Paying for Performance in Primary Medical Care: Learning about and Learning from "Success" and "Failure" in England and California

Ruth McDonald
University of Manchester

Joseph White
Case Western Reserve University

Theodore R. Marmor
Yale University

Abstract Paying physicians to hit performance targets is becoming increasingly fashionable, as evidenced by the growing number of "pay-for-performance" programs in the United States and beyond. This article compares pay-for-performance initiatives in two nations—the United Kingdom and the United States. It pays particular attention to the context in which the initiatives were conceived and implemented, factors which are largely neglected in the pay-for-performance literature. Despite some glowing reviews of the UK national pay-for-performance program for primary care doctors, we suggest that such programs face significant technical obstacles in all cases and particularly severe institutional obstacles in the United States.

Introduction

Policy fads affect every industry and medical care is no exception (Marmor 2004). One recent medical example is the enthusiasm for reforming payment methods to "pay for performance," referred to conventionally as P4P. Ample evidence of its fashionable status may be found in the large and growing number of P4P programs in the United States and beyond (Baker and Carter 2005; Roland 2004). Yet despite the popularity of P4P, evidence on the effect of P4P programs is mixed (Christianson, Leatherman, and Sutherland 2008). P4P initiatives have been shown to exacerbate (Werner, Goldman, and Dudley 2008) or to reduce disparities (Baker and

Funding was provided by the Commonwealth Fund, the NHS Service Delivery and Organisation Research and Development Programme, and the UK Department of Health. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the funders.

Journal of Health Politics, Policy and Law, Vol. 34, No. 5, October 2009
DOI 10.1215/03616878-2009-024

Middleton 2003; Doran et al. 2008a; Middleton and Baker 2003) in the United States and United Kingdom, respectively. Recent reviews of the literature report that methodological flaws in the small number of studies undertaken make drawing conclusions difficult (Christianson, Leatherman, and Sutherland 2008; Mehrotra et al. 2009). The most rigorous of these studies from the hospital sector suggests that a P4P initiative contributed to a small but statistically significant improvement in performance on clinical process measures (e.g., pneumococcal vaccine delivery).

Many P4P initiatives focus on primary medical care. After all, primary care is where much of the preventive and chronic-disease care is provided and where outcome and output metrics are relatively well developed (Terry 2005). In the primary care sector in the United Kingdom, for example, quality, as measured on a range of indicators relating to asthma, diabetes, and heart disease, was already improving rapidly before the introduction of a national P4P program (Campbell et al. 2005). Following the introduction of P4P, there were small but significant increases in improvements for asthma and diabetes, but no change in the rate of improvement for heart disease. This suggests that factors other than P4P (including other quality improvement initiatives; see Christianson, Leatherman, and Sutherland 2008) may account for changes in performance.

Felt-Lisk, Gimm, and Peterson (2007) found that the most successful plans (in terms of quality improvement) paid the highest rewards. Other studies show that payment for reaching benchmarks has resulted in most of the dollars being paid out to providers that had achieved the benchmarks prior to the P4P program's implementation (Lindenauer et al. 2007; Rosenthal et al. 2005). Payment for reaching benchmarks also brings with it the risk that when performance exceeds predetermined benchmarks by a substantial amount, the budget allocated for payments will be inadequate, as happened in the first year of the UK P4P program (Galvin 2006).

Given the popularity of P4P initiatives, there is ample reason to try to learn from the experiences of early implementers. Yet as one recent review of the literature stated, "the variation in the way in which these programs have been designed and implemented, [makes] synthesizing their findings to provide useful guidance to decision makers . . . challenging" (Christianson, Leatherman, and Sutherland 2008: 8) Furthermore, in attempting to learn from experience, one needs to separate faddish enthusiasm from factual accuracy and, as well, to treat bold cross-national conclusions with care. By care we mean especially attention to differences in context — and purpose — that attend seemingly common enterprises like paying for performance.

Cross-National Studies: Promise, Performance, and Lessons

In the United Kingdom in 2004 a large national initiative, linking remuneration to "quality" targets was introduced for primary care physicians (Roland 2004). This program, the Quality and Outcomes Framework, or QOF, led some observers to wax lyrical about the scheme. "The boldest such proposal on this scale ever attempted anywhere in the world" was the assessment made by one U.S. observer in the *British Medical Journal* (Shekelle 2003: 457). The editor of the prestigious *New England Journal of Medicine* suggested that "the British are clearly ahead of us in the adoption of financial incentives for improving the quality of care. We would do well to learn from their experience" (Epstein 2006: 408). These assessments proceeded from the assumption that the problems facing the U.S. health system powerfully resemble those of other Organisation for Economic Co-operation and Development (OECD) countries, with these including rising costs of care, the rapid pace of technological advance, changing and unmet health care needs, and deficiencies in the quality of care provided (Institute of Medicine 2001).

From that assumption, the inference is that countries with common problems will be willing and able to learn from the experiences of other health systems. However, when attempting to learn from experiences of other democracies, there is a danger that lessons will be drawn too quickly and without due regard for the contextual factors contributing to the "success," or the lack of success, of initiatives in other countries (Marmor, Freeman, and Okma 2005). Indeed, as we shall try to demonstrate in this article, the results of pay-for-performance initiatives vary substantially according to the context in which they are implemented.

This article compares pay-for-performance initiatives in two nations—the United Kingdom and the United States. In the UK program, performance against quality targets was better than anticipated. In the U.S. context, by contrast, third party payers have expressed disappointment. "Breakthrough" improvements in health care services, the declared aim of the incentive program, have not been achieved (Atoji 2008). Our aim is to identify those key factors in the NHS context that contributed to the "success" of the UK initiative. And, in parallel, we outline those contextual factors likely to reduce the impact of financial incentives for quality in U.S. primary health care settings.

Our approach was to compare the UK program with a statewide pay-for-performance initiative in California (Damberg et al. 2005). Our analysis

draws mainly on interviews with frontline clinicians (forty in total, twenty from each case). These in-depth interviews about experience and perceptions allow us to identify factors that are omitted by large-scale quantitative evaluations (e.g., Doran et al. 2006). The respondents were identified using "snowballing" (a small number of informants put the researcher in touch with others, who then nominated colleagues and other contacts, and so on). In the U.S. sample, respondents were more likely to be participating in medical group level activities in addition to providing patient care (e.g., membership of the group quality committee, group board level participation), so these physicians were likely to be more positively inclined toward financial incentives than is the case for the general population of primary care physicians in California. The interviews were all undertaken by the same researcher, digitally recorded, and transcribed verbatim. Transcripts were read and reread to identify emerging issues and interpretations, enabling the identification of key concepts and themes. Codes were created on the basis of these themes and linked to the data collected using a software package, Atlas.ti. In addition, the article reflects discussions and meetings with other key informants involved in the development and implementation of P4P in primary medical care in both countries.

The article has three further sections. The first briefly describes the background of the two financial incentive programs studied. The second compares these programs, with special attention to explaining their relative impact. The final section discusses these findings and what should, or should not, be learned from them.

Background: QOF and IHA Pay for Performance

When the UK National Health Service (NHS) was formed in 1948, primary care physicians (general practitioners, or GPs) refused to join the NHS as salaried employees and remained independent contractors (Loudon, Horder, and Webster 1998). This assured their independence, guaranteed their income from the state, and protected them by restricting the provision of primary medical care for NHS patients to members of the profession. The national GP contract is the result of a negotiated process involving representatives of both the GP profession and the state. However, in 1990 a new contract, which contained a handful of performance targets, was imposed on the profession despite its opposition. This opposition has been interpreted as a rejection of the "contract state" and related

market reforms promulgated by the Conservative Thatcher government (Lewis 1997).

The election of a Labour government in 1997, with a manifesto commitment to "save and modernise the NHS" signaled the start of a new era in health policy. In 2000 the NHS Plan (UK Department of Health 2000) outlined various policy objectives aimed at modernizing primary care. These included increased emphasis on performance-based rewards. The context was one of relatively poor morale in general practice, long hours and low pay relative to hospital doctors, and difficulties in recruiting GPs. Moreover, the medical profession was understood to accept the view that the existing general medical services (GMS) contract was outdated and inflexible. This view was widely shared by government officials as well (National Audit Office 2008).

That background sets the stage for the new GMS contract that came into effect in April 2004. The most prominent element of this, the Quality and Outcomes Framework (QOF), comprised 146 largely evidence-based process indicators of quality of care, mostly for chronic diseases. For example, diabetes indicators include "points" for offering smoking cessation advice to diabetic patients, ensuring that diabetes is well controlled, and maintaining total cholesterol within target levels in diabetic patients. Compliance with each indicator attracted a specific number of points, up to a maximum total of 1,050 for any practice. Points are converted to pounds, with a large percentage of practice income dependent on achieving these targets. That will turn out to be an important comparative difference from the California experience.

The QOF includes the concept of reporting "exception." This allows practices to exclude patients from performance calculations where, for example, patients refuse to attend for review, despite three invitations being sent or where a medication cannot be prescribed due to a contraindication or side effect. The contract reforms also offered GPs the ability to opt out of the responsibility for providing care out of hours (OOH) and resulted in significant increases in income for GP partnerships (Batty 2003). These factors help explain why, unlike in 1990, most GPs voted in favor of the new GMS contract in a national ballot.

In 2001, the year after the publication of the NHS Plan, the Institute of Medicine published a prominent report, "Crossing the Quality Chasm," that called for major change at all levels of American medical care, emphasizing particularly system redesign to improve the quality of care provided (Institute of Medicine 2001). The report highlighted the goal of better alignment of payment methods with quality goals; it lauded increased

Table 1 Pay for Performance Compared, IHA and QOF, 2007–2008

	California—IHA	England—QOF
Physicians	40,000	33,000
Patients (million)	12	50
Exception report	No	Yes
Targets	25 (approx)	135
% income	Varies	30

Notes: The patients listed for the California physicians represent only the portion of their patients covered by the IHA. IHA = Integrated Healthcare Association; QOF = Quality and Outcomes Framework

transparency and accountability and called for expanded consumer choice. (All of these are themes in the NHS Plan as well.) In 2002, the Robert Wood Johnson Foundation and the California Health Care Foundation funded seven demonstration projects to implement and evaluate incentives (financial and otherwise) for improved quality of care. The largest of these studies was Integrated Healthcare Association's (IHA) pay-for-performance program. It initially covered 6.5 million Californians, subsequently expanding to cover 12 million patients at the time of writing (Integrated Healthcare Association 2007). Unlike the UK single-payer system, the P4P program involves seven participating health plans, each with its own rules about payment. Table 1 above highlights key features of the two schemes.

Both programs include targets related to processes such as diagnosis and/or referral (e.g., cervical cancer screening), monitoring (e.g., of cholesterol on patients with diabetes), ensuring appropriate treatment (e.g., appropriate medication for asthma patients), and patient experience measures. In addition, targets also include intermediate outcomes (e.g., good control of cholesterol) that are proxies for desired health outcomes (such as avoidance of mortality or coronary events).

Factors in the Performance of "Pay for Performance"

P4P programs almost always proceed from the belief that financial incentives will motivate providers to achieve improvements in the quality of care delivered. They take for granted that behavior follows financial incentives. These financial incentives, defined as a structure of fiscal rewards or punishments, presume a fairly simple — simplistic might suffice — view of human behavior, motivation, and action.

Drawing on economic theory and "available data" to examine dimensions of P4P programs that are important determinants of their influence, Rosenthal and Dudley (2007) identify "5 aspects of program design that are likely to be the most consequential." These are "individual versus group incentives," "paying the right amount," "selecting high-impact performance measures," "making payment reward all high-quality care," and "prioritizing quality improvement for underserved populations" (741). There is a danger, however, that a reliance on economic theory and quantitative data will ignore complexity in human behavior, organizational structures, and the broader context of medical care delivery. Our study, which encountered some of this complexity, permits a different view of the factors influencing the impact of P4P initiatives.

The Extent to Which the Message Is Clear and Targeted Properly

One precondition for an incentive to have an impact is that the objects of persuasion must be aware of the measures for which they will be remunerated. Awareness as a precondition may seem an obvious point, yet we found major differences in the extent to which physicians in California and England were aware of the areas of care delivery to which financial incentives were directed.

English physicians, for instance, were very aware of the targets against which their performance was measured. Although the number of English targets is much greater, there is one nationally agreed set and a single payer. Furthermore, performance measurement is based on data extracted from the electronic medical records which are owned and maintained by the practice (a group of typically one to six physicians and their staff). Practice team members in Britain are very conscious of targets and can monitor income and performance levels on an ongoing basis. That in turn means they can try to take remedial action on a timely basis, as is suggested by one report that, "our practice manager is very good. . . . She does send us round updates . . . where we're deficient, where we should be doing better . . . about ticking the yellow [target indicator] boxes" (ID16UK). Another noted that "there's a computer programme [so] you can check, 'how am I doing?' So you can tell . . . how much you need to do by a certain date" (ID7UK).

Although there is one statewide P4P initiative in California, there are other initiatives operating within the state that incorporate financial incentives to improve quality and/or to constrain resource use. As a result,

the combination of multiple payers and complex payment rules generates a situation in which many physicians reported being unaware of and/or confused by targets and related payments. "The numbers," one responded, "the surveys, the data they take to decide what those bonuses should be, I don't quite understand it" (ID7USA). "I'm too busy to even know what this is all about," (ID20USA) said another. "It's a little bit of a black box," one explained, "because you don't know from year to year what's going to be actually measured . . . I can't say that it incentivizes you much. It just seems like it falls out of the sky" (ID9USA).

The Extent to Which Those Targets of Incentive Are Able to Respond in the Desired Way

This aspect concerns the extent to which the incentive is for something that is easily identified and acted upon. In both settings, many P4P indicators are process rather than outcome measures. In England, there are eighty "clinical domain" targets covering nineteen clinical areas (e.g., coronary heart disease, heart failure, hypertension). In addition, practices score points for organizational domain targets; these regard topics like patient satisfaction, medication management, and education and training. In contrast, in California, targets cover a smaller number of disease areas, with only thirteen clinical targets in total. For example, in IHA P4P, there are two targets relating to the measurement and control respectively of HbA1c (blood sugar) levels, compared with sixteen QOF targets for diabetes. Other targets cover patient experience and information technology (IT).

Despite the much higher number of targets in England, physicians reported greater ability to respond to incentives in the desired way. A key factor here appeared to be the availability of information technology systems designed for this purpose. In the United Kingdom by 1996, 96 percent of general practices were computerized (Jordan, Porcheret, and Croft 2004). The requirement to meet contract targets spurred practices to improve their IT systems. This facilitated data collection for payment purposes (for a detailed description, see Checkland, McDonald, and Harrison 2007). In addition to electronic disease registries that enable practices to call and recall patients for screening and treatment, general practice software provides a running total of points achieved. This permits identifying by name those patients for whom targets remain unmet (e.g., John Doe, blood cholesterol remains greater than the target) and for whom a service is overdue. GPs also described using computer templates, or standard-

ized data entry tools, which facilitate the collection of patient data and act as an aide memoire to clinicians. Computer pop-up boxes helpfully prompt clinicians, highlighting any areas where action and data entry are required.

In contrast to the United Kingdom, most practices in California have no electronic medical record (EMR). In a recent survey only 28 percent of U.S. primary care physicians reported their medical records as either fully or partially electronic (Hing, Burt, and Woodwell 2007). Thus one respondent commented that "ideally you would be able to push a button and, say, show me all my patients with congestive heart failure. Show me all my patients with systolic blood pressures over 130 . . . but nobody has that" (ID17USA). The absence of EMRs, computerized disease registers, templates, and prompts in the California practice setting reduces the ability of those physicians to meet targets.

English physicians were not, we should emphasize, unreserved enthusiasts for EMRs. A number described the requirement to enter data into the electronic medical record as distracting them from patient care. Responding to the large number of pop-up boxes linked to targets reduced eye contact and increased time spent on data collection in the consultation. Physicians also expressed fears that this would crowd out the patient's agenda, with undue focus on QOF targets. Only one English practice in our study had increased its ten-minute consultation slot to fifteen minutes, with all others attempting to collect the additional data within the pre-QOF time. "There's a system in place, you get the yellow boxes coming up," one respondent explained, "and I think that's a barrier . . . to the doctor/patient relationship. You may be seeing somebody for depression, and you really don't want to be talking about checking their weight and height and . . . when did they stop smoking" (ID17UK). As another summarized, "we have the risk of not seeing the patient as an actual patient . . . because you are so wrapped up in, 'Well, we've got to do this, we've got to do this" (ID3UK).

The Extent to Which the Desired Result Is Subject to Significant Influence by the Recipient of the Incentive

While P4P programs encompass many process measures, some process measures are much easier for physicians to influence than others. In both the U.S. and the UK settings, the content of targets is controlled by considerations of measurement. After all, if action cannot be measured, how

can performance be assessed? But in the U.S. setting, other factors greatly influenced the content of targets. In particular, the absence of an electronic health record in most of the California practices meant that the selection of targets was limited by data availability in addition to the conceptual obstacles to measurement. Often the available data were for tasks performed outside of the physicians' office, such as lab tests for diabetes, cholesterol levels, and chlamydia, or other tests such as mammography and colorectal screening. This meant the omission of tasks like hypertension measurement and management (the objects of incentives in the UK QOF).

Since hypertension is central to the work of family practice, the California program suffered accordingly. This was so despite the fact that hypertension is much more prevalent than chlamydia and highly cost-effective to treat, with treatment for hypertension likely to produce substantial health gains. The result of setting targets based on readily available data was a California program with incentives for aspects of care less amenable to physician control than in the English setting. Furthermore, the English QOF included the idea of exception reporting. This allows practices to exclude patients from performance calculations where, for example, patients refuse to attend for review, despite three invitations being sent, or where a medication cannot be prescribed due to a contraindication or side effect. In contrast, exception reporting is not allowed in the California program, which frustrated physicians who understood their limited control of the measures used. Hence, one objected to the idea that "it's a physician's responsibility for patients to access health care and do their bit for making them achieve the metrics . . . the physician is held accountable . . . I think that's flawed" (ID11USA). That leads to, in another's words, "a subtle pressure to get rid of noncompliant patients . . . And a lot of times, the noncompliant patients are the ones that need the most care" (ID17USA). Yet physicians did not necessarily blame patients; as one explained, a lot of people fail to get blood tests "because they honestly don't have time. They have to work. And the labs aren't open on weekends. So . . . when they take time off to get the blood tests done . . . that's money out of their pocket" (ID18USA).

The Extent to Which the Organizational Setting Provides the Capacity to Respond to Incentives

As outlined above, an important factor facilitating the delivery of care in line with targets is a system's IT capacity. The United Kingdom's single-

pipe financing of health care, with a large percentage of practice income guaranteed in the form of capitation payments, provides financial stability to practices. Furthermore, a combination of reimbursement incentives and subsidies reflecting a longer-term interest in and commitment to the stability of primary care providers has helped computerize general practice in the United Kingdom (Onion and Berrington 1999). As part of this process, there has been financial support for nurses and other staff working within the practice, which encourages the development of primary care teams. Compared with the physician offices in California, there are many more nurses working in the English practices. In England, these nurses handle a third of all office visits. Moreover, much of the day-to-day management of chronic disease (to which the majority of clinical indicators relate) is undertaken by these nurses (National Audit Office 2008).

English physicians therefore described working closely with nurses. In one physician's words, "I think that if you've got a good nurse . . . I've now got somebody that if I want some advice on a diabetic patient I go and talk to [names nurse] because I know she knows more about it than I do, and that's fine" (ID5UK). In some cases, they rely on nurses to take the lead in addressing targets, as with the physician who reported that "in this practice we, for whatever reason, weren't really managing diabetes very well. And the fact that we're not reaching some of those targets has been I think a bit of an impetus to change things and start getting nurses doing more chronic disease management, which I think is good and needed to happen" (ID10UK).

In contrast, in the United States many physicians have been and to a large extent remain in solo or small-practice settings (Casalino 2004), which makes it difficult to develop the infrastructure required to achieve performance targets. Health plans have moved away from contracting directly with individual physicians and small practices, with 61 percent now targeting medical groups alone (Rosenthal et al. 2006). In the United States in recent years, as a response to this trend and as a counterbalance to the might of large health plans, physicians have been increasingly organizing themselves into larger structures (Grumbach et al. 1998; Gillies et al. 2003). Most of the physicians interviewed were affiliated with one or more Independent Practice Associations (IPAs). IPAs negotiate with insurance companies; credential and inspect member physicians, institutions, and services; and disburse payment to physicians. In the California P4P initiative, payments are made to IPAs and other large groups as opposed to individual physicians.

In this structure, rather than viewing IPAs as supportive, many physi-

cians reported feeling under siege. They complained that both the health plans and their IPAs badgered them regarding their performance against targets, and some described being corralled into IPA arrangements as a result of health plans changing their contracting processes. The result was physicians feeling resentful at their perceived loss of independence rather than pleased about entering a collaborative venture, and the P4P programs were perceived as part of this new regulation. One respondent, for example, viewed the P4P incentives as "monies that were withheld in the negotiation . . . So when you talk to docs out in the field they see it as a withhold and a lot of hurdles to be met to obtain their rightful, just, and fair compensation" (ID11USA).

IPAs' business strategies introduced a further level of complexity. Even though IPAs received income based on performance against IHA P4P targets, many set local targets that differed from IHA targets. Some groups topped up the IHA P4P money to create larger sums for performance but did this by withholding money from fee-for-service payments to physicians. The percentage of income from incentives was generally much lower than in the UK initiative, but for physicians struggling to hit targets, placing more of their income at risk is not viewed positively. In contrast, the P4P funds in the United Kingdom—in the 2004 contract reforms—were viewed as and actually were new money.

Hence, California IPAs differ from an English group practice on all relevant dimensions. They do not have the organizational coherence or exclusive relationship to providers that exists within either the UK NHS or within integrated delivery systems such as Kaiser Permanente in California. Nor do the IPAs resemble English group practices as face-to-face social groups. Any influence on meeting targets that IPAs can exert is likely to be much less systematic and more limited than in these other models. (Kaiser Permanente, although not remunerated from the IHA P4P system, reports its activity against the P4P targets. It consistently outperforms the vast majority of organizations in California on these metrics despite receiving no incentive payments for doing so.)

IPAs do vary with regard to the additional organizational capacity made available to affiliated providers. In the largest IPA in our study, in which incentive payments were comparable (in percentage terms) with the English context, much greater effort was expended on providing detailed feedback to physicians, drawing attention to deficient performance and highlighting remedial action to be taken. As a result, physicians in this IPA were more likely to be aware of the targets. Such "support" from IPAs was welcomed in some cases; yet the structure of administration

and context for the incentives still could lead to negative attitudes. Greater feedback did not make up for the fact that measures came from outside the practice. "We get what are called 'drill downs,' " one physician reported. "They give us a list of the outliers. Because blank-blank-blank never got her mammogram, that's why your numbers were knocked down, and the sad thing about that blank-blank-blank [is] you may have never seen them, but they are assigned to you. You have no relationship to them, but . . . it's your responsibility" (ID14USA). Since the reports are not "owned" by the practice, "it goes in the drawer. You look at it and say, 'Ah, I'm average — that's OK.' Or, 'I used to be average, I'm a little above average — well, I don't know, what can I do about it?' " This respondent laughed and added, "we get feedback from the medical group . . . we get a list of people who haven't had their mammograms — we call them up; we send a form. They still don't get their mammogram" (ID5USA).

The Extent to Which Measures Are Perceived as Accurate

In the English context, performance measurement for each medical practice is based on data extracted from the practice EMR. This has the advantage that the practice owns the data. As part of the pay-for-performance reforms in the United Kingdom, responsibility for fulfilling contractual obligations has moved from the individual physician to the practice (a group of, typically, one to six physicians, together with nurses, administrative support, and other staff). These features, combined with the ability to exclude patients who refuse treatment or for whom treatment is contraindicated, mean that English physicians had neither the grounds nor the inclination to dispute the accuracy of data.

The California physicians, by contrast, did complain strongly about the accuracy of the data on which their performance was judged, as well as the burden of correcting errors. "I have 91 diabetics," one explained, of whom 32 were reported as "missing either a hemoglobin A1C or an LDL or [to] have elevated levels from September to August '07." But, when he went through the labs and charts, "just on the first two pages I found that six of them were incorrect" (ID9USA). Not only were there errors in the figures an IPA might have, but because a given IPA could fund only a portion of a doctor's patients, evaluations were often based on small numbers and thus were statistically unreliable. "One angry person with only 30 numbers," one physician noted, "will screw up your whole thing. One person giving you all zeros will be a very big thing and they don't

take out the high and low. . . . So it's frustrating because I don't think that's representative" (ID19USA). Another complained about "health planners . . . saying, 'OK, you only had a 65 percent mammogram success rate because two or three of these folks say, 'No, thank you.' " As a result, he commented, "pay for performance, does it work? I'm not convinced it does" (ID14USA). Not surprisingly, these data problems, some of which are built into the fragmented U.S. financing system, make the P4P initiative less legitimate to California physicians.

The Extent to Which Measures Are Perceived as Fair and Legitimate

In both contexts, developing clinical indicators involved a degree of participation by primary care professionals and their representatives. In England, all primary care physicians are formally represented in the national contract negotiations and the development of targets. They also are allowed to vote prior to the introduction of new incentives. In California, proposed indicators are published on the Internet and public comment is invited. Following this process, indicators are tried out—piloted—before being incorporated into the incentive program. Indicators are reviewed and revised annually. But there is no formal voting, and physicians in California notably complained about the content of the indicators against which their performance was measured.

Although some Californian physicians were supportive of P4P, most expressed markedly less satisfaction with it than their English counterparts. In part because it was regarded as both externally imposed and managed, it was seen as challenging physician autonomy and suggested to some that they were not trusted to perform well in the absence of incentive payments. This was viewed as ignoring professional and personal ethical motivations, such as those described by a physician who noted, "I want to do a good job. I get compared to, but not in a formal way, to my colleagues. . . . I certainly see all the time whether that colleague agrees with my referral diagnosis and continues treatment as appropriate or says, 'Doctor [X] was way off base and we think you've got something else.' That's a very strong motivation for quality of practice" (ID7USA). So, as another physician described it, it can be "kind of irritating" (ID15USA) to have one's judgment challenged and be monitored against measures that aren't perceived as useful: "One of the thing[s] that [names IPA] monitors as a clinical measure is micro albumins for diabetics. A lot of [the] time, you feel like you're ordering the micro albumin every year, and it really doesn't have a huge clinical

significance" (ibid.). The fact that physicians felt they were held accountable for results beyond their control added to their frustration, and that was compounded by their inability to exclude patients who refused treatment or for whom the targets were inappropriate.

The IHA P4P measures were especially suspect if viewed as connected to incentives to provide fewer services. The IHA was introducing, from 2009, new "efficiency targets." During the interview period, such measures from the IPAs were associated in physicians' minds with the P4P regulation. Thus, one physician mentioned in the interview that "the IPA gave a bonus at the end of year based on how much you spent on x-rays and labs and that sort of thing . . . The feeling is that if we're going to get money to order fewer tests, that's a conflict of interest . . . Even people who are well-intentioned are going to think twice about ordering the more expensive tests, because then they're going to receive less money in the end" (ID19USA). In short, how the P4P initiatives are viewed depends on the overall context of payment controls and other regulation of physicians.

The Extent to Which Measures Are Perceived as Threatening

Those receiving what they perceive to be unfair or illegitimate incentives might at times choose to ignore them. However where noncompliance represents a serious threat to a physician's livelihood and/or reputation, tensions are likely to be much stronger. In the English setting, the P4P reforms were part of a broader program of change and investment, all intended to secure "more, better paid staff using new ways of working" (UK Department of Health 2000: 10). Investment in primary care was part of an overall strategy of modernizing the NHS, investing substantial new funds to bring the UK's health expenditures more in line with other European countries. To achieve this, a large increase in expenditure on primary care was planned, from £4.9 billion in 2002–2003 to £6.9 billion in 2005–2006. The amount of new money invested exceeded the plan by £406 million, mainly due to greater than anticipated expenditures on QOF and the provision of OOH services. The result was very substantial increases in income for medical practices and improved recruitment and retention in these settings (National Audit Office 2008).

That does not mean all is well in UK primary care. But our comparison does establish much greater acceptance of quality measures in the British case, as implemented. In the UK incentives initiative, QOF scores

for every practice are publicly available on the Internet. Among English physicians interviewed, there was much greater acceptance of the target regime, compared with Californian physicians. This echoes findings from other studies, which report support for QOF targets from English GPs (McDonald et al. 2007; Campbell, McDonald, and Lester 2008). The fact that GPs voted for the 2004 contract, with its 146 targets (reduced to 135 in 2006) and rejected a 1990 variant that contained only a handful of indicators suggests that the culture of UK general medical practice has changed much in the last twenty years. The growth of evidence-based medicine, an acceptance that it is possible to define and measure aspects of high-quality care, and an increasing emphasis on accountability in professional life and the public sector in general are all factors that have contributed to an increasing acceptance of the legitimacy of performance measurement (Roland 2004; May 2007).

Other UK circumstances also altered the context in which quality issues — like those addressed by P4P — were viewed. High-profile scandals, in particular the case of Dr. Harold Shipman, a GP from England who murdered large numbers of patients, helped strengthen the case for medical accountability (Horton 2001). In the mid 1980s, UK physicians had rejected a proposal to pay them a "good practice allowance" on the grounds that quality could not be measured and there was no such thing as a bad practice (Roland 2004). The Shipman case — and the growth of evidence-based indicators — made such arguments much more difficult to sustain. Thus, a combination of additional resources and a greater acceptance of the legitimacy of performance measurement and public accountability offers a plausible account for the fact that English physicians in our study did not perceive the P4P measures as threatening.

In sharp contrast, in California, there was no such large investment in primary care. Only limited resources were available to fund the P4P initiative. Furthermore, in the United States the gap in incomes between primary care physicians and specialists is widening, not narrowing. Without radical payment reform, given a system favoring specialists (Bodenheimer, Berenson, and Rudolf 2007), it is difficult to foresee an improvement in primary care incomes relative to specialists. The condition of English GPs compared to hospital consultants, in this decade, was quite different.

Most English GPs are independent contractors, and many see themselves as more independent than salaried hospital specialists. (For a discussion of the negative reaction of hospital doctors to their new contract, see McDonald, Waring, and Harrison 2006). The new primary care contract sharply increased GP incomes, bringing them from well below to on

a par with those of hospital doctors. "Effectively," one said, "we had five years' pay rise in one go" (ID20UK). Therefore, many British GPs viewed the program as a welcome recognition of their professional worth, posing the issue as, "do we want high-quality, well-paid . . . motivated, interested, vocationally oriented GPs, or do we want a bunch of doctors who became GPs because they couldn't get the well paid jobs in eye surgery? . . . The majority of my patients have come in and said, 'Well, you do earn a lot of money,' but they have not said, 'You're not worth it,' and that's the acid test" (ID14UK). In contrast, primary care doctors in California compared their relatively low incomes with those of specialist colleagues. The fact that many specialists were not subject to P4P targets only added insult to injury. "Dermatology or other kinds of specialists," one complained, "can work short hours and make a lot of money . . . Specialty care has more or less taken over things" (ID5USA). Linked to this, many were fearful of public reporting and felt unfairly singled out for scrutiny compared with other professionals. "The physicians are resenting that kind of thing [public reporting of performance data] obviously already," as one put it, adding, "It's like the government getting into your private affairs at your bank and you not knowing about it, or this Patriot Act. . . . I think physicians are monitored more than anybody. . . . Maybe they'll be pushed to a certain point where they will rise up and say, 'No more.' I don't know" (ID20USA). Even those who supported public reporting in principle reported concerns about accuracy and relevance of data.

In summary, for many of the California physicians in our study, professional life was a struggle to maintain one's professional and commercial viability. They saw themselves in a context where performance targets worsened already increasing financial pressures and problems in recruiting staff. In this context, P4P was perceived as offering "extra pressures" with highly dubious benefits for actual health. As one articulate physician put the case, "I haven't heard that there's ever been a study done that has shown that paid performance decreases mortality or morbidity. . . . I haven't seen proof that it actually makes a difference. . . . For me, it's just 'leave me alone and let me do my work,' instead of throwing all of these road blocks at us" (ID19USA).

Our interview responses, therefore, suggest the following. The fate of quality measures has more to do with the context in which they are introduced than with technical features of the incentives. Whether initiatives are threatening or welcomed, then, is not about policy ideas abstracted from the setting in which they are introduced. Rather, the opposite holds. In California, the lack of new money added to the frustration of physicians,

764 Journal of Health Politics, Policy and Law

but the context of their relationship with IPAs and insurers had already bred frustration and distrust. The payers, in turn, were disappointed that "breakthrough" improvements in health care services, the declared aim of the IHA incentive program, were not achieved (Atoji 2008). In the United Kingdom, both extra funds and a more supportive relationship caused substantial change in practice. In fact, high levels of reaching targets in the UK program resulted in higher than expected levels of expenditure. As a result, UK primary care doctors enjoyed substantial increases in their incomes (Doran et al. 2006). In California, overspending on P4P was impossible because the funds were capped.

The UK result has prompted considerable commentary and some criticism. The absence of baseline (i.e., pre-QOF) data and the apparent ease with which practices were able to hit targets raised questions about the extent to which incentive payments produced better "quality" or merely improved recording. Media headlines followed, with some describing general practitioners as overpaid (Timmins 2005). Furthermore, the productivity increase projected in the Department of Health's business case for the new contract has not materialized. General consultation numbers have increased but at a much slower rate than costs. And the number of consultations by GPs has fallen, with increased transfers of routine cases to nurses. The result is that UK GPs are working fewer hours but earning more (National Audit Office 2008).

That, in turn, has prompted much criticism and considerable defensiveness on the part of the government. The Secretary of State for Health suggested, "If we [had] anticipated this business of GPs taking a higher share of income in profits we would have wanted to do something to try to ensure that the ratio of profits to the total income stayed the same." He went on to say that he would have wished to invest "more money . . . in even better services for patients" (Triggle 2007). The government's situation was not helped by the admission of a member of the GP negotiating team on national radio that the British Medical Association (BMA) was stunned to be offered such a generous package (Martin 2007). Although the remarks mainly concerned the offer to absolve GPs from responsibility for out of hours care, this admission added further evidence to the suggestion that GPs were overpaid and underworked and that government negotiators were deficient. Echoing the secretary of state's sentiments, the leader of the pay negotiations for NHS employers on behalf of the government alluded to the possibility of a pay cap for GPs. As part of the negotiation for 2006–2007, the Department of Health allocation contained only a small increase compared with 2005–2006.

While QOF targets and the related pay rise appeared to motivate GPs, it also appeared from GP interviews that the government's response was starting to undermine some of the positive aspects associated with the new contract. One commented, "I think what the Secretary of State said was complete nonsense. We have invested in our practices, we did a lot of that before the new contract came in and we've now reaped the rewards of that and I think it's downright dishonest of politicians to play it any other way. . . . They've not believed what's actually going on and [got] a nasty shock when they found they had to pay for it" (ID11UK). Another was concerned about "the uncertainty about how next year will look and whether we will get a zero percent pay rise next year, which in real terms here is a cut. . . . We've put things in place to help deliver [P4P performance] but now it's being undermined because in real terms the money to finance this is drifting backwards" (ID18UK). In December 2007 negotiations for revisions to the GMS contract broke down and notice was given that a new contract was to be imposed on practices if no agreement could be reached.

The issue on which the negotiations foundered concerned extended opening hours for GPs' surgeries. While increasing opening hours might be seen as part of the government's commitment to create a responsive NHS, in a survey of 2.2 million people in July 2007, 84 percent reported being satisfied with the current opening hours of their practice, suggesting that "consumers" were not clamoring for greater access. Faced with two nonnegotiated options, both of which involved extending opening hours, in March over 90 percent of GPs who voted selected the option they believed was the lesser evil. A poll conducted alongside the ballot found that English GPs have little faith in English government policy and its ability to improve the health service. The poll indicated that 97 percent of respondents reported no confidence in the government's handling of the NHS; 98 percent said they regard the government's method of negotiation as unacceptable (British Medical Association 2008). In summary, GPs have accepted a nonnegotiated contract that exerts greater control over them and leaves them feeling frustrated and demoralized. We may wonder if the special circumstances that supported the P4P initiative have changed considerably. Nevertheless, any differences between England now and England in 2005 are likely to be less significant than the differences between England and California, for the latter involve the basic structure of general practice.

Discussion

Our comparison of pay for performance in California and England provides information about both programs that can supplement large-scale quantitative reports on those programs. The comparison itself, however, adds further by highlighting aspects likely to be common among pay-for-performance reforms and what might be the effects of context and design.

Assessing the "success" of P4P programs involves some potentially controversial choices. The most obvious and internally consistent standard is the extent to which the targets are met. Analysts must ensure, however, that positive results are not due to "gaming." Beyond this difficulty, critics may argue that the real goal is quality (or cost-effectiveness) and that targets might have been met by diminishing performance on unmeasured dimensions. A second criticism is that measures often target processes (such as frequency with which blood pressure is monitored) as opposed to outcomes (e.g., levels of hypertension). A third issue involves quality for whom. While one perspective would emphasize population averages, much current health policy concern focuses on reducing disparities—so, essentially lowering the deviation more than raising the norm. A fourth issue involves results on other values. For example, when implemented, the U.S. Medicare program was a great success on its most obvious goal, improving access to care for its elderly beneficiaries. But it was swiftly criticized in the political arena because of its high costs (Marmor 2000). Similarly, we have seen that the P4P initiative in the United Kingdom has come under attack for its costs in spite of its success in terms of meeting targets.

This study's analysis focuses on the extent to which targets are met and practice changed because whether P4P initiatives meet their stated goals is logically primary: if they do not succeed on that dimension, any other successes would be pure accident. Further effects do influence the desirability of such initiatives, but they are harder to measure. A few comments, however, are possible.

If we judge the "success" of P4P programs in terms of performance against targets, then the performance in the United Kingdom appears to be superior. High rates of achievement in the UK context raise questions about gaming, since "success" may reflect an abuse of a system which allows physicians to exclude patients from target calculations. Evidence suggests that exclusion was more likely for indicators that concerned treatment (e.g., influenza vaccination) and/or intermediate outcomes (e.g.,

epilepsy—convulsion-free for twelve months). However, rates of exclusions have been generally low. In addition, characteristics of physician practices, such as whether they were in an area of socioeconomic deprivation, had only marginal effects, and high rates of achievement do not appear to have resulted from practices in deprived areas excluding many more patients (Doran et al. 2008b). However, while practices are unlikely to remove patients from their list due to targets, there is some evidence that some practices have under-recorded disease prevalence, reducing the number of patients for which a target has to be met (Gravelle, Sutton, and Ma 2008). In addition, since QOF was introduced without any prior assessment of baseline (pre-P4P) quality, judging how much extra quality has been forthcoming is difficult. Still, the totality of evidence suggests better performance at meeting targets in the United Kingdom than in California.

While it might be expected that achievement on intermediate outcomes in both countries would generally be lower than on process measures (e.g., measuring blood pressure in patients with hypertension in the United Kingdom versus reducing the blood pressure of hypertensive patients to target levels), the pattern is inconsistent in the California data. Achievement levels for some screening processes are lower than for some outcome (e.g., cholesterol measures). However, low rates of achievement on screening relate particularly to chlamydia, which is not included in the UK P4P program.

We do not have data on effects on disparities in California in particular. In the United Kingdom, variation in quality of care (as measured by targets) related to deprivation was reduced over the first three years of the program, and since all citizens can register with a practice there is little possibility that patients in areas of disadvantage will be excluded from registration. This suggests that QOF has had positive effects on reducing disparities. In part, this may be because it resulted in a relatively large injection of funds into all practices, since the payments were not made from a fixed pot that merely rewarded highest-achieving practices.

As noted, effects on cost are a major policy concern in the United Kingdom, while the contrasting concern, an increase in work without more payment, is one source of the relatively weaker target performance in California. The absence of good baseline data in the United Kingdom explains, in part at least, why expenditure exceeded projections—in the absence of data, negotiators were forced to "guesstimate" the likely level of performance, resulting in a substantial underestimate.

A simple comparison of quantitative performance data and financial

rewards in each setting might lead one to conclude that higher than expected levels of performance (on the measures) in England are due to higher levels of reward. Following this approach, a simple solution to the failure to achieve "breakthrough" improvements in measured performance in California (Atoji 2008) would be to increase the percentage of income linked to targets in the IHA P4P program. While Felt-Lisk, Gimm, and Peterson (2007) found that the most successful plans (in terms of quality improvement) paid the highest rewards in the program they studied, our work illustrates that much more complex factors must be considered.

The most evident complication is at the level of goals. Our study raises questions about not only learning from "success" or "failure," but also their definitions in our study settings. For example, had English GPs scored only 75 percent on average, which was the estimated and budgeted likely level of performance, instead of the 90-plus percent, then budgets would have been met, government would have been less defensive, the media less critical, and so on. Hence, lower rates of achievement could have received a warmer welcome, despite their (presumed) reduced impact on quality improvement. The government was not committed to paying for *too much* performance.

The government was also not exactly sure, or at least its reasoning was not compelling, as to what the system was supposed to achieve. The direct effects were clear enough. But the fact that those were not the ultimate goals is revealed by the government's subsequent doubts. In its business case to Her Majesty's Treasury, the Department of Health cites various expected benefits of its pay modernization program (National Audit Office 2008). These include the redesign of services around patients and, as part of this, the allocation of resources to local populations according to need. Yet in order to obtain support from physicians for the new incentive program, a "Minimum Practice Income Guarantee" was established, which protected practices from earning less core pay than they did under the old funding system. As a result, redistribution of resources to underserved areas has been limited.

One interpretation of such ambiguity is that the English P4P initiative was less a "solution" developed in response to a well-defined "problem" (Pawson and Tilley 1997) than a policy that had advocates and then was attached, politically, to various convenient "problems" when a policy window (the Blair government's desire to spend more money and the time to negotiate a new GP contract) opened (Kingdon 1994; March and Olsen 1976). Under such circumstances, policies will be adopted not because evidence shows they are well fitted to the problem, but because there is an

opportunity to do *something* and the chosen policy survives the assorted vetoes within the political system. An example of such a reaction in the United States might be the Institute of Medicine's report *Rewarding Provider Performance: Aligning Incentives in Medicare*, which noted that "most studies have failed to demonstrate any significant effects on processes of care," but recommended the introduction of financial incentives for quality nonetheless (Institute of Medicine 2007: 46). Under such circumstances, the standards for evaluation are not self-evident, because not all the reasons given should be taken seriously.

In a hypothetical world, the goal of the English P4P initiative would have been obvious: to increase value for the medical care dollar. Success or failure would also have been evident, because value could be measured. In fact, the subsequent controversy over spending levels suggests that, at a minimum, advocates of the initiative could not argue that it clearly achieved such a goal. Instead, it was justified on a series of conflicting or ambiguous bases, which at one point allowed for political agreement and later led to conflict.

It should be no surprise that pay-for-performance programs embody multiple and conflicting goals. Most evidently, the parties to a contract might have different goals. The conflict was present from the beginning in California. At the outset of implementation in England, such a conflict was suppressed because the government wanted to spend more, a goal with which the general practitioners naturally agreed. However, recent action by the English authorities to increase GP workload suggests that increased productivity — always a potential goal for payers — is now being prioritized. A reasonable interpretation would be that conflicts of interest between parties in a negotiation may be suppressed in good times (as in England earlier) but become more explicit under tighter circumstances (England as it is changing and California throughout the IHA program). The English case shows that physicians do not have to be so skeptical of pay for performance as U.S. experience might imply but also that the conditions for reduced skepticism might be rare (and expensive).

Yet goal conflicts occur at an operating and specification level as well. Consider the issue of how many targets should be set. There are many more targets in England than in California, and that has both causes and consequences. With as few targets as in California, pay for performance can only be a small part of payment for physicians. With as many targets as in England, it can be a major part of the system — and so attract far more attention. With as few targets as in California, some concerns are implicitly prioritized over others. And with as many as in England, it is

very difficult to avoid the sense of "tick-box" medicine, because there are so many boxes to tick (and nurses have expressed concerns similar to those of doctors; see McDonald et al. 2007; Campbell et al. 2008). A possible response in England would be to reduce the number of targets to more manageable proportions. That, however, would raise difficult choices about the conditions to be singled out (or, conversely, excluded from) "quality improvement." English policy makers do not appear eager to explicitly prioritize, for example, clinical outputs over patient satisfaction or diabetes over heart disease. That could make the measurement regime less legitimate, and California appears to illustrate the potential problem.

Goal conflict and ambiguity are endemic to policy processes. A less obvious aspect of our pay-for-performance comparison may be more significant. Pay-for-performance proposals are based on theories of incentives that take the objects of incentives, in this case physicians, out of their social and organizational contexts. Yet context is crucial. The problem is not simply that physicians and widget manufacturers are not the same—that is, that they are not motivated in identical ways, patients (unlike widgets) are not uniform, and so on. The problem is that physicians and physicians are not the same: professional autonomy, discretion, and judgment are likely to play differing roles in different contexts.

One issue frequently underemphasized in theorizing about incentives is whether the targets of incentives have the capacity to respond in the desired manner. Can they control what they are supposed to control? This is an issue in many policy areas—in the United States, perhaps the most politically salient is ratings of school performance according to the performance of students on test scores. At the most basic level, California physicians objected that they were scored on results, such as whether patients received certain tests that were not under the physicians' control. But many such problems were much more severe in California because of the differences between how medical care is organized in California and England. Capacity is largely a matter of systems, not individuals, and therefore the ways in which pay-for-performance measures work depend on organizational variables.

Organizational factors influence the extent to which the message is clear and targeted properly, the extent to which the person targeted is able to respond in the desired way, and the ability of the recipient of influence to manipulate the desired result. In our comparison, the most obvious difference in organization is how health care is purchased. Having the NHS "single payer" covering all patients makes messages about desired

behavior much more clear (since primary care physicians only face one set of measures), makes it easier for patients to comply (because they are covered for required lab services); and even makes it easier to have exception reporting (because the payer sees the entire set of exceptions and can spread costs of monitoring over the full extent of a general practice).

The comparison between California and England reveals a further and equally fundamental difference: how practice is organized. English practices are smaller and more coherent than the California IPAs, but more extensive organizations than the California offices in which many physicians actually work. The organizational effects of finance and delivery are mutually supportive: English GP practices have better information systems partly because they are organized to a scale that makes that more economic, but also because the single payer could substantially standardize the systems (giving practices a guarantee that they would be investing in the right thing) and could also subsidize their purchase (because the single payer would get all the benefit from the better information systems, unlike any American payer). The organization and payment for general practice also encourage staffing which includes the practice nurses who do so much of the work of meeting the performance standards.

This study's interviews reveal a further aspect of organization, more a matter of the informal than the formal structure. There was a striking difference in the extent to which physicians showed trust in the design and implementation of the pay-for-performance initiatives in California and England. Nobody should imagine that the NHS is a "high-trust" situation, as the quick erosion of physician belief in the good intentions of NHS authorities shows. Yet we also should not assume that the observed difference was merely due to the fact that the performance payments were extra money in England and operated virtually as part of a withhold system in California. That surely mattered, but at an operating level distrust permeated the California initiative. We could see the difference in the lack of a provision for exceptions; in comments made in interviews, about bypassing informed consent for chlamydia screening or threatening to remove patients from a doctor's practice; and in comments about the unreliability of the information.

Physician distrust of the system in California was produced, in part, by the system's distrust of physicians. "Ownership" and "buy-in" will be greater in a high-trust environment. In the English setting, the willingness to use data collected by practices has resulted in primary-care teams engaging in self-surveillance and proactive management to achieve targets. In California, the IHA P4P program permits data submission by

medical groups rather than relying on health plan data alone. However, the comments by our study physicians suggest that the relationships between the physicians and the groups with which they are affiliated are themselves fraught with distrust. The group is often seen as acting in a policing rather than a supportive role. Rather than owning the IPA data, physicians dispute information provided by IPAs. Instead of proactively managing patients, they treat prompts from external bodies (IPAs and health plans) as unjustified additions to their workload.

The distrust within the California initiative appears to be further related to the structure and dynamics of the U.S. health care system. Partly, it may be a function of change itself—whatever their problems with the NHS, English GPs are used to it, while the constancy of change in California can only destabilize expectations. Partly, it should follow from the fact that the data in California really are worse. Regardless of the explanations, the fact of distrust means that increasing the percentage of income to be derived from targets, in California, would be likely to add to the perception of targets as threatening, rather than supporting, the delivery of quality care by frontline clinicians.

There are some indications that American advocates of improved payment incentives for quality are beginning to recognize the organizational dimension. On balance, however, advocates such as the Medicare Payment Advisory Commission see the problem as one of inefficient payment systems that might inhibit delivery system reform, whereas our results suggest that attention should be paid to the organization of delivery, since this will determine how any preferred payment systems would function in practice (MedPAC 2008). A close look at the English experience shows that the record of "success" from the QOF framework in England is far more problematic than the raves from some American observers imply. At a minimum, there is doubt that the QOF framework provided extra value for the money. Yet our comparison reveals the even more fundamental point that the organization of delivery in the United States makes paying for performance even more problematic.

All policies require the political will to choose them, knowledge of techniques that will attain their goals, and the institutional power to implement them (White 2003). The eager commentary cited at the beginning of this article is evidence that pay for performance at a minimum is politically attractive enough to draw widespread interest. Yet our comparison shows that it faces significant technical obstacles in all cases and particularly severe institutional obstacles in the United States.

References

Atoji, C. 2008. Pay-for-Performance Yields Incremental Results and Outcomes-Based ROI. *Digital Healthcare and Productivity*, February 19. www.digitalhcp.com/DigitalHealthCare_Article.aspx?id=72288.

Baker, D., and E. Middleton. 2003. Cervical Screening and Health Inequality in England in the 1990s. *Journal of Epidemiology and Community Health* 57:417–423.

Baker, G., and C. Carter. 2005. *Provider Pay-for-Performance Incentive Programs: 2004 National Study Results*. San Francisco: MedVantage.

Batty, D. 2003. Q&A: GP Contract. *Guardian*, June 20. www.guardian.co.uk/society/2003/jun/20/politics.theissuesexplained.

Bodenheimer, T., R. A. Berenson, and P. Rudolf. 2007. The Primary Care–Specialty Income Gap: Why It Matters. *Annals of Internal Medicine* 146:301–306.

British Medical Association. 2008. GP Contract 2008/09: Poll Outcome. www.bma.org.uk/ap.nsf/Content/pollresults0308.

Campbell, S., R. McDonald, and H. Lester. 2008. The Experience of Pay for Performance in English Family Practice: A Qualitative Study. *Annals of Family Medicine* 6:228–234.

Campbell, S., M. Roland, E. Middleton, and D. Reeves. 2005. Improvements in the Quality of Clinical Care in English General Practice 1998–2003: Longitudinal Observational Study. *BMJ* 331:1121–1123.

Casalino, L. P. 2004. Physicians and Corporations: A Corporate Transformation of American Medicine? *Journal of Health Policy, Politics and Law* 29:869–883.

Checkland, K., R. McDonald, and S. Harrison. 2007. Ticking Boxes and Changing the Social World: Data Collection and the New UK General Practice Contract. *Social Policy and Administration* 41:693–710.

Christianson, J. B., S. Leatherman, and K. Sutherland. 2008. Lessons from Evaluations of Purchaser Pay-for-Performance Programs. *Medical Care Research and Review* 65 (suppl.): 5S –35S.

Damberg, C., K. Raube, T. Williams, and S. M. Shortell. 2005. Paying for Performance: Implementing a Statewide Project in California. *Quality Management in Health Care* 14:66–79.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland. 2006. Family Practice Performance in the First Year of the UK's New "Pay for Performance" Scheme: Good Clinical Practice or Gaming? *New England Journal of Medicine* 355:375–384.

Doran, T., C. Fullwood, E. Kontopantelis, and D. Reeves. 2008a. 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* 372:728–736.

Doran, T., C. Fullwood, D. Reeves, H. Gravelle, and M. Roland. 2008b. Should Physicians Be Able to Exclude Individual Patients from Pay-for-Performance Targets?

Analysis of Exception Reporting in the English Pay-for-Performance Scheme. *New England Journal of Medicine* 359:274–284.

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

Felt-Lisk, S., G. Gimm, and S. Peterson. 2007. Making Pay-for-Performance Work in Medicaid. *Health Affairs*, Web Exclusive 26:w516–w527. June 26. content.healthaffairs.org/cgi/content/full/hlthaff.26.4.w516v1/DC1.

Galvin, R. 2006. Pay-for-Performance: Too Much of a Good Thing? A Conversation with Martin Roland. *Health Affairs*, Web Exclusive 25:w412–w419. September 5. content.healthaffairs.org/cgi/content/full/hlthaff.25.w412v1/DC1.

Gillies, R., S. Shortell, L. Casalino, J. Robinson, and T. Rundall. 2003. How Different Is California? A Comparison of U.S. Physician Organizations. *Health Affairs* 22 (October 15): w492–w502. content.healthaffairs.org/cgi/content/full/hlthaff.w3.492v1/DC1.

Gravelle, H., M. Sutton, and A. Ma. 2008. Doctor Behaviour under a Pay for Performance Contract: Further Evidence from the Quality and Outcomes Framework. Centre for Health Economics (CHE) research paper 34. York: CHE, University of York.

Grumbach, K., J. Coffman, K. Vranizan, N. Blick, and E. H. O'Neil. 1998. Independent Practice Association Physician Groups in California. *Health Affairs* 17 (3): 227–237.

Hing, E. S., C. W. Burt, and D. A. Woodwell. 2007. Electronic Medical Record Use by Office-Based Physicians and Their Practices: United States, 2006. Advance Data from Vital and Health Statistics, no. 393. Hyattsville, MD: National Center for Health Statistics.

Horton, R. 2001. The Real Lessons from Harold Frederick Shipman. *Lancet* 357:82–83.

Institute of Medicine (IOM). 2001. *Crossing the Quality Chasm: A New Health System for the Twenty-first Century*. Washington, DC: National Academies Press.

———. 2007. *Rewarding Provider Performance: Aligning Incentives in Medicare*. Washington, DC: National Academies Press.

Integrated Healthcare Association. 2007. *California Health Plans Pay $65 Million to Improve Performance in Patient Care*. Press release, February 27, 2008. www.iha.org/pressrel/Final%20Press%20Release_2007%20Results%20Payouts.pdf.

Jordan, K., M. Porcheret, and P. Croft. 2004. Quality of Morbidity Coding in General Practice Computerized Medical Records: A Systematic Review. *Family Practice* 21:396–412.

Kingdon, J. 1994. *Agendas, Alternatives, and Public Policies*. New York: HarperCollins.

Lewis, J. 1997. Primary Care — Opportunities and Threats: The Changing Meaning of the GP Contract. *BMJ* 314:895.

Lindenauer, P. K., D. Remus, S. Roman, M. Rothberg, E. Benjamin, A. Ma, and D. Bratzler. 2007. Public Reporting and Pay for Performance in Hospital Quality Improvement. *New England Journal of Medicine* 356:486–496.

Loudon, I., J. Horder, and C. Webster, eds. 1998. *General Practice under the National Health Service, 1948–1997*. Oxford: Clarendon Press.

March, J., and J. Olsen. 1976. *Ambiguity and Choice in Organisations*. Oslo: Universitetsforlaget.

Marmor, T. 2000. *The Politics of Medicare*. 2nd ed. Edison, NJ: Aldine Transaction.

———. 2004. *Fads in Medical Care Management and Policy, Fads and Fashions in Medical Care Policy and Politics*. Rock Carling Lecture. London: The Stationery Office for Nuffield Trust.

Marmor, T., R. Freeman, and K. G. H. Okma. 2005. Comparative Perspectives and Policy Learning in the World of Health Care. *Journal of Comparative Policy Analysis* 7 (4): 331–348.

Martin, D. 2007. "Gloating" GPs "Delighted" with Lucrative Pay Deal. *Daily Mail*, January 31. www.dailymail.co.uk/news/article-432882/Gloating-GPs-delighted-lucrative-pay-deal.html.

May, C. 2007. The Clinical Encounter and the Problem of Context. *Sociology* 41: 29–45.

McDonald, R., S. Harrison, K. Checkland, S. Campbell, and M. Roland. 2007. Impact of Financial Incentives on Clinical Autonomy and Internal Motivation in Primary Care: An Ethnographic Study. *BMJ* 334:1357–1359.

McDonald, R., J. Waring, and S. Harrison. 2006. At the Cutting Edge? Modernisation and Nostalgia in a Hospital Operating Theatre Department. *Sociology* 40:1097–1115.

Medicare Payment Advisory Commission (MedPAC). 2008. *Report to the Congress: Reforming the Delivery System*. June. Washington, DC: MedPAC.

Mehrotra, A., C. L. Damberg, M. E. S. Sorbero, and S. S. Teleki. 2009. Pay for Performance in the Hospital Setting: What Is the State of the Evidence? *American Journal of Medical Quality* 24:19–28.

Middleton, E., and D. Baker. 2003. Comparison of Social Distribution of Immunisation with Measles, Mumps, and Rubella Vaccine, England, 1991–2001. *BMJ* 326:854–854.

National Audit Office. 2008. *NHS Pay Modernisation: New Contracts for General Practice Services in England*. London: Stationery Office.

Onion, D., and R. M. Berrington. 1999. Comparisons of UK General Practice and U.S. Family Practice. *Journal of the American Board of Family Practice* 12:162–172.

Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: Sage.

Roland, M. 2004. Linking Physician Pay to Quality of Care: A Major Experiment in the UK. *New England Journal of Medicine* 351:1488–1554.

Rosenthal, M., and R. A. Dudley. 2007. Pay for Performance: Will the Latest Payment Trend Improve Care? *Journal of the American Medical Association* 297:740–744.

Rosenthal, M., R. G. Frank, Z. Li, and A. Epstein. 2005. Early Experience with Pay-for-Performance: From Concept to Practice. *Journal of the American Medical Association* 294:1788–1793.

Rosenthal, M., B. Landon, S. Normand, R. Frank, and A. Epstein. 2006. Pay for Performance in Commercial HMOs. *New England Journal of Medicine* 355:1895–1902.

Shekelle, P. 2003. New Contract for General Practitioners. *BMJ* 326:457–458.

Terry, K. 2005. Pay for Performance: How Fast Is It Spreading? *Medical Economics*, November 4, medicaleconomics.modernmedicine.com/memag/article/articleDetail.jsp?id=190108&pageID=2.

Timmins, N. 2005. Do GPs Deserve Their Recent Pay Rise? *BMJ* 31:800.

Triggle, N. 2007. GP Pay "Should Have Been Capped." *BBC*, January 19, news.bbc.co.uk/2/hi/health/6276793.stm.

United Kingdom. Department of Health. 2000. *The NHS Plan: A Plan for Investment, a Plan for Reform*. London: Stationery Office.

Werner, R., D. Goldman, and R. Dudley. 2008. Comparison of Change in Quality of Care between Safety-Net and Non–Safety-Net Hospitals. *Journal of the American Medical Association* 299:2180–2187.

White, J. 2003. Three Meanings of Capacity: Why the Federal Government Is Most Likely to Lead on Health Insurance Access Issues. *Journal of Health Politics, Policy and Law* 28:217–244.