Eur J Health Econ (2013) 14:117–131
DOI 10.1007/s10198-011-0347-6

ORIGINAL PAPER

Key issues in the design of pay for performance programs

Frank Eijkenaar

Received: 26 October 2010 / Accepted: 9 August 2011 / Published online: 1 September 2011

Abstract Pay for performance (P4P) is increasingly being used to stimulate healthcare providers to improve their performance. However, evidence on P4P effectiveness remains inconclusive. Flaws in 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. 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 encourages "selection" and "teaching to the test" should not be dismissed, (3) sophisticated risk adjustment is important, especially in outcome and resource use measures, (4) involving providers in program design is vital, (5) on balance, group incentives are preferred over individual incentives, (6) whether to use rewards or penalties is context-dependent, (7) payouts should be frequent and low-powered, (8) absolute targets are generally preferred over relative targets, (9) multiple targets are preferred over single targets, and (10) P4P should be a permanent component of provider compensation and is ideally "decoupled" form 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.

Keywords Compensation methods · Incentive reimbursement · Pay for performance · Program design · Provider behavior

F. Eijkenaar (✉)
Institute of Health Policy and Management,
Erasmus University Rotterdam, Burgemeester Oudlaan 50,
3000 DR Rotterdam, The Netherlands
e-mail: eijkenaar@bmg.eur.nl

JEL classification D03 · D86 · I11 · J31 · J33

Introduction

In many countries, healthcare delivery is suboptimal. For example, McGlynn et al. [70] have shown that in the United States (US) adherence to recommended care processes is near 50 percent. In the Netherlands, this is about 67 percent, but there is large variation among providers and among specific guidelines [45]. Similar deficits were found in the United Kingdom (UK), Australia, and New Zealand [91]. 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 ([26, 41, 42, 51, 97]) 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 outcomes while mitigating unintended consequences (such as increasing disparities). By contributing to better prevention and disease management, as well as by including efficiency measures, if effective, P4P could also mitigate cost growth.

P4P is now widely being applied in the United States and the United Kingdom [4, 85, 89] and increasingly being implemented in many other countries [5, 7, 24, 46, 66, 84].

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 [13, 17, 78, 87]. Moreover, unintended and undesired effects of P4P have been demonstrated [9, 10, 34, 62, 68, 92, 99]. Nonetheless, in general, the potential of P4P to improve performance remains undisputed. There is consensus that the way in which P4P is designed has important consequences for the incentives that physicians experience and how they might respond to them [71]. As argued by several authors, the fact that P4P has not been very successful has partly been a consequence of flaws in program design [68, 78, 86, 87]. 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 of P4P programs. Other authors have already provided important contributions in this area [16, 17, 71, 86, 97, 102]. 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 providers or groups of providers ("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 reimbursement system, whether the program should pay for absolute or relative performance, the frequency of payments, and the duration of 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 to a great extent the success of the program.

What to incentivize: how is performance defined?

Dimensions and measurement of performance

Depending on the goals of the stakeholders involved, programs will vary in how "good performance" is defined [59]. Cost and utilization control were the main focus of early P4P programs in the United States (e.g., [74]), 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, payers and 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 registration system for diabetics), processes (e.g., regularly performing blood sugar checks on diabetics), and (intermediate) outcomes (e.g., optimal blood sugar levels in diabetics) [19]. 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 [26]. 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 [100].

The number and characteristics of included performance measures are likely to affect the eventual effect of the program on overall performance [97]. If a program only includes one or a few measures pertaining to one specific performance aspect (e.g., diabetes care), this could result in a disproportionate focus on a specific behavior (i.e., improving care for diabetics). 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 [97]. Thus, a balance is needed between "narrow and shallow" and "broad and deep." It also seems important to combine objective measures (e.g., adherence to clinical guidelines) with subjective measures such as patient satisfaction and continuity of care [38]. 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 [4, 17, 88, 93]. 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 complement claims data, purchasers may require providers to provide additional performance information based on extractions of medical records and by

administering patient satisfaction 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 [65]. 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 United Kingdom, primary care practices receive substantial financial rewards for scoring well on a large number of performance measures [85]. For each practice, performance information is extracted automatically via a uniform system of EMRs. 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 United Kingdom, primary care practices were largely compensated for health IT [22], 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 United States, EMRs are increasingly used for the purpose of data collection, although still on a relatively small scale [17, 65, 93].

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 case mix. 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 demographic characteristics, socioeconomic status, 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 [17]. 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 [17, 88, 93].

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. [12], "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 [54]. 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 [23]. 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 [63, 76]. In addition to clinical outcomes, this will often also hold for measures of utilization and resource use [54, 63, 73, 76].

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 United Kingdom, 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" [20, 43], although more research is needed to confirm this. Extensive inspections and severe penalties for fraud may have contributed to preventing this behavior.

Risk selection is not just a theoretical concept. Hofer et al. [54] showed empirically that the easiest way for physicians being profiled on the blood sugar levels of their diabetic 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 year t-1 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 year t. In their analysis, about half of this improvement was due to patient selection. Shen [92] 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. The data showed that after implementation of performance-based contracting, 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 [23]: 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 [68], 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 both 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 [10].

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 [38, 55]. 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 leads to positive spillover effects on unincentivized aspects of performance. As noted by Mullen et al. [75], "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 HMO, Mullen et al. [75] did not find evidence of positive or negative spillovers on unincentivized aspects of care, although some rewarded performance measures improved. Another US study [39] found that among hospitals participating in a quality-improvement program, P4P had limited incremental impact on quality of care for acute myocardial infarction. 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 other studies have addressed teaching to the test with respect to the QOF in the United Kingdom, with more than 130 measures in about thirty different areas the most comprehensive P4P program in the world. Steel et al. [94] found neither improvement nor deterioration in unincentivized conditions. However, Campbell et al. [9] 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 United States, while purchasers underscore the importance of a broad set of measures, sets are typically narrow [4, 88]. However, the somewhat stronger evidence of teaching to the test in the United Kingdom may also have been a result of the magnitude of rewards, which can be up to 30 percent of practice income. Rewards of this size may have "crowded out" practices' intrinsic motivation, hence 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. [75] 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, i.e., measures pertaining to conditions with a high prevalence and/or disease burden. However, especially with respect to clinical quality, lack of data often hampers inclusion of important performance 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 information technology 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 that are tied to financial incentives [71]. Yet, in P4P it is particularly important to carefully monitor the more indeterminate aspects such as continuity of care and patient centredness (both core features of good patient care) because these aspects will be among the first aspects that may be neglected when the extrinsic motivation of providers is emphasized [67]. However, adequate measurement of these aspects is often more difficult and more expensive than measurement of e.g., clinical processes or resource use. Consequently, even monitoring may be not feasible. It is important, therefore, that providers are actively involved in measure selection and program design.

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 [48]. 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 [25]. 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 [18]. 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 [32]. The introduction of P4P could then play a trivializing role regarding the nonfinancial motivation [6, 13]. 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 [67], 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 [31]. Outcrowding is more likely to occur in creative tasks, in overly bureaucratic schemes, and in the more indeterminate aspects of professional practice [67]. To prevent outcrowding, purchasers should make sure that the incentives are viewed as legitimating and reinforcing of internal motivators [15, 33]. If the incentives are aligned with providers' internal value framework, the likelihood that the program will be successful increases [67]. 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 [102]. 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).

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.

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 [37, 86, 97]. However, health care is increasingly provided in settings in which professionals from diverging medical disciplines cooperate in the treatment of patients. Consequently, it is becoming increasingly difficult to ascribe a "good performance" to an individual practitioner. Therefore, often it would be logical to target P4P at groups of physicians rather than individual physicians. (In this paper, we follow Town et al.'s [97] definition of a medical group, i.e., 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, rather than 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 [97]. 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 [29, 58] and because of enabling factors like assistance of other professional and support staff [102], collaboration, peer review, and available infrastructure. However, it is important to assess whether and how incentives are passed along to group members [35]. 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 [3, 36]. Free riding on the efforts of peers may then be difficult to detect and penalize. As noted by Town and colleagues [97], 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 [37]. Next to diluted incentives, from a purchaser perspective, a potential disadvantage of directing P4P at groups is that groups generally have more bargaining power than individuals and are more effective in defying or negotiating the terms of external incentive programs [77, 97]. Based on interviews with sponsors of hospital P4P programs in the United States, Damberg et al. [17] 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 [64]. 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 [63]. Patient panels of individual physicians and small groups are typically too small to measure performance reliably [2, 54, 56, 63, 76, 90]. Thus, if P4P targets individual physicians or small groups, measured performance is likely to reflect to a significant degree random variation [13, 73], possibly resulting in misclassification of providers and incorrect allocation of incentive payments [2, 76]. Constructing composite scores could increase low reliability due to small sample size per measure [8] 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 [80]. 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 [90]. Aggregating data across purchasers may also be an option [50]. However, for several reasons (e.g., possible violations of anti-trust regulation, technical difficulties, patient confidentiality), this does not occur on a large and systematic scale yet.

On balance, group incentives seem preferred over individual incentives, mainly because performance profiles are more likely to be reliable [56]. However, when performance is compared across groups, it is important that there are sufficient numbers of physicians in each comparison group to detect meaningful differences. Nonadjustment 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 [44]. In addition, groups differ considerably in size and

composition, and it is unclear how to treat the many providers working in small practices with small numbers of patients for many measures [65, 73]. Although health care is increasingly provided in group settings, small practice settings will likely remain important, necessitating strategies to facilitate inclusion of small practices [65]. As methods for data aggregation and constructing composite scores continue to evolve [50], 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 [97], perhaps because they are less transparent and therefore less visible to providers.

How to incentivize: how is the program structured?

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 [61]. 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 in case performance targets are met will elicit a stronger behavioral response than offering providers a $1000 bonus for good performance [17]. However, research has shown that incentive schemes incorporating losses tend to be perceived as unfair and may result in negative reactions among those incentivized [60]. 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., a health plan) is relatively low and if providers can choose from among multiple plans to contract with [1]. But even if providers can be convinced or enforced to participate, the behavioral response to financial penalties may not necessarily be a desired response. The prospect of a loss may cause physicians to behave opportunistically, and incentives for gaming and other undesired behavior may be large. (Importantly, 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. Yet, negative reactions will be stronger in case of absolute financial penalties).

A possible way to still take advantage of the expected strong provider response while limiting the possibility 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 [71]. In case the provider chooses the second option and fails to reach the target, it 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 1 displays the characteristics of four possible schemes.

Despite the advantages of using rewards, purchasers may opt for using "old" money (e.g., redistributing money to high performers based on generically reduced base payments). They could argue that programs using rewards may not be sustainable and object to investing additional resources in settings with substantial inefficiencies [13]. 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

Table 1 Characteristics 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 than in 2	Depends on choice	Moderately high	Moderately low

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, in case 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 programs has been declining rapidly. In the United States, although withholds are still applied in ten to twenty 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 [4, 17, 93]. Also in other countries, P4P programs typically only provide positive incentives.

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 [16, 47, 101]. These costs will vary by the base payment system and the set of performance measures, so the payment level sufficient to realize improvements is not a static figure [12]. 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- or target- income hypothesis is applicable to physicians [81, 82], 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 with small payments they are more likely to impair providers' intrinsic motivation [18, 33]. Consequently, the likelihood of undesired behavior increases because positive net gains of this behavior are more likely. Monitoring for this behavior may be costly and difficult, so in determining incentive size purchasers will often be confronted with a trade-off between an increased (but at a certain 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 [17]. This is illustrated by Gneezy and Rustichini [40], 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 [55, 79]. Therefore, relatively low-powered payments seem to be preferred, provided that they are based on performance measures that are aligned with providers' professional norms and values.

Empirical research on the influence of incentive size is scarce. Hillman et al. [52, 53] 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 2 years) and lack of physician awareness. Conversely, Mullen et al. [75] 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 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 increased the size of potential bonuses for medical groups compared to what they could potentially earn under QIP. The authors found that while the QIP alone had not been able to generate improvements in quality, after the other plans also adopted P4P some quality measures did improve. Thus, the authors concluded that payment size matters [75]. 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 [22]. However, it is unclear to what extent observed improvements can be attributed to these generous payments. In addition, as shown by McDonald and Roland [68], 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 to P4P incentives (i.e., the gains forgone of doing the next best alternative) are determined largely by the base payment system [35]. Especially in fee-for-service, these costs can be substantial because time and effort put in improving performance cannot be used to treat patients and to perform tests. 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 so that mixed payment is appointed [27]. 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 [97]. Performance-related payments will therefore be supplemental to base payments. In addition, it seems warranted to "decouple" incentive payments from base payment as much as possible [71]. 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 incentive payment may be perceived as negligible compared to the base payment and the behavioral response may be small [95]. However, decoupling adds to administrative complexity [71].

Absolute versus relative performance

Incentive payments can be based on absolute performance (e.g., performing a foot examination 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 [86]. The goal gradient hypothesis predicts that a goal should be perceived attainable by providers; otherwise, little response can be expected [49]. 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 [86]. 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 [16]. 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 can be calculated ex ante [86], which gives providers the prospect of certain payment in case 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. This is exactly what happened in the QOF in the United Kingdom. By 2006–2007 (the third year), primary care practices scored on average more than 95 percent of the points available, which exceeded the predictions of the Department of Health, which had anticipated 75 percent attainment [22]. 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 [16]. If there is not much flexibility in increasing the pool of incentive payments, the pool may 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 solved 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. In that case, 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 penalty. Another option is to eliminate targets altogether and to use a continuous gradient [71]. 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 [21]. 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 [86]. 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., its simplicity and certainty for providers, as well as less incentives for risk selection compared to explicit targets). A recent study by Chien et al. [11] 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. Also, the program did not exacerbate disparities nor have a negative effect on children with chronic conditions.

In sum, differentiating required absolute performance levels 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 in a single program should be weighted against increased complexity and reduced incentive salience.

Frequency of payments

Providing a monthly $100 bonus with an additional payment of $500 based on overall improvement will 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 [30]. In addition, people generally discount losses at lower rates than gains and large outcomes more than small outcomes [30, 96]. 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 often incurred without much delay. A high frequency becomes even more important in case providers experience uncertainty regarding the net gains of improvement efforts (as in relative schemes) 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 lump-sum 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 [17, 95]. Finally, a high payment frequency increases incentive salience. In practice, however, 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. [14] investigated whether the impact of P4P is larger when payments are provided quarterly as opposed to 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 quarterly P4P and quarterly 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 more generic 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 be used only for short-term objectives. For example, in long-term contracts with hospitals, payment could be linked to 5-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.

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 performance improvement (e.g., adopting an expensive IT infrastructure) requires making projections about future payment rates and expectations

about return on investment [97]. Thus, the duration of the program as well as providers' expectations hereof 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 [16, 18]. 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 incentives emanating from the base payments. Therefore, once implemented, performance-related payment should ideally remain a permanent component of providers' compensation. However, it is questionable 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, i.e., the duration of incentivizing specific aspects of performance within the program, is also of relevance [102]. A high frequency can be demoralizing for providers, especially if 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 in 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 [63, 90].

Discussion

This paper provides an overview of key issues in the design of P4P programs by synthesizing theoretical and empirical literature. The design of P4P programs is important since it determines the way in which the behavior of providers is influenced. To prevent undesired behavior, careful consideration of how the incentives are framed is vital, especially in multitasking environments [55]. Although the idea underlying P4P is simple, this paper has shown that designing a fair and effective P4P program is a complex undertaking requiring consideration of many interrelated aspects and potential pitfalls. Nonetheless, several tentative conclusions can be made, which are summarized in Table 2.

However, conclusions on appropriate program 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. Next to the base payment system, other relevant contextual factors are the characteristics of the practice environment (e.g., the level of information technology); whether P4P is implemented in a single-purchaser healthcare system or in a system with

Table 2 Conclusions with respect to P4P-program design

What to incentivize
Performance is ideally defined broadly, provided that the set of measures remains comprehensible
Concerns that P4P encourages "risk selection" and "teaching to the test" should not be dismissed
Outcome and resource use measures should be included provided that risk adjustment is sophisticated and sample size is sufficient. Other strategies to minimize incentives for risk selection may still be necessary
Measure sets should at least incorporate "high-impact" measures; the more indeterminate aspects of care such as patient satisfaction and continuity of care are ideally also included or monitored
P4P incentives should be aligned with professional norms and values; it is vital that providers are actively involved in program design and in the selection of performance measures
Monitoring, structured feedback, and sophisticated information technology will remain important in preventing undesired provider behavior
Whom to incentivize
On balance, group incentives are preferred over individual incentives, mainly because performance profiles are then more likely to be reliable
Individual or small-group incentives as well as using measures with small sample size will become increasingly feasible as methods for constructing composite scores evolve
Caution should be upheld in applying hybrid schemes
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 considered
Increasing the size of the incentive increases their strength up to a certain point. Yet, relatively low-powered payments are preferred, provided that providers' costs of improving performance are covered
Differentiated absolute targets across groups and/or a tiered series of absolute targets, possibly combined with additional "piece-rates" for each appropriately managed patient, are preferred over single targets and schemes using relative targets
The time lag between care delivery and payment should be minimized
P4P should be a permanent component of compensation and is ideally decoupled from base payments. Measures should be reevaluated periodically and be replaced or updated as necessary

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 United States, 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 have more possibilities to directly influence providers' behavior and align providers' incentives than purchasers that operate more or less independently from providers); whether providers have fixed patient panels (if not, computerized algorithms are necessary to attribute care to involved providers 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 be in violation with anti-trust regulation). Recently, research has begun to address the influence of specific contextual factors (e.g., [69, 98]. 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 including 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 [71]. 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 [13, 35, 72, 78, 87]. 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. [78] and Frølich et al. [35], 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 predominantly 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 [17].

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 [28]. Nonetheless, P4P may be able to mitigate cost growth through better prevention and disease management and through inclusion of efficiency measures. Recently, purchasers have begun to incorporate efficiency measures in their P4P programs [57, 83]. 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 improvement strategies in their efforts to achieve more value for money.

Acknowledgments I would like to thank Wynand van de Ven, Erik Schut, and two anonymous referees for their helpful comments on earlier drafts of this paper.

References

1. Arrow, K.J.: Agency and the market. In: Arrow, K.J., Intriligator, M.D. (eds.) Handbook of Mathematical Economics, vol. 3, pp. 1183–1195. Elsevier, Amsterdam (1986)
2. Adams, J.L., Mehrotra, A., Thomas, J.W., McGlynn, E.A.: Physician cost profiling—reliability and risk of misclassification. N. Engl. J. Med. **362**(11), 1014–1021 (2010)
3. Alchian, A.A., Demsetz, H.: Production, information costs, and economic organization. Am. Econ. Rev. **62**(5), 777–795 (1972)
4. Baker, G., Delbanco, S.: Pay for Performance: National Perspective. 2006 Longitudinal Survey Results with 2007 Market Updates. Med-Vantage, San Francisco (2007)
5. Benavent, J., Juan, C., Clos, J., Sequeira, E., Gimferrer, N., Vilaseca, J.: Using pay-for-performance to introduce changes in primary healthcare centers in Spain: first year results. Qual. Prim. Care **17**(2), 123–131 (2009)
6. Berwick, D.M.: The toxicity of pay for performance. Qual. Manag. Healthc. **4**(1), 27 (1995)
7. Buetow, S.: Pay-for-performance in New Zealand primary health care. J. Health Organ. Manag. **22**(1), 36–47 (2008)
8. Caldis, T.: Composite health plan quality scales. Health Care Financ. Rev. **28**(3), 95–107 (2007)
9. Campbell, S.M., Reeves, D., Kontopantelis, E., Sibbald, B., Roland, M.: Effects of pay for performance on the quality of primary care in England. N. Engl. J. Med. **361**(4), 368–378 (2009)
10. Chen, T.T., Chung, K.P., Lin, I.C., Lai, M.S.: The unintended consequence of diabetes mellitus pay-for-performance (P4P) program in Taiwan: are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? Health Serv. Res. (28 Sept 2010)
11. Chien, A.T., Li, Z., Rosenthal, M.B.: Improving timely childhood immunizations through pay for performance in medicaid-managed care. Health Serv. Res. **45**(6, part II), 1934–1947 (2010)
12. Christianson, J., Leatherman, S., Sutherland, K.: Financial Incentives, Healthcare Providers and Quality Improvements: A Review of the Evidence. The Health Foundation, London (2007)
13. Christianson, J.B., Leatherman, S., Sutherland, K.: Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence. Med. Care Res. Rev. **65**(6 Suppl), 5S–35S (2008)
14. Chung, S., Palaniappan, L., Wong, E., Rubin, H., Luft, H.: Does the frequency of pay-for-performance payment matter? Experience from a randomized trial. Health Serv. Res. **45**(2), 553–564 (2009)
15. Conrad, D.A., Christianson, J.B.: Penetrating the" black box": financial incentives for enhancing the quality of physician services. Med. Care Res. Rev. **61**(3), 37–68 (2004)
16. Conrad, D.A., Perry, L.: Quality-based financial incentives in health care: can we improve quality by paying for it? Annu. Rev. Public Health **30**, 357–371 (2009)
17. Damberg, C.L., Sorbero, M.E., Mehrotra, A., Teleki, S.S., Lovejoy, S., Bradley, L.: An environmental scan of pay for performance in the hospital setting: final report. Rand Health working paper WR-474-ASPE/CMS, Rand Health, Santa Monica, CA (2007)
18. Deci, E.L., Koestner, R., Ryan, R.M.: A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol. Bull. **125**, 627–668 (1999)
19. Donabedian, A.: The quality of care: how can it be assessed? J. Am. Med. Assoc. **260**(12), 1743–1748 (1988)
20. Doran, T., Fullwood, C., Reeves, D., Gravelle, H., Roland, M.: Exclusion of patients from pay-for-performance targets by English Physicians. N. Engl. J. Med. **359**(3), 274–284 (2008)
21. Doran, T., Fullwood, C., Kontopantelis, E., Reeves, D.: Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. Lancet **372**(9640), 728–736 (2008)
22. Doran, T., Roland, M.: Lessons from major initiatives to improve primary care in the United Kingdom. Health Aff. **29**(5), 1023–1029 (2010)
23. Dranove, D., Kessler, D., McClellan, M., Satterthwaite, M.: Is more information better? The effects of "Report Cards" on health care providers. J. Political Econ. **111**(3), 555–588 (2003)
24. Duckett, S., Daniels, S., Kamp, M., Stockwell, A., Walker, G., Ward, M.: Pay for performance in Australia: Queensland's new clinical practice improvement payment. J. Health Serv. Res. Policy **13**(3), 174–177 (2008)
25. Dudley, R.A., Frolich, A., Robinowitz, D.L., Talavera, J.A., Broadhead, P., Luft, H.S., McDonald, K.: Strategies to support quality-based purchasing: a review of the evidence. Technical review 10. 04–0057. Agency for Healthcare Research and Quality, Rockville (2004)
26. Dudley, R.A., Miller, R.H., Korenbrot, T.Y., Luft, H.S.: The impact of financial incentives on quality of health care. Milbank Q. **76**(4), 649–686 (1998)
27. Eggleston, K.: Multitasking and mixed systems for provider payment. J. Health Econ. **24**(1), 211–223 (2005)
28. Emmert, M., Eijkenaar, F., Kempter, H., Esslinger, S., Schöffski, O.: Economic evaluation of pay for performance in health care: a systematic review. Forthcoming in the Eur. J. Health Econ. (2011)
29. Enthoven, A.C., Tollen, L.A.: Competition in health care: It takes systems to pursue quality and efficiency. Health Aff. Web Exclusives **W5**, 420–433 (2005)
30. Frederick, S., Loewenstein, G., O'Donoghue, T.: Time discounting and time preference: a critical review. J. Econ. Lit. **40**(2), 351–401 (2002)
31. Freedman, J.L., Cunningham, J.A., Krismer, K.: Inferred values and the reverse-incentive effect in induced compliance. J. Pers. Soc. Psychol. **62**(3), 357–368 (1992)
32. Freidson, E.: Professionalism: the third logic. Polity Press, London (2001)
33. Frey, B.S.: On the relationship between intrinsic and extrinsic work motivation. Int. J. Ind. Organ. **15**(4), 427–439 (1997)
34. Friedberg, M.W., Safran, D.G., Coltin, K., Dresser, M., Schneider, E.C.: Paying for performance in primary care: potential impact on practices and disparities. Health Aff. **29**(5), 926–932 (2010)
35. Frølich, A., Talavera, J.A., Broadhead, P., Dudley, R.A.: A behavioral model of clinician responses to incentives to improve quality. Health Policy **80**(1), 179–193 (2007)
36. Gaynor, M., Gertler, P.: Moral hazard and risk spreading in partnerships. Rand J. Econ. **26**(4), 591–613 (1995)
37. Gaynor, M., Rebitzer, J.B., Taylor, L.J.: Physician incentives in health maintenance organizations. J. Political Econ. **112**(4), 915–931 (2004)
38. Gibbons, R.: Incentives in organizations. J. Econ. Perspect. **12**, 115–132 (1998)
39. Glickman, S.W., Ou, F., DeLong, E.R., Roe, M.T., Lytle, B.L., Mulgund, J., Rumsfeld, J.S., Gibler, W.B., Ohman, E.M., Schulman, K.A., Peterson, E.D.: Pay for performance, quality of care, and outcomes in acute myocardial infarction. J. Am. Med. Assoc. **297**(21), 2373–2380 (2007)
40. Gneezy, U., Rustichini, A.: Pay enough or don't pay at all. Quart. J. Econ. **115**(3), 791–810 (2000)
41. Gosden, T., Forland, F., Kristiansen, I. S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., Pedersen, L. :Capitation, salary, fee-for-service and mixed systems of payment: effects on the

behaviour of primary care physicians. Cochrane Database Syst. Rev. **3**(3) (2000)
42. Gosden, T., Forland, F., Kristiansen, I. S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., Pedersen, L.: Impact of payment method on behaviour of primary care physicians: a systematic review. J. Health Serv. Res. Policy **6**(1), 44–55 (2001)
43. Gravelle, H., Sutton, M., Ma, A.: Doctor behaviour under a pay for performance contract: further evidence from the quality and outcomes framework. CHE research paper 34. Center for Health Economics, York (2008)
44. Greenfield, S., Kaplan, S.H., Kahn, R., Ninomiya, J., Griffith, J.L.: Profiling care provided by different groups of physicians: effects of patient case-mix (bias) and physician-level clustering on quality assessment results. Ann. Intern. Med. **136**(2), 111–121 (2002)
45. Grol, R.: Successes and failures in the implementation of evidence-based guidelines for clinical practice. Med. Care **39**(8 Suppl 2), II46–II54 (2001)
46. Gross, R., Elhaynay, A., Friedman, N., Buetow, S.: Pay-for-performance programs in Israeli sick funds. J. Health Organ. Manag. **22**(1), 23–35 (2008)
47. Grossman, S.J., Hart, O.D.: An analysis of the principal-agent problem. Econometrica **51**, 7–45 (1983)
48. Grumbach, K., Osmond, D., Vranizan, K., Jaffe, D., Bindman, A.B.: Primary care physicians' experience of financial incentives in managed-care systems. N. Engl. J. Med. **339**(21), 1516–1521 (1998)
49. Heath, C., Larrick, R.P., Wu, G.: Goals as reference points. Cogn. Psychol. **38**(1), 79–109 (1999)
50. Higgins, A., Zeddies, T., Pearson, S.D.: Measuring the performance of individual physicians by collecting data from multiple health plans: the results of a two-state test. Health Aff. **30**(4), 673–681 (2011)
51. Hillman, A. L., Pauly, M. V., Kerstein, J. J.: How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations? N. Engl. J. Med. **321**(2), 86–92 (1989)
52. Hillman, A.L., Ripley, K., Goldfarb, N., Nuamah, I., Weiner, J., Lusk, E.: Physician financial incentives and feedback: failure to increase cancer screening in medicaid managed care. Am. J. Public Health **88**(11), 1699 (1998)
53. Hillman, A.L., Ripley, K., Goldfarb, N., Weiner, J., Nuamah, I., Lusk, E.: The use of physician financial incentives and feedback to improve pediatric preventive care in medicaid managed care. Pediatrics **104**(4), 931–935 (1999)
54. Hofer, T.P., Hayward, R.A., Greenfield, S., Wagner, E.H., Kaplan, S.H., Manning, W.G.: The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. J. Am. Med. Assoc. **281**(22), 2098–2105 (1999)
55. Holmstrom, B., Milgrom, P.: Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. J. Law Econ. Organ. **7**(1), 24–52 (1991)
56. Huang, I.C., Diette, G.B., Dominici, F., Frangakis, C., Wu, A.W.: Variations of physician group profiling indicators for asthma care. Am. J. Manag. Care **11**(1), 38–44 (2005)
57. Institute of Medicine: Rewarding Provider Performance: Aligning Incentives in Medicare. The National Academies Press, Washington, DC (2007)
58. Institute of Medicine: Crossing the Quality Chasm: A New Health System for the 21st Century. National University Press, Washington, DC (2001)
59. Ittner, C.D., Larcker, D.F.: Determinants of performance measure choices in worker incentive plans. J. Labor Econ. **20**(2, Pt. 2), S58–S90 (2002)
60. Kahneman, D., Knetsch, J.L., Thaler, R.: Fairness as a constraint on profit seeking: entitlements in the market. Am. Econ. Rev. **76**(4), 728–741 (1986)
61. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**(2), 263–291 (1979)
62. Karve, A.M., Ou, F.S., Lytle, B.L., Peterson, E.D.: Potential unintended financial consequences of pay-for-performance on the quality of care for minority patients. Am. Heart J. **155**(3), 571–576 (2008)
63. Krein, S.L., Hofer, T.P., Kerr, E.A., Hayward, R.A.: Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. Health Serv. Res. **37**(5), 1159–1180 (2002)
64. Landon, B.E., Normand, S.L., Blumenthal, D., Daley, J.: Physician clinical performance assessment: prospects and barriers. J. Am. Med. Assoc. **290**(9), 1183–1189 (2003)
65. Landon, B.E., Normand, S.T.: Performance Measurement in the small office practice: challenges and potential solutions. Ann. Intern. Med. **148**, 353–357 (2008)
66. Lee, T.T., Cheng, S.H., Chen, C.C., Lai, M.S.: A pay-for-performance program for diabetes care in Taiwan: a preliminary assessment. Am. J. Manag. Care **16**(1), 65–69 (2010)
67. Marshall, M., Harrison, S.: It's about more than money: financial incentives and internal motivation. Qual. Saf. Health Care **14**(1), 4–5 (2005)
68. McDonald, R., Roland, M.: Pay for performance in primary care in England and California: comparison of unintended consequences. Ann. Fam. Med. **7**(2), 121–127 (2009)
69. McDonald, R., White, J., Marmor, T.R.: Paying for performance in primary medical care: learning about and learning from "success" and "failure" in England and California. J. Health Politics Policy Law **34**(5), 747–776 (2009)
70. McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., Kerr, E.A.: The quality of health care delivered to adults in the United States. N. Engl. J. Med. **348**(26), 2635–2645 (2003)
71. Mehrotra, A., Adams, J.L., Thomas, J.W., McGlynn, E.A.: Cost profiles: should the focus be on individual physicians or physician groups? Health Aff. **29**(8), 1532–1538 (2010)
72. Mehrotra, A., Damberg, C.L., Sorbero, M.E., Teleki, S.S.: Pay for performance in the hospital setting: What is the state of the evidence? Am. J. Med. Qual. **24**(1), 19–28 (2009)
73. Mehrotra, A., Sorbero, M.E., Damberg, C.L.: Using the lessons of behavioral economics to design more effective pay-for-performance programs. Am. J. Manag. Care **16**(7), 497–503 (2010)
74. Moore, S.H., Martin, D.P., Richardson, W.C., Riedel, D.C.: Cost containment through risk-sharing by primary care physicians: A history of the development of united healthcare. Health Care Financ. Rev. **1**(4), 1–13 (1980)
75. Mullen, K.J., Frank, R.G., Rosenthal, M.B.: Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. RAND J. Econ. **41**(1), 64–91 (2010)
76. Nyweide, D.J., Weeks, W.B., Gottlieb, D.J., Casalino, L.P., Fisher, E.S.: Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. J. Am. Med. Assoc. **302**(22), 2444–2450 (2009)
77. Oliver, P.: Rewards and punishments as selective incentives for collective action: theoretical investigations. Am. J. Sociol. **85**(6), 1356–1375 (1980)
78. Petersen, L.A., Woodard, L.D., Urech, T., Daw, C., Sookanan, S.: Does pay-for-performance improve the quality of health care? Ann. Intern. Med. **145**(4), 265–272 (2006)
79. Prendergast, C.: The provision of incentives in firms. J. Econ. Lit. **37**(1), 7–63 (1999)
80. Reeves, D., Campbell, S.M., Adams, J., Shekelle, P.G., Kontopantelis, E., Roland, M.O.: Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. Med. Care **45**(6), 489–496 (2007)

81. Rizzo, J.A., Blumenthal, J.A.: Is the target income hypothesis an economic heresy? Med. Care Res. Rev. **53**(3), 243–266 (1996)
82. Rizzo, J.A., Zeckhauser, R.J.: Reference incomes, loss aversion, and physician behavior. Rev. Econ. Stat. **85**(4), 909–922 (2003)
83. Robinson, J.C., Williams, T., Yanagihara, D.: Measurement of and reward for efficiency in California's pay-for-performance program. Health Aff. **28**(5), 1438–1447 (2009)
84. Rochon, M., Pink, G.H., Brown, A.D., 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.: *HealthcarePapers* **6**(4) (2006)
85. Roland, M.: Linking physicians' pay to the quality of care–a major experiment in the United Kingdom. N. Engl. J. Med. **351**(14), 1448–1454 (2004)
86. Rosenthal, M.B., Dudley, R.A.: Pay-for-performance: will the latest payment trend improve care? J. Am. Med. Assoc. **297**(7), 740–744 (2007)
87. Rosenthal, M.B., Frank, R.G.: What is the empirical basis for paying for quality in health care? Med. Care Res. Rev. **63**(2), 135–157 (2006)
88. Rosenthal, M.B., Landon, B.E., Howitt, K., Song, H.R., Epstein, A.M.: Climbing up the pay-for-performance learning curve: where are the early adopters now? Health Aff. **26**(6), 1674–1682 (2007)
89. Rosenthal, M.B., Landon, B.E., Normand, S.L., Frank, R.G., Epstein, A.M.: Pay for performance in commercial HMOs. N. Engl. J. Med. **355**(18), 1895–1902 (2006)
90. Scholle, S.H., Roski, J., Adams, J.L., Dunn, D.L., Kerr, E.A., Dugan, D.P., Jensen, R.E.: Benchmarking physician performance: Reliability of individual and composite measures. Am. J. Manag. Care **14**(12), 833–838 (2008)
91. Seddon, M.E., Marshall, M.N., Campbell, S.M., Roland, M.O.: Systematic review of studies of quality of clinical care in general practice in the UK, Australia and New Zealand. Qual. Health Care **10**(3), 152–158 (2001)
92. Shen, Y.: Selection incentives in a performance-based contracting system. Health Serv. Res. **38**(2), 535–552 (2003)
93. Sorbero, M.E., Damberg, C.L., Shaw, R., Teleki, S., Lovejoy, S., Decristofaro, A., Dembosky, J., Schuster, C.: Assessment of pay-for-performance options for medicare physician services: final report. RAND health, RAND working paper WR-391-ASPE, Santa Monica, CA (2006)
94. Steel, N., Maisey, S., Clark, A., Fleetcroft, R., Howe, A.: Quality of clinical primary care and targeted incentive payments: an observational study. Br. J. Gen. Pract. **57**, 449–454 (2007)
95. Thaler, R.H.: Mental accounting and consumer choice. Market. Sci. **4**(3), 199–214 (1985)
96. Thaler, R.H.: Some empirical evidence on dynamic inconsistency. Econ. Lett. **8**(3), 201–207 (1981)
97. Town, R., Wholey, D.R., Kralewski, J., Dowd, B.: Assessing the influence of incentives on physicians and medical groups. Med. Care Res. Rev. **61**(3 Suppl), 80S–118S (2004)
98. Van Herck, P., De Smedt, D., Annemans, L., Remmen, R., Rosenthal, M.B., Sermeus, W.: Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Serv. Res. **10**(1), 247 (2010)
99. Werner, R.M., Goldman, L.E., Dudley, R.A.: Comparison of change in quality of care between safety-net and non-safety-net hospitals. J. Am. Med. Assoc. **299**(18), 2180–2187 (2008)
100. Weyer, S.M., Bobiak, S., Stange, K.C.: Possible unintended consequences of a focus on performance: insights over time from the research association of practices network. Qual. Manag. Health Care **17**(1), 47–52 (2008)
101. Young, G.J., Conrad, D.A.: Practical issues in the design and implementation of pay-for-quality programs. J. Healthc. Manag. **52**(1), 10–18 (2007)
102. Young, G.J., White, B., Burgess Jr, J.F., Berlowitz, D., Meterko, M., Guldin, M.R., Bokhour, B.G.: Conceptual issues in the design and implementation of pay-for-quality programs. Am. J. Med. Qual. **20**(3), 144–150 (2005)