competition in health care, such as possible violations of antitrust legislation and the free-rider problem for health insurers?

8.3 SUGGESTIONS FOR FURTHER RESEARCH

As this thesis has shown, designing a successful P4P-program is a complex undertaking that requires consideration of many interrelated aspects, contextual factors, and potential pitfalls. In part, the disappointing success of P4P is a reflection of this complexity and the fact that we still know little about which specific configuration works best in a given setting. Although the (theoretical) analysis in chapter 3 provides useful directions, the paucity of empirical evidence on the effect of specific design choices hampered strong conclusions on adequate P4P-program design. Quantitative and qualitative empirical research is therefore necessary to gain more insight in:

- Which specific design elements contribute to desired effects;
- The relative importance of different design elements in reaching these effects;
- Adequate design of specific elements, such as the appropriate payment size for individual physicians, groups of physicians, and organizations (e.g., hospitals);
- How unintended consequences such as risk selection, teaching to the test, and widening inequalities in quality of care among specific population subgroups can be prevented;
- How these results vary according to relevant contextual factors, such as providers' base payment system, the organization of care purchasing, and regulation.

To get the most out of P4P, it is crucially important that programs are implemented in the context of rigorous evaluation, using convincing control groups to disentangle the P4P effect from the effect of concurrent improvement efforts. The conclusion drawn in chapter 5 (that the evidence on effects of P4P is insufficient to recommend widespread implementation of P4P) is partly a result of the difficult to interpret findings from the many observational studies. New P4P-programs should first be pilot-tested among a selection of providers before they are implemented on a large scale. This enables comparison of changes in performance between participants and non-participants. Such pilots also allow for experimenting with various design options, providing input for answering the questions above. In addition, program evaluations should also assess the long-term impact on health outcomes and costs, which thus far has largely been ignored. To allow for rigorous cost-effectiveness analysis, costs should be defined broadly, including at least the (transaction) costs of program administration and the impact on healthcare costs. Furthermore, evaluations should contain

quantitative and qualitative studies on unintended effects such as neglect of unrewarded aspects of performance, risk selection, diminished intrinsic motivation of providers, and widening inequalities among specific population subgroups. Finally, researchers should be thorough in reporting the results (e.g., regarding P4P-program design, relevant contextual factors, etc.) to facilitate external scrutiny and to provide possibilities for policymakers in other settings to learn from the results.

Regarding performance measurement, this thesis has revealed several insights that are useful for implementing P4P in practice. However, additional research is necessary to confirm the conclusions. First, the results of our empirical analyses may not generalize to other settings and measures. We looked at a specific group of providers (Dutch gatekeeping GPs with fixed patient panels) using administrative data from one insurer. Given the widespread use of comparative performance assessment and P4P, further research should investigate whether the results are confirmed in other settings (e.g., other providers, other data sources, etc.) and for other performance measures for the same as well as other conditions. Second, more insight is required in the merits and drawbacks of methods and strategies for improving reliability, such as combining different data sources, creating composite scores, aggregating individual physicians into practice groups, aggregating data over multiple time periods, and/or aggregating data over multiple (competing) purchasers. Finally, our set of risk adjusters was not developed for comparative provider performance assessment but for explaining patient-level variation in costs. Consequently, more than 90 percent of the patient-level variation in the quality measures remained unexplained. Also, some of the risk adjusters (e.g., the pharmacy-based cost groups, which are based on the prior use of prescription medication) can be directly influenced by providers and thus provide opportunities for gaming. Incentives for such behavior may be especially large when performance assessment results have consequences for providers' income. Thus, research on the development of customized risk-adjustment models merits high priority. In this respect, more insight is also required in how such models can be effectively implemented; although risk adjustment for health outcomes has become increasingly sophisticated (in part resulting from the fact that more clinical data have become available in electronic format), there is still a lot to learn about how they can be applied transparently to prevent providers from viewing it as an arbitrary black box and from being suspicious of its validity.

REFERENCES

Adams, J.L., Mehrotra, A., Thomas, J.W., McGlynn, E.A. 2010a. Physician cost profiling--reliability and risk of misclassification. *The New England Journal of Medicine* 362(11): 1014-1021.

Adams, J.L., McGlynn, E., Thomas, J.W., Mehrotra, A. (2010). Incorporating statistical uncertainty in the use of physician cost profiles. *BMC Health Services Research* 10: 57.

AHRQ. 2001. *AHRQ Quality Indicators – Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions*. Publication number 02-R0203. Rockville: Agency for Healthcare Research and Quality.

Alchian, A.A., Demsetz, H. 1972. Production, information costs, and economic organization. *The American Economic Review* 62(5): 777-795.

Algemene Rekenkamer. 2013. Indicatoren voor kwaliteit in de zorg. Den Haag: Sdu Uitgevers.

Alshamsan, R., A. Majeed, M. Ashworth, J. Car, Millett, C. 2010. Impact of pay for performance on inequalities in health care: systematic review. *Journal of Health Services Research & Policy* 15(3): 178-84.

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

An, L.C., Bluhm, J.H., Foldes, S.S., Alesci, N.L., Klatt, C.M., Center, B.A., Nersesian, W.S., Larson, M.E., Ahluwalia, J., Manley, M. 2008. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. *Archives of Internal Medicine* 168: 1993-1999.

Anderson, K.K., Sebaldt, R.J., Lohfeld, L., Burgess, K., Donald, F.C., Kaczorowski, J. 2006. Views of family physicians in southwestern Ontario on preventive care services and performance incentives. *Family Practice* 23: 469-471.

Armour, B.S., Pitts, M.M., Maclean, R., Cangialose, C., Kishel, M., Imai, H., Etchason, J. 2001. The effect of explicit financial incentives on physician behavior. *Archives of Internal Medicine* 161: 1261-1266.

Armour, B.S., Pitts, M.M. 2003. Physician financial incentives in managed care: Resource use, quality and cost implications. *Disease Management and Health Outcomes* 11(3): 139-147.

Arrow, K.J. 1986. *Agency and the market*. In: Arrow, K.J., Intriligator, M.D. (Eds.), *Handbook of mathematical economics*, Vol. 3 (pp. 1183-1195). Amsterdam: Elsevier.

Ash, A.S., Ellis, R.P. 2012. Risk-adjusted Payment and Performance Assessment for Primary Care. *Medical Care* 50(8): 643-653.

Austin, P.C., Alter, D., Anderson, G., Tu, J.V. 2004. Impact of the choice of benchmark on the conclusions of hospital report cards. *American Heart Journal* 148(6): 1041-1046.

Australian National Audit Office. 2010. *Practice incentives program*. Audit Report No. 5 2010-11. Canberra: Commonwealth of Australia.

Bailey, C.J., Kodack, M. 2011. Patient adherence to medication requirements for therapy of type 2 diabetes. *International Journal of Clinical Practice* 65(3): 314-322.

Bailit Health Purchasing. 2008. *The feasibility and cost effectiveness of making pay-for-performance opportunities available to Texas Medicaid providers*. http://www.hhsc.state.tx.us/reports/Pay-for-Performance_0209.pdf.

Baker, G., Delbanco, S. 2007. *Pay for performance: National perspective. 2006 longitudinal survey results with 2007 market updates*. San Francisco: Med-Vantage.

Balicer, R. D., Shadmi, E., Lieberman, N., Greenberg-Dotan, S., Goldfracht, M., Jana, L., Jacobson, O. 2011. Reducing health disparities: Strategy planning and implementation in Israel's largest health care organization. *Health Services Research* 46(4): 1281-1299.

Benavent, J., Juan, C., Clos, J., Sequeira, E., Gimferrer, N., Vilaseca, J. 2009. Using pay-for-performance to introduce changes in primary healthcare centers in Spain: First year results. *Quality in Primary Care* 17(2): 123-131.

Benning, A., Ghaleb, M., Suokas, A., Dixon-Woods, M., Dawson, J., Barber, N., Franklin, B., Girling, A., Hemming, K., Carmalt, M., Rudge, G., Naicker, T., Nwulu, U., Choudhury, S., Lilford, R. 2011. Large scale organizational intervention to improve patient safety in four UK hospitals: a mixed method evaluation. *British Medical Journal* 342: d195.

Berlowitz, D.R., Anderson, J.J., Ash, A.S., Brandeis, G.H., Brand, H.K., Moskowitz, M.A. 1998. Reducing random variation in reported rates of pressure ulcer development. *Medical Care* 36(6): 818-825.

Berwick, D. 1995. The toxicity of pay for performance. *Quality Management in Healthcare* 4: 27-33.

Blomqvist, A. 1991. The doctor as double agent: Information asymmetry, health insurance, and medical care. *Journal of Health Economics* 10: 411-432

Blomqvist, A., Léger, P.T. 2005. Information asymmetry, insurance, and the decision to hospitalize. *Journal of Health Economics* 24: 775:793

Blustein, J., Weissman, J.S., Ryan, A.M., Doran, T., Hasnain-Wynia, R. 2011. Analysis raises questions on whether pay-for-performance in Medicaid can efficiently reduce racial and ethnic disparities. *Health Affairs* 30(6): 1165-1175.

Braspenning, J., Kirschner, K., Batenburg, J., van de Rijt, D., Grol, R. 2008. Loon naar werken loont: bonus stimuleert kwaliteitsverbetering huisartsenzorg. *Medisch Contact* 24: 1042-1045.

Brennan, T.A., Spettell, C.M., Fernandes, J., Downey, R.L., Carrara, L.M. 2008. Do managed care plans' tiered networks lead to inequities in care for minority patients? *Health Affairs* 27(4): 1160-1166.

Briesacher, B., Field, T., Baril, J., Gurwitz, J. 2009. Pay-for-performance in nursing homes. *Health Care Financing Review* 30(3): 1-13.

Buetow, S. 2008. Pay-for-performance in New Zealand primary health care. *Journal of Health Organization and Management* 22 (1): 36-47.

Caldis, T. 2007. Composite health plan quality scales. *Health Care Financing Review* 28(3): 95-107.

Campbell, S.M., Kontopantelis, E., Reeves, D., Valderas, J.M., Gaehl, E., Small, N., Roland, M. 2010. Changes in patient experiences of primary care during health service reforms in England between 2003 and 2007. *Annals of Family Medicine* 8(6): 499-506.

Campbell, S.M., McDonald, R., Lester, H. 2008. Experience of pay for performance in English family practice: A qualitative study. *Annals of Family Medicine* 6: 228-234.

Campbell, S.M., Reeves, D., Kontopantelis, E., Sibbald, B., Roland, M. 2009. Effects of pay for performance on the quality of primary care in England. *New England Journal of Medicine* 361(4): 368-378.

Canadian Medical Association. 2010. H*ealth care transformation in Canada. Change that works, care that lasts*. Ottawa, Ontario: Canadian Medical Association.

Casalino, L.P. 2003. Markets and medicine–barriers to creating a "business case for quality". *Perspectives in Biology and Medicine* 46: 38–51.

Casalino, L.P., Elster, A., Eisenberg, A., Lewis, E., Montgomery, J., Ramos, D. 2007. Will pay-for-performance and quality reporting affect health care disparities? *Health Affairs* 26(3): w405-14.

Chaix-Couturier, C., Durand-Zaleski, I., Jolly, D., Durieux, P. 2000. Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. *International Journal for Quality in Health Care* 12(2): 133-142.

Chang, H.J. 2004. Quality-based payment. Taiwan's experience. Presentation at Academy Health Annual Research Meeting (June 6-8), San Diego, California, United States.

Chang, R., Lin, S., Aron, D.C. 2012. A pay-for-performance program in Taiwan improved care for some diabetes patients, but doctors may have excluded sicker ones. *Health Affairs* 31: 93-102.

Chen, T.T., Chung, K.P., Lin, I.C., Lai, M. 2011. The unintended consequence of a diabetes mellitus pay-for-performance program in Taiwan: Are patients with more comorbidities or more severe conditions likely to be excluded from the P4P-program? *Health Services Research* 46: 47-60.

Cheng, S., Lee, T., Chen, C. 2012. A longitudinal examination of a pay-for-performance program for diabetes care: evidence from a natural experiment. *Medical Care* 50(2): 109-116.

Cheng, T.M. 2006. P4P in Taiwan. Presentation at Academy Health Annual Research Meeting (June 25-27), Seattle, Washington, United States.

Chien, A.T., Chin, M.H., Davis, A.M., Casalino, L.P. 2007. Pay for performance, public reporting, and racial disparities in health care: How are programs being designed? *Medical Care Research and Review* 64(5): S283-S304.

Chien, A.T., Li, Z., Rosenthal, M.B. 2010. Improving timely childhood immunizations through pay for performance in Medicaid-managed care. *Health Services Research* 45(6): 1934-1947.

Christiansen, C.L., Morris, C.N. 1997. Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* 127(8): 764-768.

Christianson, J.B., Conrad, D. 2011. *Provider Payment and Incentives*. In: Glied, S., Smith, P. (Eds.), *The Oxford Handbook of Health Economics* (Chapter 26; pp. 624-648). New York: Oxford University Press Inc.

Christianson, J., Leatherman, S., Sutherland, K. 2007. *Financial incentives, healthcare providers and quality improvements: A review of the evidence*. London: The Health Foundation.

Christianson, J., Leatherman, S., Sutherland, K. 2008. Lessons from evaluations of purchaser pay-for-performance programs: A review of the evidence. *Medical Care Research and Review* 65(6): S5-35.

Chung, S., Palaniappan, L., Wong, E., Rubin, H., Luft, H. 2009. Does the frequency of pay-for-performance payment matter? Experience from a randomized trial. *Health Services Research* 45(2): 553-564.

Clinical Practice Improvement Centre. 2008. *Clinical practice improvement payment (CPIP): Implementation plan*. Brisbane: Queensland Health.

Clinical Practice Improvement Centre. 2010. *Clinical practice improvement payment: User guide V2.0, pilot scheme--phase two*. Brisbane: Queensland Health.

Committee Modernisering curatieve zorg (Committee Biesheuvel). 1994. *Gedeelde zorg: betere zorg*. Rijswijk: Ministerie van WVC.

Committee Toekomstige financieringsstructuur huisartsenzorg (Committee Tabaksblat). 2001. *Een gezonde spil in de zorg*. Den Haag: Ministerie van VWS.

Conrad, D.A., Christianson, J.B. 2004. Penetrating the "black box": Financial incentives for enhancing the quality of physician services. *Medical Care Research and Review* 61(3): 37-68.

Conrad, D.A., Perry, L. 2009. Quality-based financial incentives in health care: Can we improve quality by paying for it? *Annual Review of Public Health* 30: 357-371.

Cowen, M.E., Strawderman, R.L. 2002. Quantifying the Physician Contribution to Managed Care Pharmacy Expenses. A Random Effects Approach. *Medical Care* 40(8): 650-661.

Craig, D., Rice, S. 2007. *NHS Economic Evaluation Database Handbook*. York: Centre for Reviews and Dissemination, University of York.

Curtin, K., Beckman, H., Pankow, G., Milillo, Y., Greene, R.A. 2006. Return on investment in pay for performance: a diabetes case study. *Journal of Healthcare Management* 51: 365-376.

Damberg, C., Raube, K., Williams, T., Shortell, S. 2005. Pay for performance: Implementing a statewide project in California. *Quality Management in Health Care* 14(2): 66-79.

Damberg, C.L., Sorbero, M.E., Mehrotra, A., Teleki, S.S., Lovejoy, S., Bradley, L. 2007. *An environmental scan of pay for performance in the hospital setting: Final report*. Rand Health working paper WR-474-ASPE/CMS. Santa Monica: Rand Health.

De Bakker, D.H., Struijs, J.N., Baan, C.B., Raams J., de Wildt, J.E., Vrijhoef, H.J., Schut, F.T. 2012. Early results from adoption of bundled payment for diabetes care in the Netherlands show improvement in care coordination. *Health Affairs* 31(2): 426-33.

De Brantes, F., Rastogi, A., Painter, M. 2010. Reducing potentially avoidable complications in patients with chronic diseases: The Prometheus payment approach. *Health Services Research*, 45(6): 1854-1871.

Deb, P., Manning, W.G., Norton, E.C. 2011. *Modeling health care costs and counts.* Workshop. IHEA World Congress on Health Economics, Toronto, July.

Deci, E.L., Koestner, R., Ryan, R.M. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125: 627-68.

DeLong, E.R., Peterson, E.D., DeLong, D.M., Muhlbaier, L.H., Hackett, S., Mark, D.B. 1997. Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* 16(23): 2645-2664.

District Health Boards New Zealand. PHO performance programme. http://www.dhbnz.org.nz/Site/SIG/pho/Default.aspx

Donabedian, A. 1988. The quality of care: How can it be assessed? *Journal of the American Medical Association* 260(12): 1743-8.

Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U., Roland, M. 2006. Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine* 355: 375-384.

Doran, T., Fullwood, C., Reeves, D., Gravelle, H., Roland, M. 2008a. Exclusion of patients from pay-for-performance targets by English physicians. *The New England Journal of Medicine* 359(3): 274-284

Doran, T., Fullwood, C., Kontopantelis, E., Reeves, D. 2008b. 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* 372(9640): 728-736.

Doran, T., Kontopantelis, E., Valderas, J.M., Campbell, S.M., Roland, M., Salisbury, C., Reeves, D. 2011. Effect of financial incentives on incentivized and non-incentivized clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *British Medical Journal* 342: d3590.

Doran, T., Roland, M. 2010. Lessons from major initiatives to improve primary care in the United Kingdom. *Health Affairs* 29(5): 1023-1029.

Douven, R., Mocking, R., Mosca, I. 2012. *The effect of physician fees and density differences on regional variation in hospital treatments*. CPB Discussion paper 208. The Hague: CPB Netherlands Bureau for Economic Policy Analysis.

Dreier, M., Borutta, B., Stahmeyer, J., Krauth, C., Walter, U. 2010. *Vergleich von Bewertungsinstrumenten für die Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum*. First edition. Köln: DIMDI.

Dranove, D. 1988. Demand inducement and the physician-patient relationship. *Economic Inquiry* 26: 281-298.

Dranove, D., Kessler, D., McClellan, M., Satterthwaite, M. 2003. Is more information better? The effects of "report cards" on health care providers. *Journal of Political Economy* 11: 555-588.

Dranove, D., White, W.D. 1987. Agency and the organization of health care delivery. *Inquiry* 24(4): 405-415.

Drummond, M.F., Jefferson, T.O. 1996.Guidelines for authors and peer reviewers of economic submissions to the BMJ. *British Medical Journal* 313: 275-283.

Drummond, M.F., Sculpher, M.J., Torrance, G., O'Brien, B.J., Stoddart, G.L. 2005. *Methods for the economic evaluation of health care programs*. Oxford: Oxford University Press.

Duckett, S., Daniels, S., Kamp, M., Stockwell, A., Walker, G., Ward, M. 2008. Pay for performance in Australia: Queensland's new clinical practice improvement payment. *Journal of Health Services Research & Policy* 13(3): 174-177.

Dudley, R.A., Frølich, A., Robinowitz, D.L., Talavera, J.A., Broadhead, P., Luft, H., McDonald, K. 2004. *Strategies to support quality-based purchasing: A review of the evidence. Technical review 10*. Rockville: Agency for Healthcare Research and Quality.

Dudley, R.A., Miller, R.H., Korenbrot, T.Y., Luft, H. 1998. The impact of financial incentives on quality of health care. *The Milbank Quarterly* 76(4): 649-686.

Eggleston, K. 2005. Multitasking and mixed systems for provider payment. *Journal of Health Economics* 24(1): 211-23.

EIB. 2012. Integrale bekostiging van zorg: Werk in uitvoering. Den Haag: Evaluatiecommissie Integrale Bekostiging.

Eijkenaar, F., Edgar, P. 2012. Inkoop huisartsenzorg nog een gok. *Medisch Contact* 67(7): 392-396.

Eijkenaar, F., van de Ven, W.P.M.M., Schut, F.T. 2012. *Uitkomstbekostiging in de zorg. Internationale voorbeelden en relevantie voor Nederland*. Rotterdam: iBMG, Erasmus University Rotterdam.

Eisenhardt, K.M. 1989. Agency theory: an assessment and review. *The Academy of Management Review* 14(1): 57-74.

Ellis, R., McGuire, T.G. 1990. Optimal payment systems for health services. *Journal of Health Economics* 9: 375-396.

Eldridge, C., Palmer, N. 2009. Performance-based payment: some reflections on the discourse, evidence and unanswered questions. *Health Policy and Planning* 24: 160-166.

Emmert, M. 2008. Pay for Performance (P4P) im Gesundheitswesen. Ein Ansatz zur Verbesserung der Gesundheitsversorgung? Dissertation. Norderstedt: Books on Demand.

Emmert, M., Eijkenaar, F., Kempter, H., Esslinger, S., Schöffski, O. 2012. Economic evaluation of pay for performance in health care: a systematic review. *European Journal of Health economics* 13(6): 755-767.

Endsley, S., Kirkegaard, M., Baker, G., Murcko, A.C. 2004. Getting rewards for your results: pay-for performance programs. *Family Practice Management* 11: 45-50.

Enthoven, A.C., Tollen, L.A. 2005. Competition in health care: It takes systems to pursue quality and efficiency. *Health Affairs* 5: W420-433.

Epstein, A.M. 2006. Paying for Performance in the United States and Abroad. *New England Journal of Medicine* 355: 406-408.

Evans, R.G. 1974. *Supplier-induced demand: some empirical evidence and implications*. In: Perlman, M. (Ed.), *The Economics of Health and Medical Care* (pp. 163-173). New York: John Wiley and Sons.

Fachklinik Herzogenaurach. 2010. *Qualitätsbericht rehabilitation 2009*. Hopfen am See: m&i-Klinikgruppe Enzensberg.

Fairbrother, G., Hanson, K.L., Friedman, S., Butts, G.C. 1999. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization rates. *American Journal of Public Health* 89(2): 171-175.

Fiorentini, G., Iezzi, E., Lippi Bruni, M., Ugolini, C. 2011. Incentives in primary care and their impact on potentially avoidable admissions. *European Journal of Health Economics* 12: 297-309.

Flodgren, G., Eccles, M.P., Shepperd, S., Scott, A., Parmelli, E., Beyer, F.R. 2011. An Overview of reviews evaluation the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. *Cochrane Database on Systematic Reviews* 7: CD009255.

Frederick, S., Loewenstein, G., O'Donoghue, T. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40(2): 351-401.

Freedman, J.L., Cunningham, J.A., Krismer, K. 1992. Inferred values and the reverse-incentive effect in induced compliance. *Journal of Personality and Social Psychology* 62(3): 357-368.

Freidson, E. 2001. *Professionalism: The third logic*. London: Polity Press.

Frey, B.S. 1997. On the relationship between intrinsic and extrinsic work motivation. *International Journal of Industrial Organization* 15(4): 427-439.

Friedberg, M.W., Damberg, C.L. 2011. *Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting. A Guide for Community Quality Collaboratives.* AHRQ Publication No. 11-0093. Rockville: Agency for Healthcare Research and Quality.

Friedberg, M.W., Damberg, C.L. 2012. A five-point checklist to help performance reports incentivize improvement and effectively guide patients. *Health Affairs* 31(3): 612-618.

Friedberg, M.W., Safran, D.G., Coltin, K., Dresser, M., Schneider, E.C. 2010. Paying for performance in primary care: Potential impact on practices and disparities. *Health Affairs* 29(5): 926-932.

Friedman, N. L., Kokia, E., Shemer, J. 2003. Health value added: Linking strategy, performance, and measurement in healthcare organizations. *Israel Medical Association Journal* 5: 3-8.

Frohlich, N., Katz, A., De Coster, C., Dik, N., Soodeen, R.-A., Watson, D., Bogdanovic, B. 2006. Profiling primary care physician practice in Manitoba. Winnipeg: Manitoba Centre for Health Policy.

Frølich, A., Talavera, J.A., Broadhead, P., Dudley, R.A. 2007. A behavioral model of clinician responses to incentives to improve quality. *Health Policy* 80(1): 179-193.

Fuchs, V.R. 2004. More variation in use of care, more flat-of-the-curve medicine. *Health Affairs*, *Variations Revisited: Web-Exclusive Collection* 104-107.

Fung, C.H., Lim, Y., Mattke, S., Damberg, C., Shekelle, P.G. 2008. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Annals of Internal Medicine* 148(2): 111-123.

Fung, V., Schmittdiel, J.A., Fireman, B., Meer, A., Thomas, S., Smider, N., Hsu, J., Selby, J. 2010. Meaningful variation in performance. A systematic literature review. *Medical Care* 48(2): 140-148.

Gaynor, M., Gertler, P. 1995. Moral hazard and risk spreading in partnerships. *Rand Journal of Economics* 26(4): 591-613.

Gaynor, M., Rebitzer, J.B., Taylor, L.J. 2004. Physician incentives in health maintenance organizations. *Journal of Political Economy* 112(4): 915-31.

Gené-Badia, J., Escaramis-Babiano, G., Sans-Corrales, M., Sampietro-Colom, L., Aguado-Menguy, F., Cabezas-Pena, C., Gallo de Puelles, P. 2007. Impact of economic incentives on quality of professional life and on end-user satisfaction in primary care. *Health Policy* 80: 2-10.

Gerdes, N., Funke, U., Schuwer, U., Themann, P., Kunze, H., Walle, E., von Ameln, M. 2008. Ergebnis-orientierte vergütung der rehabilitation nach schlaganfall. http://forschung.deutsche-rentenversicherung.de/ForschPortalWeb/ressource?key=05_Gerdes.pdf

Gerdes, N., Funke, U., Schuwer, U., Kunze, H., Walle, E., Kleinfeld, A., Jäckel, W.H. 2009. Ergebnisorientierte vergutung der rehabilitation nach schlaganfall. Entwicklungsschritte eines modellprojekts 2001-2008. *Die Rehabilitation* 48: 190-201.

Gibbons, R. 1998. Incentives in organizations. *Journal of Economic Perspectives* 12: 115-132.

Ginsburg, P.B., Grossman, J.M. 2005. When the price isn't right: how inadvertent payment incentives drive medical care. *Health Affairs* 24(5): W5376-5384.

Giuffrida, A., Gosden, T., Forland, F., Kristiansen, I.S., Sergison, M., Leese, B., Pedersen, L., Sutton, M. 2000. Target payments in primary care: effects on professional practice and health care outcomes. *Cochrane Database on Systematic Reviews* 4: CD000531.

Glance, L.G., Dick, A., Osler, T.M., Li, Y, Mukamel, D.B. 2006. Impact of changing the Statistical Methodology on Hospital and Surgeon Ranking. The Case of the New York State Cardiac Surgery Report Card. *Medical Care* 44(4): 311-319.

Glickman, S.W., Ou, F., DeLong, E.R., Roe, M., Lytle, B., Mulgund, J., Rumsfeld, J., Gibler, W., Ohman, E.M., Schulman, K., Peterson, E.D. 2007. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *Journal of the American Medical Association* 297(21): 2373-2380

Gneezy, U., Rustichini, A. 2000. Pay enough or don't pay at all. *Quarterly Journal of Economics* 115(3): 791-810.

Goldstein, H., Spiegelhalter, D.J. 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*. Series A 190(3): 385-443.

Gonzalez-Perez, J.G. 2002. Developing a scoring system to quality assess economic evaluations. *European Journal of Health Economics* 3: 131-136.

Gosden, T., Forland, F., Kristiansen, I.S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., Pedersen, L. 2000. Capitation, salary, fee-for-service and mixed systems of payment: Effects on the behaviour of primary care physicians. *Cochrane Database on Systematic Reviews* 3(3): CD002215.

Gosden, T., Forland, F., Kristiansen, I.S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., Pedersen, L. 2001. Impact of payment method on behaviour of primary care physicians: A systematic review. *Journal of Health Services Research & Policy* 6(1): 44-55.

Graf von der Schulenburg, J., Greiner, W., Jost, F., Klusen, N., Kubin, M., Leidl, R., Mittendorf, T., Rebscher, H., Schöffski, O., Vauth, C., Volmer, T., Wahler, S., Wasem, J., Weber, C. 2007. Deutsche empfehlungen zur gesundheitsökonomischen evaluation-dritte und aktualisierte fassung des Hannoveraner Konsens. *Gesundheitsökonomie Qualitätsmanagement* 12: 285-290.

Gravelle, H., Sutton, M., Ma, A. 2008. *Doctor behaviour under a pay for performance contract: Further evidence from the quality and outcomes framework*. CHE Research Paper 34. York: Centre for Health Economics, University of York.

Gravelle, H., Sutton, M., Ma, A. 2010. Doctor behaviour under a pay for performance contract: Treating, cheating and case finding? *The economic journal* 120(542): F129-156.

Greene, R.A., Beckman, H.B., Mahoney, T. 2008. Beyond the efficiency index: finding a better way to reduce overuse and increase efficiency in physician care. *Health Affairs* 27: w250-259.

Greene, S.E., Nash, D.B. 2009. Pay for performance: an overview of the literature. *American Journal of Medical Quality* 24: 140-163.

Greenfield, S., Kaplan, S.H., Kahn, R., Ninomiya, J., Griffith, J.L. 2002. Profiling care provided by different groups of physicians: Effects of patient case-mix (bias) and physician-level clustering on quality assessment results. *Annals of Internal Medicine* 136(2): 111-121.

Grol, R. 2001. Successes and failures in the implementation of evidence-based guidelines for clinical practice. *Medical Care* 39(8): II46-54.

Gross, R., Elhaynay, A., Friedman, N., Buetow, S. 2008. Pay-for-performance programs in Israeli sick funds. *Journal of Health Organization and Management* 22(1): 23-35.

Grossman, S.J., Hart, O.D. 1983. An analysis of the principal-agent problem. *Econometrica* 51: 7-45.

Gruisen, W.J.A.M., Muijrers, P.E.M. 2002. Doelmatig voorschrijven: verzekeraar stimuleert huisartsen met geld. *Medisch Contact* 57: 714-717.

Grumbach, K., Osmond, D., Vranizan, K., Jaffe, D., Bindman, A.B. 1998. Primary care physicians' experience of financial incentives in managed-care systems. *The New England Journal of Medicine* 339(21): 1516-1521.

Guthrie, B., McLean, G., Sutton, M. 2006. Workload and reward in the quality and outcomes framework of the 2004 GP contract. *British Journal of General Practice* 56: 836-41.

Habets, P., Bruggeman, F., Lock, B. 2009. Meer wortel, minder stok. Beloon huisarts voor voorschrijven eerstekeusmiddel. *Medisch Contact* 64(14): 585-587.

Hahn, J. 2006. Pay-for-performance in health care. http://www.vascularweb.org/professionals/Government_Relations/PDF_Doc/CRS%20report%20on%20P4P.pdf.

Hanchak, N.A., Schlackman, N. 1995. The measurement of physician performance. *Quality Management in Health Care* 4(1): 1-12.

Hanchak, N.A., Schlackman, N., Harmon-Weiss, S. 1996. US Healthcare's quality-based compensation model. *Health Care Financing Review* 17: 143-159.

Health and Human Services. Hospital Inpatient Value-Based Purchasing Program. Final Rule. 2011. *Federal Register* 88(88): 26490-26547.

Health and Social Care Information Centre. 2009. *General and Personal Medical Services in England: 1998-2008*. London: Health and Social Care Information Centre.

Heath, C., Larrick, R.P., Wu, G. 1999. Goals as reference points. *Cognitive Psychology* 38: 79-109.

Hellinger, F.J. 1996. The impact of financial incentives on physician behavior in managed care plans: a review of the evidence. *Medical Care Research and Review* 53: 294-314.

Het Financieele Dagblad. 2011. Nieuw ziekenhuiscontract levert besparing op van ruim €1 mrd. *Het Financieele Dagblad* April 4.

Higgins, A., Zeddies, T., Pearson, S.D. 2011. Measuring the performance of individual physicians by collecting data from multiple health plans: the results of a two-state test. *Health Affairs* 30(4): 673-681.

Higgins, J.P.T., Green, S. (Eds.). 2008. *Cochrane handbook for systematic reviews of interventions*. Chichester: Wiley.

Hillman, A.L., Pauly, M.V., Kerstein, J.J. 1989. How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations? *The New England Journal of Medicine* 321(2): 86-92.

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

Hillman, A.L., Ripley, K., Goldfarb, Weiner, J., N., Nuamah, I., Lusk, E. 1999. The use of physician financial incentives and feedback to improve pediatric preventive care in medicaid managed care. *Pediatrics* 104 (4): 931-935.

Hochrhein-Institut. *Ergebnisorientierte vergütung der rehabilitation nach schlaganfall (ERGOV)*. Bad Säckingen: Hochrhein-Institut. www.hri.de/index.php?menuid=1&reporeid=59.

Hofer, T.P., Hayward, R.A., Greenfield, S., Wagner, E.H., Kaplan, S.H., Manning, W.G. 1999. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *Journal of the American Medical Association* 281(22): 2098-2105.

Holmstrom, B., Milgrom, P. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7(1): 24-52.

Holmboe, E.S., Weng, W., Arnold, G.K., Kaplan, S.H., Normand, S.L., Greenfield, S., Hood, S., Lipner, M. 2010. The comprehensive care project: Measuring physician performance in ambulatory practice. *Health Services Research* 45(6): 1912-1933.

Huang, I.C., Diette, G.B., Dominici, F., Frangakis, C., Wu, A.W. 2005a. Variations of physician group profiling indicators for asthma care. *The American Journal of Managed Care* 11(1): 38-44.

Huang, I.C., Dominici, F., Frangakis, C., Diette, G.B., Damberg, C.L., Wu, A.W. 2005b. Is risk-adjustor selection more important than statistical approach for provider profiling? Asthma as an example. *Medical Decision Making* 25(1): 20-34.

Iezzoni, L.I. 2003. *Risk-adjustment for measuring health care outcomes*. Third edition. Chicago: Health Administration Press.

Iezzoni, L.I., Ash, A.S., Shwartz, M., Daley, J., Hughes, J.S., Mackiernan, Y.D. 1996. Judging hospitals by severity-adjusted mortality rates: The influence of the severity-adjustment method. *American Journal of Public Health* 86(10): 1379-1387.

Institute for Quality and Efficiency in Health Care / Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. 2008. *Allgemeine Methoden, Version 3*. Köln: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen.

Institute of Medicine. 1999. *To err is human: Building a safer health system*. Washington, DC: National University Press.

Institute of Medicine. 2001. *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National University Press.

Institute of Medicine. 2007. *Rewarding provider performance: aligning incentives in Medicare*. Washington, DC: The National Academies Press.

Ittner, C.D., Larcker, D.F. 2002. Determinants of performance measure choices in worker incentive plans. *Journal of Labor Economics* 20(2): S58-S90

IVM (Institute for rational use of medicines). 2012. *Monitor Voorschrijfgedrag Huisartsen 2012*. Utrecht: IVM.

Jha, A.K., Joynt, K.E., Orav, E.J., Epstein, A.M. 2012. The long-term effect of Premier pay for performance on patient outcomes. *The New England Journal of Medicine* 366(17): 1606-1615.

Jha, A.K. 2013. Time to get serious about pay for performance. *Journal of the American Medical Association* 309(4): 347-348.

Jones, A.M. 2000. *Health econometrics*. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, vol. 1A (Chapter 6, pp. 265-344). Amsterdam: North-Holland.

Jones, H.E., Spiegelhalter, D.J. 2011. The identification of "unusual" health care providers from a hierarchical model. *The American Statistician* 65(2): 154-163.

Kahn, J.M., Scales, D.C., Au, D.H., Carson, S.S., Curtis, J.R., Dudley, R.A., Iwashyna, T.J., Krishnan, J.A., Maurer, J.R., Mularski, R., Popovich Jr, J., Rubenfeld, G.D., Heffner, J.E. 2010. An official American thoracic society policy statement: pay-for-performance in pulmonary, critical care, and sleep medicine. *American Journal of Respiratory and Critical Care* 181: 752-761.

Kahneman, D., Knetsch, J.L., Thaler R,. 1986. Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review* 76(4): 728-741.

Kahneman, D., Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47(2): 263-291.

Kane, R.L., Johnson, P.E., Town, R.J., Butler, M. 2004. *Economic incentives for preventive care. Evidence Report/Technology Assessment* (101). Publication 04-E024-2. Rockville: Agency for Healthcare Research and Quality.

Kang, H.C., Hong, J.S. 2011. Do differences in profiling criteria bias performance measurements? Economic profiling of medical clinics under the Korea national health insurance program: An observational study using claims data. *BMC Health Services Research* 11: 189.

Kantarevic, J., Kralj, B., Weinkauf, D. 2010. *Enhanced fee-for-service model and access to physician services: Evidence from family health groups in Ontario*. IZA Discussion Paper No. 4862. Bonn: Institute for the Study of Labor.

Kaplan, S.H., Griffith, J.L., Price, L.L., Pawlson, L.G., Greenfield, S. 2009. Improving the reliability of physician performance assessment: Identifying the "physician effect" on quality and creating composite measures. *Medical Care* 47: 378-387.

Karve, A.M., Ou, F.S., Lytle, B.L., Peterson, E.D. 2008. Potential unintended financial consequences of pay-for-performance on the quality of care for minority patients. *American Heart Journal* 155(3): 571-6.

Katon, W., Rutter, C.M., Lin, E., Simon, G., Von Korff, M., Bush, T., Walker, E., Ludman, E. 2000. Are there detectable differences in quality of care or outcome of depression across primary care providers? *Medical Care* 38(6): 552-561.

Katz, A., Bogdanovic, B., Soodeen, R. 2010. *Physician integrated network baseline evaluation: Linking electronic medical records and administrative data*. Winnipeg: Manitoba Centre for Health Policy.

Kilpatrick, K.E., Lohr, K.N., Leatherman, S., Pink, G., Buckel, J.M., Legarde, C., Whitener, L. 2005. The insufficiency of evidence to establish the business case for quality. *International Journal of Quality in Health Care* 17: 347-355.

Kirschner, K., Braspenning, J., Batenburg, J., van de Rijt, D., Muijrers, P., van Everdingen, C., Grol, R. 2008. Value for money: Een model voor honoreren van kwaliteit in de huisartsenpraktijk. Nijmegen: Scientific Institute for Quality of Healthcare.

Kirschner, K., Braspenning, J., Gootzen, T., van Everdingen, C., Batenburg, J., Verstappen, W., Grol, R. 2009. Pay-for-performance in de huisartsenpraktijk. Een experiment in Zuid-Nederland. Nijmegen: Scientific Institute for Quality of Healthcare.

Kizer, K.W. 2001. Establishing health care performance standards in an era of consumerism. *Journal of the American Medical Association* 286: 1213-1217.

Koolman, X., Luijendijk, D., Boonen, L. 2011. *Op weg naar meer betrouwbare prestatieberekening in verpleeghuizen, verzorgingshuizen en thuiszorgorganisaties*. Rotterdam: SiRM.

Kouides, R.W., Bennett, N., Lewis, B., Cappuccio, J., Barker, W.H., LaForce, F.M. 1998. Performance-based physician reimbursement and influenza immunization rates in the elderly. The primary-care physicians of Monroe County. *American Journal of Preventive Medicine* 14: 89-95.

Krein, S.L., Hofer, T.P., Kerr, E.A., Hayward, R.A. 2002. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. *Health Services Research* 37(5): 1159-1180.

Krieger, N. 2003. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures– the public health disparities geocoding project. *American Journal of Public Health* 93: 1655-1671.

Kuo, R.N.C., Chung, K.-P., Lai, M.-S. 2011. Effect of the pay-for-performance program for breast cancer care in Taiwan. *American Journal of Managed Care* 17(5): e203-211.

Landis, J.R., Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174.

Landon, B.E., Normand, S.L., Blumenthal, D., Daley, J. 2003. Physician clinical performance assessment: prospects and barriers. *Journal of the American Medical Association* 290(9): 1183-1189.

Landon, B.E., Normand, S.L. 2008. Performance measurement in the small office practice: challenges and potential solutions. *Annals of Internal Medicine* 148: 353-357.

Landon, B.E., Schneider, E.C., Tobias, C., Epstein, A.M. 2004. The evolution of quality management in Medicaid managed care. *Health Affairs* 23: 245-254.

Langelaan, M., Baines, R.J., Broekens, M.A., Siemerink, K.M., van de Steeg, L., Asscheman, H., de Bruijne, M.C., Wagner, C. *Monitor Zorggerelateerde Schade 2008. Dossieronderzoek in Nederlandse ziekenhuizen*. Utrecht, Amsterdam: NIVEL, EMGO.

Leatherman, S., Berwick, D., Iles, D., Lewin, L.S., Davidoff, F., Nolan, T., Bisognano, M. 2003. The business case for quality: case studies and an analysis. *Health Affairs* 22: 17-30.

Lee, J.T., Netuveli, G., Majeed, A., Millett, C. 2011. The effects of pay for performance on disparities in stroke, hypertension, and coronary heart disease management: interrupted time series study. *PLoS ONE* 6(12): e27236.

Lee, T.T., Cheng, S.H., Chen, C.C., Lai, M.S. 2010. A pay-for-performance program for diabetes care in Taiwan: A preliminary assessment. *The American Journal of Managed Care* 16(1): 65-69.

Li, J., Hurley, J., DeCicca, P., Buckley, G. 2010b. *Physician response to pay-for-performance: Evidence from a natural experiment*. Hamilton: McMaster University.

Li, Y.-H., Tsai, W.-C., Khan, M., Yang, W., Lee, T.-F., Wu, Y.-C., Kung, P.-T. 2010a. The effects of pay-for-performance on tuberculosis treatment in Taiwan. *Health Policy and Planning* 25: 334-341.

Lindenauer, P.K., Remus, D., Roman, S., Rothberg, M.B., Benjamin, E.M., Ma, A., Bratzler, D.W. 2007. Public reporting and pay for performance in hospital quality improvement. *The New England Journal of Medicine* 356: 486-496.

Long, J.A., Helweg-Larsen, M., Volpp, K.G. 2008. Patient opinions regarding 'pay for performance for patients'. *Journal of General Internal Medicine* 23: 1647-1652.

Ludwig Boltzmann Institut für Health Technology Assessment. 2009. *Schweregraddifferenzierung in der neurologischen und trauma-rehabilitation: Internationale erfahrungen zur qualitats-, performancemessung und vergütung*. Vienna: Ludwig Boltzmann Institut für Health Technology Assessment.

Lyratzopoulos, G., Elliott, M.N., Barbiere, J.M., Staetsky, L., Paddison, C.A., Campbell, J., Roland, M. 2011. How can health care organizations be reliably compared? Lessons from a national survey of patient experience. *Medical Care* 49(8): 724-733.

Ma, C.A. 1994. Health care payment systems: cost and quality incentives. *Journal of Economics & Management Strategy* 3(1): 93-112.

Ma, C.A., McGuire, T.G. 1997. Optimal health insurance and provider payment. *American Economic Review* 87: 685-704.

MacDonald, G.M. 1984. New directions in the economic theory of agency. *Canadian Journal of Economics* 17(3): 415-440.

Manitoba Health. 2007. *Physician integrated network: Evaluation plan*. Winnipeg: Manitoba Health.

Manitoba Health. 2010. *PIN information management guide 1.5*. Winnipeg: Manitoba Health.

Manitoba Health. *Physician integrated network*. http://www.gov.mb.ca/health/phc/pin/index.html.

Manning, W.G., Mullahy, J. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics* 20: 461-494.

Marshall, M., Harrison, S. 2005. It's about more than money: financial incentives and internal motivation. *Quality & Safety in Health Care* 14(1): 4-5.

Mattke, S., Seid, M., Ma, S. 2007. Evidence for the effect of disease management: Is $1 billion a year a good investment? *American Journal of Managed Care* 13: 670-676.

McCullagh, P., Nelder, J.A. 1989. *Generalized linear models*. Second edition. London: Chapman & Hall/CRC.

McDermott, S., Williams, T., Lempert, L., Yanagihara, D. 2006. *Advancing quality through collaboration: The California pay for performance program. A report on the first five years and a strategic plan for the next five years*. Oakland: Integrated Healthcare Association.

McDonald, R., Roland, M. 2009. Pay for performance in primary care in England and California: Comparison of unintended consequences. *Annals of Family Medicine* 7(2): 121-127.

McDonald, R., White, J., Marmor, T.R. 2009. Paying for performance in primary medical care: Learning about and learning from "success" and "failure" in England and California. *Journal of Health Politics, Policy and Law* 34(5): 747-776.

McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., Kerr, E.A. 2003. The quality of health care delivered to adults in the United States. *The New England Journal of Medicine* 348(26): 2635-2645.

McGuire, T.G. 2000. *Physician agency*. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, vol. 1A (Chapter 9, pp. 461-536). Amsterdam: North-Holland.

McGuire, T.G. 2011. Physician Agency and Payment for Primary Medical Care. In: Glied, S., Smith, P. (Eds.), *The Oxford Handbook of Health Economics* (Chapter 25, pp. 602-623). New York: Oxford University Press Inc.

McNamara, P. 2005. Quality-based payment: six case examples. *International Journal for Quality in Health Care*, 17(4): 357-362.

Medicare Australia. 2011. *Practice incentives program*. http://www.medicareaustralia.gov.au/provider/incentives/pip/index.jsp

Med-Vantage. 2009. *4th annual national P4P survey results (App. B)*. San Francisco: Med-Vantage.

Mehrotra, A., Damberg, C.L., Sorbero, M.E., Teleki, S.S. 2009. Pay for performance in the hospital setting: What is the state of the evidence? *American Journal of Medical Quality* 24: 19-28.

Mehrotra, A., Sorbero, M.E., Damberg, C.L. 2010a. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *The American Journal of Managed Care* 16(7): 497-503.

Mehrotra, A., Adams, J.L., Thomas, J.W., McGlynn, E.A. 2010b. Cost profiles: should the focus be on individual physicians or physician groups? *Health Affairs* 29(8): 1532-1538.

Metfessel, B.A., Greene, R.A. 2012. A nonparametric statistical method that improves physician cost of care analysis. *Health Services Research* 47(6): 398-417.

Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G. 2011. Review of statistical methods for analyzing healthcare resources and costs. *Health Economics* 20: 897-916.

Moore, S.H., Martin, D.P., Richardson, W.C., Riedel, D.C. 1980. Cost containment through risk-sharing by primary care physicians: A history of the development of united healthcare. *Health Care Financing Review* 1(4): 1-13.

Mukamel, D.B., Brower, C.A. 1998. The influence of risk-adjustment methods on conclusions about quality of care in nursing homes based on outcome measures. *The Gerontologist*, 38(6): 695-703.

Mukamel, D.B., Glance, L.G., Li, Y., Weimer, D.L., Spector, W.D., Zinn, J.S., Mosqueda, L. 2008. Does risk-adjustment of the CMS quality measures for nursing homes matter? *Medical Care* 46(5): 532-541.

Mullen, K.J., Frank, R.G., Rosenthal, M.B. 2010. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *RAND Journal of Economics* 41(1): 64-91.

Nahra, T.A., Reiter, K.L., Hirth, R.A., Shermer, J.E., Wheeler, J.R. 2006. Cost-effectiveness of hospital pay-for-performance incentives. *Medical Care Research and Review* 63: S49-72.

Nyweide, D.J., Weeks, W.B., Gottlieb, D.J., Casalino, L.P., Fisher, E.S. 2009. Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. *Journal of the American Medical Association* 302(22): 2444-50.

Nelson, A.R. 2007. Pay-for-performance programs: ethical questions and unintended consequences. *Journal of Family Practice* 56: 16A-18A.

New Zealand Ministry of Health. 2010. *Primary health care*. http://www.health.govt.nz/our-work/primary-health-care.

Newhouse, J.P. 1996. Reimbursing health plans and health providers: selection versus efficiency in production. *Journal of Economic Literature* 34: 1236-1263.

Newhouse, J.P., 2002. *Pricing the Priceless: A Health Care Conundrum*. Cambridge: MIT Press.

NHS North West. 2008. *A North West health system approach to advancing quality*. Manchester: NHS North West.

NHS North West. *Advancing quality*. http://www.advancingqualitynw.nhs.uk.

Nicholson, S., Pauly, M.V., Wu, A.Y., Murray, J.F., Teutsch, S.M., Berger, M.L. 2008. Getting real performance out of pay-for-performance. *Milbank Quarterly* 86: 435–457.

Norton, E.C. 1992. Incentive regulation of nursing homes. *Journal of Health Economics*. 11: 105-128.

Nunnally, J.C., Bernstein, I.H. 1994. *Psychometric theory*. Third edition. New York: McGraw-Hill.

Nyweide, D.J., Weeks, W.B., Gottlieb, D.J., Casalino, L.P., Fisher, E.S. 2009. Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. *Journal of the American Medical Association* 302 (22): 2444-2450.

NZa, 2012. Monitor medisch-specialistische zorg. Weergave van de markt 2008-2012. Utrecht: Nederlandse Zorgautoriteit.

OECD, 2012. *OECD Health Data 2012 - Frequently Requested Data*. http://www.oecd.org/els/healthpoliciesanddata/oecdhealthdata2012-frequentlyrequesteddata.htm.

O'Kane, M.E. 2007. Performance-based measures: the early results are in. *Journal of Managed Care Pharmacy* 13: 3-6.

Oliver, P. 1980. Rewards and punishments as selective incentives for collective action: theoretical investigations. *American Journal of Sociology* 85(6): 1356-1375.

Parke, D.W. 2007. Impact of a pay-for-performance intervention: financial analysis of a pilot program implementation and implications for ophthalmology. *Transactions of the American Ophthalmological Society* 105: 448-460.

Pauly, M.V. 2000. *Insurance reimbursement*. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, vol. 1A (Chapter 10, pp. 537–560). Amsterdam: North-Holland.

Peabody, J., Shimkhada, R., Quimbo, S., Florentino, J., Bacate, M., McCulloch, C. E., Solon, O. 2011. Financial incentives and measurement improved physicians' quality of care in the Philippines. *Health Affairs* 30: 773-781.

Pedrós, C., Vallano, A., Cereza, G., Mendoza-Aran, G., Agustí, A., Aguilera, C., Arnau, J.M. 2009. An intervention to improve spontaneous adverse drug reaction reporting by hospital physicians: A time series analysis in Spain. *Drug Safety* 32: 77-83.

Peeters, B., van Tongelen, I., Boussery, K., Mehuys, E., Remon, J.P., Willems, S. 2011. Factors associated with medication adherence to oral hypoglycaemic agents in different ethnic groups suffering from type 2 diabetes: a systematic literature review and suggestions for further research. *Diabetes Medicine* 28(3): 262-275.

Peiro, S., Garcia-Altes, A. 2008. Possibilities and limitations of results-based management, pay-for-performance and the redesign of incentives. *Gaceta Sanitaria* 22: 143-155.

Petersen, L.A., Woodard, L.D., Urech, T., Daw, C., Sookanan, S. 2006. Does pay-for-performance improve the quality of health care? *Annals of Internal Medicine* 145(4): 265-272.

Pope, G.C., Kautter, J. 2007. Profiling efficiency and quality of physician organizations in Medicare. *Health Care Financing Review* 29(1): 31-43.

Pregibon, D.1980. Goodness of Link Tests for Generalized Linear Models. *Applied Statistics* 29: 15-24.

Premier, Inc. 2010. *NHS Northwest and premier advancing quality program. Composite quality score and outcome methodologies year one*. Charlotte: Premier, Inc.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37(1): 7-63.

Price Waterhouse Coopers. 2001. *Evaluation of primary care reform pilots in Ontario phase 2 interim report*. Toronto: Ontario Ministry of Health and Long-Term Care.

Prinsze, F., & van Vliet, R.C.J.A. 2007. Health-based risk-adjustment: Improving the pharmacy-based cost group model by adding diagnostic cost groups. *Inquiry* 44: 469-480.

Queensland Health. 2010. *Clinical practice improvement payment*. http://www.health.qld.gov.au/cpic/service_improve/cpip.asp.

Racz, M.J. Sedransk, J. 2010. Bayesian and frequentist methods for profiling using risk-adjusted assessments of medical outcomes. *Journal of the American Statistical Association* 105(489): 48-58.

Reeves, D., Campbell, S.M., Adams, J., Shekelle, P.G., Kontopantelis, E., Roland, M. 2007. Combining multiple indicators of clinical quality: An evaluation of different analytic approaches. *Medical Care* 45(6): 489-496.

Reiter, K.L., Kilpatrick, K.E., Greene, S.B., Lohr, K.N., Leatherman, S. 2007. How to develop a business case for quality. *International Journal of Quality in Health Care* 19: 50-55.

Rice, N., Jones, A. 1997. Multilevel models and health economics. *Health Economics* 6: 561-575.

Rizzo, J.A., Blumenthal, J.A. 1996. Is the target income hypothesis an economic heresy? *Medical Care Research and Review* 53(3): 243-266.

Rizzo, J.A., Zeckhauser, R.J. 2003. Reference incomes, loss aversion, and physician behavior. *Review of Economics and Statistics* 85(4): 909-922.

Rhoads, K.F., Konety, B.M., Dudley, R.A. 2009. Performance measurement, public reporting, and pay-for-performance. *Urology Clinics of North America* 36: 37-48.

Robinowitz, D.L., Dudley, R.A. 2006. Public reporting of provider performance: can its impact be made greater? *Annual Review of Public Health* 27: 517-536.

Robinson, J.C. 2001. Theory and Practice in the Design of Physician Payment Incentives. *Milbank Quarterly* 79(2): 149-177.

Robinson, J.C., Williams, T., Yanagihara, D. 2009. Measurement of and reward for efficiency in California's pay-for-performance program. *Health Affairs* 28(5): 1438-1447.

Rochon, M., Pink, G.H., A.D. Brown, Studer, M.L., Reiter, K.L., Leatt, P., Landon, B.E., Culyer, T., Golden, B.R., Feasby, T.E., Gerdes, C., Halparin, E., Davis, D., Greengarten, M., Hundert, M., Vertesi, L., Hudson, A.R. 2006. *HealthcarePapers* 6(4).

Rodriguez, H.P., Perry, L., Conrad, D.A., Maynard, C., Martin, D.P., Grembowski, D.E. 2012. The reliability of medical group performance measurement in a single insurer's pay for performance program. *Medical Care* 50(2): 117-123.

Roeg, D., van de Goor, I., Garretsen, H. 2005. Towards quality indicators for assertive outreach programmes for severely impaired substance abusers: concept mapping with Dutch experts. *International Journal of Quality in Health Care* 17: 203-208.

Roland, M. 2004. Linking physicians' pay to the quality of care – a major experiment in the United Kingdom. *The New England Journal of Medicine* 351(14): 1448-1454.

Roland, M. 2006. Pay-for-performance: Too much of a good thing? A conversation with Martin Roland. Interview by Robert Galvin. *Health Affairs* 25(5): w412-419.

Roland, M. 2012. Pay-for-performance: not a magic bullet. *Annals of Internal Medicine* 157(12): 912-913.

Roland, M., Elliott, M., Lyratzopoulos, G., Barbiere, J., Parker, R. A., Smith, P., Bower, P., Campbell, J. 2009. Reliability of patient responses in pay for performance schemes: Analysis of national general practitioner patient survey data in England. *British Medical Journal* 339(7727): b3851.

Rosen, A.K, Wu, J., Chang, B.H., Berlowitz, D., Rakovski, C., Ash, A., Moskowitz, M. 2001. Risk-adjustment for measuring health outcomes: An application in VA long-term care. *American Journal of Medical Quality* 16(4): 118-127.

Rosen, A.K., Rakovski, C.C., Loveland, S.A., Anderson, J.J., Berlowitz, D. R. 2002. Profiling resource use: Do different outcomes affect assessments of provider efficiency? *The American Journal of Managed Care* 8(12): 1105-1115.

Rosen, A.K, Loveland, S., Rakovski, C., Christiansen, C., Berlowitz, D. 2003. Do different case-mix measures affect assessments of provider efficiency? Lessons from the department of Veterans Affairs. *Journal of Ambulatory Care Management* 26: 229-242.

Rosenthal, M.B. 2007a. Pay for performance and beyond. *Expert Review of Pharmacoeconomics & Outcomes Research* 7: 351-355.

Rosenthal, M.B. 2007b. Pay for performance: Rumors of its demise may be exaggerated. *American Journal of Managed Care* 13(5):238-239.

Rosenthal, M.B. 2008. Beyond pay for performance – emerging models of provider-payment reform. *The New England Journal of Medicine* 359: 1197-1200.

Rosenthal, M.B., Dudley, R.A. 2007. Pay-for-performance: Will the latest payment trend improve care? *Journal of the American Medical Association* 297(7): 740-744.

Rosenthal, M.B., Frank, R.G. 2006. What is the empirical basis for paying for quality in health care? *Medical Care Research and Review* 63(2): 135-157.

Rosenthal, M.B., Landon, B.E., Normand, S.L., Frank, R.G., Epstein, A.M. 2006. Pay for performance in commercial HMOs. *The New England Journal of Medicine* 355: 1895-1902.

Rosenthal, M.B., Landon, B.E., Normand, S.T. 2007a. Employers' use of value-based purchasing strategies. *Journal of the American Medical Association* 298: 2281-2288.

Rosenthal, M.B., Landon, B.E., Howitt, K., Song, H.R., Epstein, A. 2007b. Climbing up the pay-for-performance learning curve: Where are the early adopters now? *Health Affairs* 26: 1674-1682.

Rosenthal, M.B., Li, Z., Robertson, A.D., Milstein, A. 2009. Impact of financial incentives for prenatal care on birth outcomes and spending. *Health Services Research* 44: 1465-1479.

Roski, J., Jeddeloh, R., An, L., Lando, H., Hannan, P., Hall, C., Zhu, S. 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: 291-299.

Rowe, J.W. 2006. Pay-for-performance and accountability: related themes in improving health care. *Annals of Internal Medicine* 145: 695-699.

Rubinstein, A., Rubinstein, F., Botargues, M., Barani, M., Kopitowski, K. 2009. A multimodal strategy based on pay-per-performance to improve quality of care of family practitioners in Argentina. *Journal of Ambulatory Care Management* 32: 103-114.

RVZ. 2011. *Sturen op gezondheidsdoelen*. Den Haag: Raad voor de Volksgezondheid en Zorg.

Ryan, A.M. 2009a. Hospital-based pay-for-performance in the United States. *Health Economics* 18(10): 1109-1113.

Ryan, A.M. 2009b. Effects of the premier hospital quality incentive demonstration on Medicare patient mortality and cost. *Health Services Research* 44: 821-842.

Ryan, A.M. 2010. Has pay-for-performance decreased access for minority patients? *Health Services Research* 45(1): 6-23.

Ryan, A.M., Blustein, J., Casalino, L.P. 2012a. Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Affairs* 31(4): 797-805.

Ryan, A.M., Blustein, J., Doran, T., Michelow, M.D., Casalino, L.P. 2012b. The effect of phase 2 of the premier hospital quality incentive demonstration on incentive payments to hospitals caring for disadvantaged patients. *Health Services Research* 47(4): 1418-1436.

Sabatino, S., Habarta, N., Baron, R., Coates, R., Rimer, B., Kerner, J., Coughlin, S.S., Kalra, G.P., Chattopadhyay, S. 2008. Interventions to increase recommendation and delivery of screening for breast, cervical, and colorectal cancers by healthcare providers systematic reviews of provider assessment and feedback and provider incentives. *American Journal of Preventive Medicine* 35(1): S67-74.

Safran, D. G., Karp, M., Coltin, K., Chang, H., Li, A., Ogren, J., Roger, W.H. 2006. Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. *Journal of General Internal Medicine* 21(1): 13-21.

Salisbury, C., Wallace, M., Montgomery, A. 2010. Patients' experience and satisfaction in primary care: Secondary analysis using multilevel modelling. *British Medical Journal 341*: c5004.

SBEG / Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen. 2007. *Gutachten 2007 des Sachverständigenrates zur Begutachtung der Entwicklung im Gesundheitswesen. Kooperation und Verantwortung – Voraussetzungen einer zielorientierten Gesundheitsversorgung.* http://dipbt.bundestag.de/dip21/btd/16/063/1606339.pdf.

Schatz, M. 2008. Does pay-for-performance influence the quality of care? *Current Opinion in Allergy and Clinical Immunology* 8: 213-221.

Schatz, M., Blaiss, M., Green, G., Aaronson, D. 2007. Pay for performance for the allergist-immunologist: potential promise and problems. *Journal of Allergy and Clinical Immunology* 120: 769-775.

Schlingensiepen, I. 2009. Gute Reha-Ergebnisse, da gibt einen Bonus. *Ärzte Zeitung* July 23.

Schöffski, O. 2008. *Grundformen gesundheitsökonomischer Evaluationen*. In: Schöffski, O., Graf von der Schulenburg, J. (Eds.), *Gesundheitsökonomische Evaluationen* (pp. 65-94). Berlin: Springer.

Scholle, S.H., Roski, J., Adams, J.L., Dunn, D.L., Kerr, E.A., Dugan, D.P., Jensen, R.E. 2008. Benchmarking physician performance: Reliability of individual and composite measures. *The American Journal of Managed Care* 14(12): 833-838.

Scholle, S., Roski, J., Dunn, D., Adams, J., Dugan, D., Pawlson, L., Kerr, E.A. 2009. Availability of data for measuring physician quality performance. *The American Journal of Managed Care* 15(1): 67-72.

Schut, F.T., van Doorslaer, E.K.A. 1999. Towards a reinforced agency role of health insurers in Belgium and the Netherlands. *Health Policy* 48: 47-67.

Scott, I.A. 2007. Pay for performance in health care: Strategic issues for Australian experiments. *Medical Journal of Australia* 187: 31-35.

Scott, A., Sivey, P., Ait Ouakrim, D., Willenberg, L., Naccerella, L., Furler, J., Young, D. 2010. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database of Systematic Reviews* 9: CD008451.

Seddon, M.E., Marshall, M.N., Campbell, S.M., Roland, M. 2001. Systematic review of studies of quality of clinical care in general practice in the UK, Australia and New Zealand. *Quality in Health Care* 10(3): 152-158.

Setodji, C.M., Shwartz, M. 2013. Fixed-effect or Random-effect Models: What are the Key Inference Issues? *Medical Care* 51(1): 25-27.

Selby, J.V., Schmittdiel, J.A., Lee, J., Fung, V., Thomas, S., Smider, N., Crosson, F.J., Hsu, J., Fireman, B. 2010. Meaningful variation in performance: What does variation in quality tell us about improving quality? *Medical Care* 48(2): 133-139.

Sequist, T., Schneider, E., Li, A., Rogers, W., Safran, D. 2010. Reliability of medical group and physician performance measurement in the primary care setting. *Medical Care* 49(2): 126-131.

Shavell, S. 1979. Risk sharing and incentives in the principal and agent relationship. *The Bell Journal of Economics* 10: 55-73.

Shahian, D.M., Normand, S.L., Torchiana, D.F., Lewis, S.M., Pastore, J.O., Kuntz, R.E., Dreyer, P.I. 2001. Cardiac surgery report cards: Comprehensive review and statistical critique. *Annals of Thoracic Surgery* 72(6): 2155-2168.

Shekelle, P.G., Pronovost, P.J., Wachter, R.M., Taylor, S. L., Dy, S.M., Foy, R., Hempel, S., McDonald, K.M., Ovretveit, J., Rubenstein, L.V., Adams, A.S., Angood, P.B., Bates, D.W. , Bickman, L., Carayon, P., Donaldson, L., Duan, N., Farley, D.O., Greenhalgh, T., Haughom, J., Lake, E.T., Lilford, R., Lohr, K.N., Meyer, G., Miller, M.R., Neuhauser, D., Ryan, G., Saint, S., Shojania, K.G., Shortell, S.M., Stevens, D.P., Walshe, K. 2011. Advancing the science of patient safety. *Annals of Internal Medicine* 154: 693-696.

Snijders, T.A.B., Bosker, R.J. 2011. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Second edition. London: Sage Publications Ltd.

Shen, Y. 2003. Selection incentives in a performance-based contracting system. *Health Services Research* 38(2): 535-552.

Smith, A.K., Sussman, J.B., Bernstein, S.J., Hayward, R.A. 2013. Improving the reliability of physician "report cards". *Medical Care* 51(3): 266-274.

Smith, A.L. 2007. Merging P4P and disease management: how do you know which one is working? *Journal of Managed Care Pharmacy* 13: 7-10.

Solomon, L.S., Zaslavsky, A.M., Landon, B.E., Cleary, P. 2002. Variation in patient-reported quality among health care organizations. *Health Care Financing Review* 23(4): 85-100.

Sorbero, M.E., Damberg, C.L., Shaw, R., Teleki, S.S., Lovejoy, S., Decristofaro, A., Dembosky, J., Schuster, C. 2006. *Assessment of pay-for-performance options for Medicare physician services: Final report*. RAND Working Paper WR-391-ASPE. Santa Monica: RAND Health.

Sorian, R. 2006. *Measuring, Reporting, and Rewarding Performance in Health Care*. New York: The Commonwealth Fund.

Steel, N., Maisey, S., Clark, A., Fleetcroft, R., Howe, A. 2007. Quality of clinical primary care and targeted incentive payments: an observational study. *British Journal of General Practice* 57: 449-454

Steel, N., Willems, S. 2010. Research learning from the UK Quality and Outcomes Framework: a review of existing research. *Quality in Primary Care* 18(2):117–25.

Steuten, L., Lemmens, K., Vrijhoef, B. 2007. Health technology assessment of asthma disease management programs. *Current Opinion in Allergy and Clinical Immunology* 7: 242-248.

Stevens, D.P., Shojania, K.G. 2011. Tell me about the context, and more. *BMJ Quality & Safety* 20: 557-559.

Sutton, M., Nikolova, S., Boaden, R., Lester, H., Roland, M. 2012. Reduced mortality with hospital pay for performance in England. *New England Journal of Medicine* 367(19): 1821-1828.

Thaler, R. 1985. Mental accounting and consumer choice. *Marketing Science* 4(3): 199-214.

Thaler, R. 1981. Some empirical evidence on dynamic inconsistency. *Economics Letters* 8(3):201-207.

Thomas, J.W., Grazier, K.L., Ward, K. 2004. Economic profiling of primary care physicians: Consistency among risk-adjusted measures. *Health Services Research* 39(4): 985-1003.

Thomas, J.W. 2006. Economic profiling of physicians: Does omission of pharmacy claims bias performance measurement? *American Journal of Managed Care* 12(6): 341-351.

Thomas, J.W., Ward, K. 2006. Economic profiling of physician specialists: Use of outlier treatment and episode attribution rules. *Inquiry* 43(3): 271-282.

Town, R., Wholey, D.R., Kralewski, J., Dowd, B. 2004. Assessing the influence of incentives on physicians and medical groups. *Medical Care Research and Review* 61: S80-118.

Town, R., Kane, R., Johnson, P., Butler, M. 2005. Economic incentives and physicians' delivery of preventive care: a systematic review. *American Journal of Preventive Medicine* 28(2): 234-240.

Tsai, W.-C., Kung, P.-T., Khan, M., Campbell, C., Yang, W.-T., Lee, T.-F., Li, Y.-H. 2010. Effects of pay-for-performance system on tuberculosis default cases control and treatment in Taiwan. *Journal of Infection* 61: 235-243.

Tucker, A.M., Weiner, J.P., Honigfeld, S., Parton, R.A. 1996. Profiling primary care physician resource use: Examining the application of case mix adjustment. *The Journal of Ambulatory Care Management* 19(1): 60-80.

Turenne, M.N., Hirth, R.A., Pan, Q., Wolfe, R.A., Messana, J.M., Wheeler, J.R. 2008. Using knowledge of multiple levels of variation in care to target performance incentives to providers. *Medical Care* 46(2): 120-126.

Tweede Kamer. 2012. Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport voor het jaar 2013. Kamerstuk 33 400 XVI, no. 15. Den Haag: Tweede Kamer.

Van de Ven, W.P.M.M., van Vliet, R.C. J.A., Lamers, L.M. 2004. Health-adjusted premium subsidies in the Netherlands. *Health Affairs* 23(3): 45-55.

Van de Ven, W.P.M.M., Schut, F.T. 2009. Managed competition in the Netherlands: Still work-in-progress. *Health Economics* 18(3): 253-255.

Van der Lucht, F, Verweij, A. 2010. *Etniciteit en zorggebruik. In: Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid.* Bilthoven: RIVM. http://www.nationaalkompas.nl/bevolking/etniciteit/allochtonen-en-zorggebruik/.

Van der Veer, S.N., De Keizer, N.F., Ravelli, A.C.J., Tenkink, S., Jager, K.J. 2010. Improving quality of care. A systematic review on how medical registries provide information feedback to health care providers. *International Journal of Medical Informatics* 79(5): 305-323.

Van Dishoeck, A-M., Lingsma, H.F., Mackenbach, J.P., Steyerberg, E.W. 2011. Random variation and rankability of hospitals using outcome indicators. *British Medical Journal Quality and Safety* 20: 869-874.

Van Herck, P., De Smedt, D., Annemans, L., Remmen, R., Rosenthal, M.B., Sermeus, W. 2010. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research* 10(1): 247.

Van Kleef, R.C., van Vliet, R.C.J.A. 2011. Prior use of durable medical equipment as a risk adjuster for health-based capitation. *Inquiry* 47(4): 343-358

Van Tits, M.H.L., Nuyens, W.J.F.I. 1987. Een bonus-malus experiment onder huisartsen: tussentijds verslag uit Tilburg. *Medisch Contact* 42(9): 276-279.

Van Tits, M.H.L. 1989. Experiment huisartsenhonorering: belang en risico huisartsen en ziekenfondsen gedeeld. *Medisch Contact* 44: 255-257.

Vermaas, A. 1995. Doelmatigheidsopslagen voor huisartsen: het remgeld- en substitutie-effect. *Medisch Contact* 50(2): 47-48.

Vermaas, A. 2006. *Agency, Managed Care and Financial-Risk Sharing in General Medical Practice*. Dissertation. Erasmus University Rotterdam. Enschede: Gildeprint Drukkerijen.

Volpp, K.G., Pauly, M.V., Loewenstein, G., Bangsberg, D. 2009. P4P4P: an agenda for research on pay-for-performance for patients. *Health Affairs* 28: 206–214.

Vosselman, E.G.J. 1996. *De Agentschapstheorie*. In: Vosselman, E.G.J. (Ed.), O*ntwerp van management control-systemen: een economische benadering* (Chapter 7.4). Deventer: Kluwer Bedrijfswetenschappen.

Walker, S., Mason, A.R., Claxton, K., Cookson, R., Fenwick, E., Fleetcroft, R., Sculpher, M. 2010. Value for money and the Quality and Outcomes Framework in primary care in the UK NHS. *British Journal of General Practice* 60: 213-220.

Walle, E. 2009. ERGOV: Ergebnisorientierte Vergütung in der neurologischen Rehabilitation. Hopfen am See: m&i- Klinikgruppe Enzensberg.

Wang, H., Zhang, L., Yip, W., Hsiao, W. 2011. An experiment in payment reform for doctors in rural China reduced some unnecessary care but did not lower total costs. *Health Affairs* 30: 2427-2436.

Ward, M., Daniels, S.A., Walker, G.J., Duckett, S. 2007. Connecting funds with quality outcomes in health care: A blueprint for a clinical practice improvement payment. *Australian Health Review* 31(1): S54-58.

Wennberg, J.E. 2010. *Tracking medicine: a researcher's quest to understand health care*. New York: Oxford University Press.

Werner, R.M., Bradlow, E.T. 2010. Public reporting on hospital process improvements is linked to better patient outcomes. *Health Affairs* 29: 1319-1324.

Werner, R.M., Goldman, L.E., Dudley, R.A. 2008. Comparison of change in quality of care between safety-net and non-safety-net hospitals. *Journal of the American Medical Association* 299(18): 2180-2187.

West, D. 2008. Advancing quality in the North West. *Health Service Journal* November 27.

Weyer, S.M., Bobiak, S., Stange, K.C. 2008. Possible unintended consequences of a focus on performance: Insights over time from the research association of practices network. *Quality Management in Health Care* 17(1): 47-52.

Wheeler, J.R., White, B., Rauscher, S., Nahra, T.A., Reiter, K.L., Curtin, K.M., Damberg, C.L. 2007. Pay-for-performance as a method to establish the business case for quality. *Journal of Health Care Finance* 33: 17-30.

Wikipedia. *Hospital Italiano de Buenos Aires*. http://en.wikipedia.org/wiki/Hospital_Italiano_de_Buenos_Aires

Williams, T., Yanagihara, D., McDermott, S., Rose, L. 2009. *The California Pay for Performance Program. The second chapter. Measurement years 2006-2009*. Oakland: Integrated Healthcare Association.

Wilson, J.F. 2007. Lessons for health care could be found abroad. *Annals of Internal Medicine* 146: 473-476.

Wilson, R. 2006. Primary care renewal in Ontario—Focus on remuneration. Presentation at the Primary Care Forum, College of Family Physicians of Canada (November), Ontario, Canada.

Witter, S., Fretheim, A., Kessy, F.L., Lindahl, A.K. 2012. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database of Systematic Reviews* 2: CD007899.

Young, G.J., Conrad, D.A.. 2007. Practical issues in the design and implementation of pay-for-quality programs. *Journal of Healthcare Management* 52(1): 10-18.

Young, G.J., White, B., Burgess Jr., J.F., Berlowitz, D., Meterko, M., Guldin, M.R., Bokhour, B.G. 2005. Conceptual issues in the design and implementation of pay-for-quality programs. *American Journal of Medical Quality* 20(3): 144-150.

Zaslavsky, A.M., Epstein, A.M. 2005. How patients' sociodemographic characteristics affect comparisons of competing health plans in California on HEDIS quality measures. *International Journal for Quality in Health Care* 17(1): 67-74.

Zhang, H., Lu, N., Feng, C., Thurston, S.W., Xia, Y., Zhu, L., Tu, X.M. 2011. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine* 30: 2562-2572.

SAMENVATTING

In veel landen is de verlening van gezondheidszorg suboptimaal. Zo is vaak aantoonbaar sprake van een tekortschietende kwaliteit van zorg en onnodig gebruik van (dure) zorg. Perverse prikkels in de bekostiging van zorgaanbieders (zoals bij betalingen per verrichting) worden gezien als een belangrijke oorzaak van dit probleem. Bij de hervorming van bestaande bekostigingssystemen wordt in steeds meer landen gekozen voor een systeem gebaseerd op *pay-for-performance* (P4P), een bekostigingsprincipe waarbij zorgaanbieders expliciete financiële prikkels ontvangen voor het verbeteren van de kwaliteit en doelmatigheid van zorg. Het afgelopen decennium is wereldwijd de interesse in P4P sterk toegenomen. Omdat zorgaanbieders reageren op financiële prikkels en omdat de afgelopen jaren forse vooruitgang is geboekt met het meten van kwaliteit van zorg met behulp van indicatoren, wordt P4P door veel beleidsmakers gezien als een veelbelovende en logische verbeterstrategie. Desondanks blijkt uit internationaal onderzoek dat P4P in de meeste gevallen nog niet heeft geresulteerd in de gewenste verbeteringen. Dit beperkte succes is voor een belangrijk deel een gevolg van onvoldoende kennis over hoe P4P het beste kan worden vormgegeven en geïmplementeerd. *Het doel van dit proefschrift is om inzicht te verkrijgen in cruciale conceptuele en praktische aspecten van de vormgeving en implementatie van P4P en om daarover aanbevelingen te doen.* Dit doel wordt gerealiseerd door (i) een theoretische verkenning van de kernaspecten van de vormgeving en implementatie van P4P, (ii) analyse van empirische literatuur over (onbedoelde) effecten van P4P en (iii) beantwoording van enkele belangrijke empirische vragen ten aanzien van het meten van de prestaties van zorgaanbieders.

In een aantal landen (bijvoorbeeld de Verenigde Staten en het Verenigd Koninkrijk) wordt al jaren geëxperimenteerd met P4P in de zorg. Ondanks deze ervaring is er nog weinig bekend over welke aspecten van de vormgeving en implementatie van belang zijn voor het behalen van gewenste effecten. Daarnaast is nog weinig bekend over hoe P4P-programma's in de praktijk worden vormgegeven en geïmplementeerd. In **hoofdstuk 2** wordt op basis van theoretische en empirische literatuur een aantal kernaspecten van de vormgeving en implementatie van P4P geïdentificeerd. Deze kernaspecten zijn geclassificeerd in drie categorieën: *wat* te belonen? (definitie en meting van de prestatie, casemix-correctie en betrokkenheid van zorgaanbieders), *wie* te belonen? (individuele artsen versus groepen van artsen) en *hoe* te belonen? (positieve versus negatieve prikkels, omvang van de betaling, aantal en type prestatiedoelen, frequentie van betalen en duur van het P4P-programma). De analyse laat zien dat de vormgeving en implementatie van een P4P-programma complexe aangelegenheden zijn, waarbij rekening moet worden gehouden met veel verschillende aspecten en mogelijke valkuilen. Het is bijvoorbeeld mogelijk dat zorgaanbieders zich vooral

zullen richten op die onderdelen van de zorg waarvoor beloningen te behalen zijn (zoals het volgen van medische richtlijnen en protocollen). Dit zou dan ten koste kunnen gaan van de onbeloonde onderdelen (zoals patiënttevredenheid en continuïteit van zorg). Een andere mogelijke valkuil is dat P4P zorgaanbieders zou kunnen aanzetten tot het aantrekken van relatief gezonde patiënten voor wie het gemakkelijker is om goede prestaties te leveren (risicoselectie). In het hoofdstuk wordt een aantal conclusies getrokken over hoe bepaalde aspecten van de vormgeving en implementatie van P4P zouden kunnen bijdragen aan gewenste effecten. Zo verdienen absolute prestatiedoelen (bijvoorbeeld: bij minimaal 80 procent van de ingeschreven diabetespatiënten wordt regelmatig de bloedsuikerwaarde gecontroleerd) de voorkeur boven relatieve prestatiedoelen (bijvoorbeeld: behoren tot de 20 procent zorgaanbieders met het hoogste percentage diabetespatiënten bij wie regelmatig de bloedsuikerwaarde wordt gecontroleerd). Daarnaast is een goede correctie voor relevante patiëntkenmerken (zoals leeftijd en gezondheid) cruciaal, vooral voor uitkomstindicatoren (dat wil zeggen: indicatoren die betrekking hebben op de uitkomsten van de zorg, zoals complicaties tijdens of na een operatie) en zorgkosten. In **hoofdstuk 3** zijn bestaande P4P-programma's geïdentificeerd aan de hand van een systematisch literatuuronderzoek. In totaal zijn dertien programma's uit negen landen (waaronder Nederland) geanalyseerd, gebruikmakend van de inzichten uit hoofdstuk 2. Vrijwel alle programma's belonen zorgaanbieders voor medisch-inhoudelijke kwaliteit. Hierbij spelen uitkomstindicatoren geen of slechts een beperkte rol. Daarnaast hanteren de meeste programma's alleen positieve prikkels (beloningen), betrekken zij zorgaanbieders actief bij de vormgeving en richten zij zich voornamelijk op zorgaanbieders in de eerstelijn. Er zijn ook belangrijke verschillen, bijvoorbeeld ten aanzien van de gebruikte prestatie-indicatoren, het gebruik van methoden ter beperking van het (financieel) risico voor zorgaanbieders, de omvang van de betalingen, en het type en aantal prestatiedoelen. Hoewel de heterogeniteit in de vormgeving ook veroorzaakt wordt door contextuele verschillen, lijkt deze vooral een gevolg van praktische implementatieproblemen en onvoldoende kennis over wat werkt in de praktijk. De vormgeving en implementatie van bestaande P4P-programma's zijn in de meeste gevallen voor verbetering vatbaar, vooral ten aanzien van inbouwen van waarborgen ter voorkoming van ongewenste neveneffecten.

Vanaf eind jaren 90 heeft het empirisch onderzoek naar de effecten van P4P een grote vlucht genomen. De resultaten hiervan zijn gebundeld in verschillende literatuuroverzichten. In deze overzichten blijft de kosteneffectiviteit van P4P meestal buiten beschouwing. Dat is opvallend aangezien P4P gebruik maakt van alternatief aanwendbare middelen. Dit onderwerp verdient daarom meer aandacht. In **hoofdstuk 4** wordt een systematisch literatuuronderzoek naar de kosteneffectiviteit van P4P-programma's gepresenteerd. Van de negen geïncludeerde studies konden er drie worden geclassificeerd als economische evaluaties waarbij een expliciete link wordt gelegd tussen kosten en effecten, bijvoorbeeld door middel van het berekenen van kosteneffectiviteitsratio's. Over het geheel genomen laten de

resultaten zien dat P4P kosteneffectief kan zijn. Het bewijs is echter allerminst overtuigend. Zo laten veel studies relevante kostensoorten buiten beschouwing en/of analyseren zij slechts één of enkele prestatie-indicatoren. Daarnaast is de methodologische kwaliteit van veel studies beperkt (bijvoorbeeld door afwezigheid van een controlegroep) en zijn er grote verschillen tussen de geëvalueerde P4P-programma's. Om deze redenen kan een definitieve conclusie over de kosteneffectiviteit van P4P niet worden getrokken. **Hoofdstuk 5** bevat een uitgebreid literatuuronderzoek naar effecten van P4P. De bestaande literatuuroverzichten verschillen nogal in focus en daarom ook in de getrokken conclusies. Als gevolg hiervan is de kennis ten aanzien van effecten van P4P nog betrekkelijk onoverzichtelijk. Het doel van hoofdstuk 5 is om de informatie uit bestaande overzichten te synthetiseren. De meeste empirische studies hebben gekeken naar het effect van P4P op de kwaliteit van preventieve en chronische zorg in de eerstelijn. De resultaten van experimenteel onderzoek laten een inconsistent beeld zien: sommige studies vonden (kleine) verbeteringen in kwaliteit van zorg, terwijl andere studies geen effect vonden. Observationele studies vonden meestal verbeteringen voor minimaal één indicator waarbij de resultaten positiever worden naarmate de methodologische kwaliteit van de studies afneemt. De effectiviteit van P4P lijkt hoger als (i) indicatoren worden gebruikt met veel verbeterpotentieel, (ii) het programma gericht is op individuele artsen of kleine groepen, (iii) de betalingen zijn gebaseerd op absolute prestaties en (iv) het programma wordt vormgegeven en geïmplementeerd in samenspraak met betrokken zorgaanbieders. Over het geheel genomen is het empirische bewijs op dit moment echter onvoldoende om brede toepassing van P4P aan te kunnen bevelen, temeer omdat in een aantal studies is gevonden dat P4P weinig toevoegt als aanbieders reeds worden voorzien van feedback op hun prestaties. Daarnaast is in een aantal studies empirisch aangetoond dat P4P ongewenste neveneffecten kan hebben, zoals een verschraling van de onbeloonde onderdelen van de zorg(kwaliteit). Overigens lijkt P4P ook positieve neveneffecten te kunnen hebben. Zo is in het Verenigd Koninkrijk de ongelijkheid in de kwaliteit van de verleende zorg tussen sociaaleconomische groepen afgenomen na de introductie van een grootschalig P4P-project.

Bij het meten en vergelijken van de kwaliteit en doelmatigheid van zorgaanbieders is het van belang om te corrigeren voor verschillen in patiëntenpopulaties (casemix) tussen zorgaanbieders. Een goed statistisch model is hiervoor essentieel. Met een dergelijk model kan per zorgaanbieder een vooraf verwacht prestatieniveau worden bepaald waarmee de werkelijke prestatie achteraf kan worden vergeleken. Vervolgens kan de beloning aan de zorgaanbieder afhankelijk worden gesteld van het verschil tussen de verwachte prestatie en de werkelijke prestatie. Het is dus van belang dat het statistische model de werkelijkheid goed weergeeft. De entiteit die de metingen uitvoert (bijvoorbeeld een zorgverzekeraar) zal de voorkeur geven aan relatief eenvoudige statistische methoden die gemakkelijk te implementeren, onderhouden en uit te leggen zijn. Een voorbeeld hiervan is de kleinstekwadratenmethode. Prestatiegegevens in de zorg voldoen echter vaak niet aan de statistische voorwaarden om

dergelijke eenvoudige methoden te kunnen toepassen. Niettemin is het goed mogelijk dat de verschillen in uitkomsten tussen eenvoudige en complexe methoden verdwijnen wanneer de verwachte waarden worden bekeken op het niveau van de zorgaanbieder in plaats van op patiëntniveau. De mate waarin dit het geval is, is onderzocht in **hoofdstuk 6**. Door middel van analyses op declaratiegegevens van een grote Nederlandse zorgverzekeraar is nagegaan of verschillende statistische methoden (waaronder de kleinstekwadratenmethode) leiden tot verschillende rangschikkingen van huisartsen en gezondheidscentra ten aanzien van hun scores op diverse indicatoren voor kwaliteit van zorg, zorggebruik en kosten. Uit de analyses blijkt dat er inderdaad sprake is van verschillen, vooral bij uitkomstindicatoren en zorgkosten. De verschillen zijn echter vrij klein, en mogelijk klein genoeg voor een verzekeraar om te kiezen voor de eenvoudige methode. Voorzichtigheid is echter wel degelijk geboden omdat de verschillen vooral optreden in de extremen van de rangschikkingen. De methoden verschillen dus vooral in welke zorgaanbieders worden bestempeld als zeer goed en welke als zeer slecht.

Om misclassificatie van zorgaanbieders (en daardoor onjuiste verdeling van beloningen) als gevolg van toevalsfluctuaties te voorkomen dienen prestatiemetingen voldoende betrouwbaar te zijn. Een betrouwbare prestatievergelijking van zorgaanbieders vereist zowel voldoende patiënten per zorgaanbieder als voldoende variatie in prestaties tussen zorgaanbieders. Prestatievergelijkingen in de zorg zijn vaak gericht op individuele artsen en worden meestal uitgevoerd door individuele private zorginkopers, zoals zorgverzekeraars. Op deze manier is een betrouwbare vergelijking echter twijfelachtig door de relatief kleine patiëntaantallen per zorgaanbieder. In **hoofdstuk 7** is onderzocht in hoeverre individuele huisartsen betrouwbaar kunnen worden vergeleken ten aanzien van hun scores op dertien verschillende indicatoren die zijn afgeleid van de declaratiegegevens van één zorgverzekeraar. De analyse laat zien dat dit alleen mogelijk is voor indicatoren waarvan huisartsen de uitkomsten direct kunnen beïnvloeden. Voor vijf indicatoren (twee proces- en drie zorggebruik-indicatoren) is de betrouwbaarheid onvoldoende om bruikbare input te kunnen leveren voor verbeterinitiatieven. Deze indicatoren hebben vooral betrekking op het gebruik van ziekenhuiszorg, zowel in het algemeen als door patiënten met specifieke chronische aandoeningen. Voor vijf andere indicatoren (twee proces- en drie zorggebruik-indicatoren) is de betrouwbaarheid voldoende voor gebruik voor niet-vergaande toepassingen, zoals het geven van feedback aan zorgaanbieders op hun prestaties. Voor drie indicatoren (verschillende typen huisartskosten) is de betrouwbaarheid tevens hoog genoeg voor vergaande toepassingen, zoals P4P en het openbaar maken van de scores. De hoge betrouwbaarheid voor deze indicatoren is niet alleen een gevolg van de relatief grote patiëntaantallen, maar ook een gevolg van de substantiële variatie tussen huisartsen. Zelfs voor deze indicatoren is voorzichtigheid echter geboden indien zij worden gebruikt voor P4P. Zo moet ervoor worden gewaakt dat huisartsen niet worden beloond voor het onnodig doorverwijzen van patiënten naar de tweedelijn en niet worden gestraft voor het zoveel mogelijk behande-

len van patiënten in de eerstelijn. Voor de overige indicatoren zou de betrouwbaarheid verhoogd zou kunnen worden door het vergroten van de patiëntaantallen. Hiervoor zijn verschillende opties, zoals het combineren van de gegevens van meerdere zorginkopers. Het is vooralsnog echter onduidelijk wat hiervan de invloed zal zijn op de variatie tussen en binnen zorgaanbieders, die beide ook bepalend zijn voor de betrouwbaarheid. Tevens zal een toename in patiëntaantallen vaak niet genoeg zijn om voldoende betrouwbaarheid te realiseren. De nadruk zal daarom vooral moeten liggen op het gebruiken en ontwikkelen van indicatoren met substantiële systematische variatie tussen zorgaanbieders.

P4P wordt in toenemende mate ook in Nederland toegepast. Tot nu toe is dit echter beperkt gebleven tot tijdelijke en relatief kleinschalige experimenten. De Raad voor de Volksgezondheid en Zorg heeft in 2011 aanbevolen om hier verandering in aan te brengen. Ook de huidige Minister van Volksgezondheid, Welzijn en Sport (Schippers) onderschrijft de noodzaak tot hervorming van de bestaande bekostigingssystemen en heeft aangekondigd de komende jaren in te gaan zetten op de ontwikkeling en implementatie van innovatieve bekostigingssystemen gericht op het stimuleren van goede uitkomsten van medische behandelingen in termen van zowel kwaliteit als kosten. Dit proefschrift laat echter zien dat de verwachtingen ten aanzien van het rendement op de investeringen in P4P niet overschat moeten worden. Daarnaast is transparantie in zorgkwaliteit een cruciale voorwaarde voor een succesvolle toepassing van P4P waar in Nederland op dit moment nog onvoldoende aan is voldaan. Ook hebben zorgverzekeraars te maken met potentieel meeliftgedrag van concurrenten wat hen mogelijk terughoudend maakt bij het doen van investeringen in prestatie-indicatoren en P4P. Het is in dit kader van belang dat aandacht wordt besteed aan het ontwikkelen en ontsluiten van gestandaardiseerde sets van indicatoren (op gecoördineerde wijze met input van zorgaanbieders, patiëntenorganisaties, zorgverzekeraars en overheidsinstanties) die door alle zorgverzekeraars en zorgaanbieders gebruikt kunnen worden. Het nieuwe Kwaliteitsinstituut (onderdeel van het Zorginstituut Nederland, voorheen het College voor Zorgverzekeringen) zou hierbij een belangrijke faciliterende rol kunnen spelen. Verder vereisen grote verbeteringen in kwaliteit en doelmatigheid effectieve zorgcoördinatie en afstemming tussen verschillende typen zorgaanbieders. Het gebruik van prospectieve betalingen per zorgbundel (dat wil zeggen: een bundeling van verschillende typen zorg) gecombineerd met effectieve P4P voor kwaliteit lijkt een veelbelovende strategie om dit te realiseren. In Nederland zijn in dit kader reeds belangrijke stappen gezet met de invoering van diagnose behandeling combinaties (DBC's) in zowel de tweedelijns- als in de eerstelijnszorg. Grote uitdagingen voor de toekomst zijn het slaan van een brug tussen de bestaande bekostigingsvormen richting *integrale* bekostiging en het zetten van de stap van betaling per *patiënt* naar betaling per *verzekerde*. Hierbij dient goed gekeken te worden naar ervaringen uit andere landen, zoals de Verenigde Staten, het Verenigd Koninkrijk en Duitsland.

Het onderzoek in dit proefschrift leidt tot een aantal suggesties voor vervolgonderzoek. Om meer inzicht te krijgen in welke aspecten van de vormgeving en implementatie van P4P

in de praktijk wel en welke niet bijdragen aan gewenste effecten is kwantitatief en kwalitatief empirisch onderzoek noodzakelijk. Het is in dit kader van belang dat P4P-programma's uitgebreid worden geëvalueerd, gebruikmakend van controlegroepen. Idealiter wordt hierbij ook de kosteneffectiviteit en de lange-termijn invloed op gezondheidsuitkomsten meegenomen. Nieuwe programma's dienen eerst op kleine schaal te worden getest (bijvoorbeeld binnen een bepaalde regio voor een beperkt aantal zorgaanbieders), waarna bij bewezen positieve resultaten gefaseerde uitbreiding mogelijk is. Ten aanzien van de resultaten gevonden in hoofdstuk 6 en 7 dient te worden nagegaan of deze worden bevestigd in andere omgevingen (bijvoorbeeld de tweedelijnszorg) en voor andere indicatoren. Ook is meer inzicht nodig in de voor- en nadelen van specifieke methoden voor het verhogen van de betrouwbaarheid van prestatiemetingen. Tenslotte is onderzoek nodig naar de vormgeving en implementatie van statistische methoden voor casemix-correctie. Hoewel dit proefschrift laat zien dat een dergelijke correctie van groot belang is bij het vergelijken van zorgaanbieders op kwaliteit en kosten, is nog relatief weinig bekend over hoe dergelijke methoden op indicatorniveau dienen te worden vormgegeven en geïmplementeerd om risicoselectie te voorkomen en zowel een eerlijke als betekenisvolle vergelijking te waarborgen.

DANKWOORD

Najaar 2007, een e-mail van professor van de Ven verschijnt op mijn scherm met de vraag of ik geïnteresseerd zou zijn in het uitvoeren van een vooronderzoek voor een promotietraject. Ik moest maar eens langskomen voor een oriënterend gesprek en alvast nadenken over een onderwerp, mits ik geïnteresseerd was natuurlijk. Ik besloot op het aanbod in te gaan en startte begin 2008 als student-assistent bij het iBMG. Ik had lange tijd niet gedacht dat ik het promotieonderzoek vervolgens zelf zou gaan uitvoeren, ondanks dat mijn begeleider (Wynand van de Ven) al een aantal keren had geopperd dat dit wellicht een logisch vervolg zou zijn. Uiteindelijk ben ik de uitdaging aan gegaan. Nu, zes jaar na het e-mailtje, kan ik met een gerust hart zeggen dat het traject met succes is afgerond, iets waar ik behoorlijk trots op ben.

Zonder de steun van verschillende personen had ik het echter nooit gekund. Die personen wil ik hier graag bedanken. In de eerste plaats natuurlijk mijn promotor Wynand en copromotor René. Wynand, heel veel dank voor het in mij gestelde vertrouwen. Jouw deur stond en staat altijd open en je prettige en informele manier van begeleiden zijn ontzettend belangrijk geweest bij de afronding van mijn proefschrift. Ondanks je drukke agenda ben je er altijd in geslaagd om mijn teksten binnen korte tijd van gedetailleerd en constructief commentaar te voorzien. En René, jouw kennis van statistiek, econometrie en SAS zijn van groot belang geweest bij de totstandkoming van de twee empirische hoofdstukken van het proefschrift. Ik ben blij dat we onze samenwerking voorlopig nog even kunnen voortzetten.

Verder ben ik veel dank verschuldigd aan de leden van de beoordelingscommissie (Prof. dr. F.T. Schut, Prof.dr. E.W. Steyerberg en Dr. J.C.C. Braspenning) en de promotiecommissie (Prof.dr. D.H. de Bakker en Prof.dr. N.S. Klazinga) voor het lezen en beoordelen van mijn proefschrift en voor het opponeren tijdens de verdediging.

Ook wil ik Martin Emmert, Heike Kemter, Manfred Scheppach, Oliver Schöffski en Susanne Esslinger bedanken voor hun bijdrage aan twee hoofdstukken van mijn proefschrift. Martin, all the hours on Skype haven't been for nothing! Thanks a lot for all your work and for involving me in your research. I'm sure we'll stay in touch!

Een belangrijk deel van dit proefschrift had nooit tot stand kunnen komen zonder de door Achmea beschikbaar gestelde databestanden. In het bijzonder wil ik Matthijs Hagenaars en Geert Groenenboom hartelijk bedanken voor hun openheid en het in mij gestelde vertrouwen. Geert, bedankt voor alle inkijkjes en natuurlijk voor je bijdrage aan het artikel!

Ik wil hier ook graag mijn collega's bedanken, in het bijzonder die van de sectie ZKV: Anne-Fleur, Daniëlle, Erik, Ilaria, Kayleigh, Lieke, Marco, Ramsis, Richard, Rudy, Stéphanie, Suzanne, Trea en Weiwei. Dank jullie wel voor de fijne werkomgeving en de gezellige tijden

tijdens de lunchpauzes en congressen. Anne-Fleur, jouw gezelligheid heeft veel bijgedragen aan de fijne tijd die ik de afgelopen jaren bij ZKV heb gehad. Ik weet zeker dat er binnen afzienbare tijd ook een prachtig boek met jouw naam erop van de pers rolt!

Verder wil ik, zonder hier iedereen te noemen, mijn vaste vriendenclub bedanken. Michael, Remco, Arjen en Marjolein, Maarten en Karlijn, Marcel en Kim: bedankt voor de leuke weekendjes, festivals, concerten en andere uitstapjes! Ik weet niet of ik het zonder die momenten gered zou hebben. Remco en Robin, dank voor de vele uren op de squashbaan (voor mij onmisbaar, er zullen er nog veel volgen!) en de leuke en 'diepgaande' gesprekken en discussies over tennis en voetbal.

Dan natuurlijk mijn familie. Pa en ma, Paul en Sonja (en Fleur), Ellen en Wilrik (en Max), Gillis en Jacoline en Remco, bedankt voor jullie steun de afgelopen jaren en voor de getoonde interesse in mijn onderzoek. Ik kan me geen betere (schoon)familie wensen! En Gillis, Jacoline en Remco: die wisselbeker blijft voorlopig nog in huize Eijkenaar!

Tenslotte Anita: jouw steun, oprechte interesse en geduld zijn voor mij onmisbaar geweest, niet alleen de afgelopen vier jaar. Ik weet zeker dat ik sommige hordes zonder jouw duwtjes in de rug nooit had durven nemen. Onze reizen naar Australië, Canada en vele andere plaatsen zie ik nog altijd als de mooiste tijden van mijn leven. Maar vooral omdat ze met jou waren. Ik kan niet wachten om met jou de rest de wereld te gaan verkennen. Dank je wel voor alles!

CURRICULUM VITAE

Frank Eijkenaar (1986) is a research fellow at the institute of Health Policy and Management (iBMG), Erasmus University Rotterdam. From 2004 to 2008 he studied Health Sciences at Erasmus University and obtained his master's degree (cum laude) in health economics, policy, and law. During 2008, he also worked on an international explorative study on the theoretical basis and effects of financial-risk sharing in primary care, which also included a proposal for subsequent PhD-research.

After graduation, Frank started working as a PhD-student at iBMG. From 2009 to 2013 he worked on his dissertation and published his research in peer-reviewed scientific journals, including the European Journal of Health Economics (2), Health Policy (1), Medical Care (1), Medical Care Research and Review (1), and Medical Decision Making (1). In addition, he presented his work at several international scientific conferences, including the National Pay for Performance Summit in 2010 (San Francisco, United States), the World Congress on Health Economics in 2011 (Toronto, Canada), the European Conference on Health Economics in 2012 (Zurich, Switzerland), and the Risk Adjustment Network Meeting in 2012 (Lucerne, Switzerland). In addition to the research included in his dissertation, Frank also worked on an international comparative study on outcome-based payment in health care, commissioned by the Dutch Ministry of Health. After completion of his dissertation, his research continued to focus on the design and implementation of bundled payment and pay-for-performance programs, as well as on performance measurement and risk adjustment.

As a teacher, Frank has been involved in the bachelor program Health Sciences and the master program Health Economics, Policy, and Law at Erasmus University where he gives workgroups for the courses Health Economics and Multivariate Analysis (both bachelor) and lectures for the courses Health Insurance and Healthcare Systems (bachelor) and Economics and Financing of Healthcare Systems (master). In addition, he (co-)supervises bachelor and master theses.

PHD PORTFOLIO

PhD student: Frank Eijkenaar
Department: institute of Health Policy and Management
PhD period: 2009-2013
Promotor: Prof.dr. W.P.M.M. van de Ven
Copromotor: Dr. R.C.J.A. van Vliet

Training

- SAS programming I and II, SAS institute, Huizen (2009)
- Regression analysis (Erasmus summer program), Erasmus University Rotterdam (2011)
- Health Policy Workshop: physician incentives, Ministry of VWS, The Hague (2012)
- Academic writing for graduate students, Erasmus University Rotterdam (2011)
- Klaar in vier jaar, Erasmus University Rotterdam (2010)
- How to make effective presentations, Erasmus University Rotterdam (2011)
- Geven van onderwijs, kleine groepen, Erasmus University Rotterdam (2012)

Conferences

- LOLA Health Economics Study Group, Maastricht (2009)
- Academy Health Annual Research Meeting, Chicago (2009)
- 5th World Congress of the international Health Economics Association, Beijing (2009)
- Risk Adjustment Network Meeting, Jerusalem (2009)
- 5th National Pay For Performance Summit, San Francisco (2010)
- 6th World Congress of the international Health Economics Association, Toronto (2011)
- Risk Adjustment Network Meeting, Lucerne (2012)
- 9th European Conference on Health Economics, Zürich (2012)

Presentations

- 6th World Congress of the international Health Economics Association, Toronto (2011)
- 9th European Conference on Health Economics, Zürich (2012)
- Risk Adjustment Network Meeting, Lucerne (2012)
- Erasmus University Rotterdam, iBMG seminar (2010 and 2012)
- Achmea Zorg en Gezondheid, Amersfoort, seminar (2011)
- Dutch Healthcare Authority (NZa), Utrecht, seminar (2011)
- Ministry of VWS, The Hague, seminar (2012)
- iBMG alumni event Zorg voor Kennis II, Rotterdam, seminar (2012)
- Dutch Hospital Association (NVZ), Utrecht, seminar (2013)
- VGZ, Alkmaar, seminar (2013)

Teaching at iBMG

- Health economics, first year bachelor, workgroups (2009-2013)
- Multivariate analysis (M&T 3), second year bachelor, workgroups (2008-2013)
- Health insurance & healthcare systems, third year bachelor and "schakel", lecturer (2010-2013)
- Economics & financing of healthcare systems, master Health Economics, Policy, and Law, lecturer (2011-2013)
- Statistics and quantitative research, "schakel aanschuifvariant", supervisor (2010-2012)
- Bachelor and Master theses, (co-)supervisor (2010-2013)