PAY-FOR-PERFORMANCE: TOXIC TO QUALITY? INSIGHTS FROM BEHAVIORAL ECONOMICS

David U. Himmelstein, Dan Ariely, and
Steffie Woolhandler

Pay-for-performance programs aim to upgrade health care quality by tailoring financial incentives for desirable behaviors. While Medicare and many private insurers are charging ahead with pay-for-performance, researchers have been unable to show that it benefits patients. Findings from the new field of behavioral economics challenge the traditional economic view that monetary reward either is the only motivator or is simply additive to intrinsic motivators such as purpose or altruism. Studies have shown that monetary rewards can undermine motivation and worsen performance on cognitively complex and intrinsically rewarding work, suggesting that pay-for-performance may backfire.

People respond to rewards (1). This basic tenet of both economics and psychology underlies pay-for-performance (P4P) programs that aim to upgrade health care quality and efficiency by offering carefully tailored financial incentives for desirable behaviors. P4P has been adopted as a key strategy by the English National Health Service, U.S. Medicare, and many private insurers.

Traditionally, economists have viewed extrinsic (i.e., monetary) reward either as the only motivator or as simply additive to intrinsic motivators such as purpose, altruism, or autonomy. According to this view, higher pay induces better performance.

Interestingly, while higher pay clearly increases performance for straightforward manual tasks such as installing windshields (2), a growing body of evidence from behavioral economics and social psychology indicates that rewards

International Journal of Health Services, Volume 44, Number 2, Pages 203–214, 2014

doi: http://dx.doi.org/10.2190/HS.44.2.a
http://baywood.com

sometimes undermine motivation and worsen performance on complex cognitive tasks, especially when intrinsic motivation is high.

THE LOGIC OF PAY-FOR-PERFORMANCE

Though only rarely made explicit (3, 4), P4P rests on several assumptions:

1. Performance can be accurately ascertained; that is, measurements of clinicians' performance actually reflect their performance, not the nature of their patients, practice setting, or ability to game the system.
2. The current payment system is too simple; more detailed contracts rewarding specific, important aspects of performance will improve quality.
3. Variation in performance is caused by variation in motivation. (If a clinician's poor performance is not intentional, incentive pay is unlikely to improve it.)
4. Financial incentives will add to total motivation, not undermine it.
5. Hospitals and physicians currently delivering poor-quality care should get fewer resources.

Under scrutiny, each of these assumptions appears flawed.

1. Performance can be Accurately Ascertained

Health outcomes such as death or disability are the most salient indicators of performance. Yet patients' health is shaped by myriad genetic, social, behavioral, and random factors that are beyond the clinician's control—not just (or even mostly) the quality of medical care. Isolating the "signal" of medical quality amidst the "noise" of so many other factors is devilishly difficult: bad outcomes sometimes follow good decisions and good outcomes may follow bad decisions. Moreover, the time lag between treatment and ultimate outcome is often long, further clouding the link between performance and results. Even strikingly effective interventions such as blood pressure treatment may take decades to bear fruit, but P4P cannot be effective in such long time horizons.

Finding the needle of performance amidst the haystack of other outcome determinants requires, *inter alia*, robust adjustment for patients' pre-existing health status. Mortality rates are the most obvious candidate for performance measurement, and hospital mortality provides the best-case scenario for development of risk adjusters. Inpatient deaths are frequent, unambiguous, and likely to reflect quality differences. Moreover, time horizons are short and data are available for millions of hospitalizations, facilitating statistical analyses to identify risk adjusters. Yet despite years of effort to develop inpatient risk adjustment, four widely used algorithms yield strikingly divergent rankings of hospital performance (5). Hospitals that appear excellent according to one algorithm can appear downright dangerous according to another.

One reason that risk adjustment is so hard is that the diagnosis—the foundation for risk adjustment—is not solely a patient characteristic, but also reflects the aggressiveness of diagnostic workups (6, 7) and coding practices. In regions of the United States where medical spending is high, intensive diagnostic testing is common and patients are labeled with more diagnoses and comorbidities—what might be called "overdiagnosis." But careful analysis indicates that this apparently greater severity of illness is entirely artifactual (7). It is not hard to imagine how more aggressive investigation might label patients with more diagnoses. More spirometry (a breathing test) or prostate-specific antigen (PSA) testing would generate more chronic obstructive pulmonary disease and prostate cancer diagnoses, inflating the denominators when calculating mortality rates for these conditions. These inflated denominators result in apparently better outcomes, even if aggressive diagnostic testing results in treatments that do more harm than good (as is often the case in prostate cancer). In sum, diagnosis-based risk adjustment falsely inflates the quality scores of providers who do more diagnostic testing.

Similarly, intensive coding—that is, tailoring diagnoses to maximize payment under per-case or risk-adjusted capitation schemes—also makes patients appear sicker on paper, and hence boosts risk-adjusted quality scores. Under U.S. Medicare's diagnosis-related group hospital payment system, recoding a diagnosis as "aspiration pneumonia with acute on chronic systolic heart failure" rather than the synonymous terms "pneumonia with CHF" triples the payment and ups the risk score (8). Such "upcoding" is endemic among health maintenance organizations (HMOs) that contract with Medicare for risk-adjusted capitation payments (9), as well as among hospitals (10). Medicare Advantage plans extracted overpayments totaling $30 billion (all dollar amounts in U.S. dollars) in 2007, largely by gaming the risk adjustment formula (9). One Maryland hospital reportedly urged physicians to document "protein malnutrition" in patients' charts, allowing the hospital to code (and bill for) 287 cases of comorbid "kwashiorkor" in 2007 (up from zero cases in 2004) (11).

Process-based quality metrics, although easier to calculate than risk-adjusted outcomes, are pale proxies for global quality performance. Even seemingly clear-cut measures have hidden complexity: total hospital readmission rates correlate poorly with avoidable readmission rates (12); initiating therapy for pneumonia patients within four hours of arrival correlates with quality, yet paying hospitals to do so caused the willy-nilly administration of antibiotics to almost any emergency department patient with a cough (13).

Patients' social characteristics also confound process-based measures; even excellent doctors who care for disadvantaged or difficult patients often look bad on current P4P metrics. Among physicians at Massachusetts General Hospital (a flagship Harvard teaching hospital), those caring for more minority, non-English-speaking, poor, and uninsured patients, as well as patients with infrequent visits, scored low on P4P metrics (14). A simulation in Massachusetts estimated that P4P would penalize physicians caring for the poor by $7,100 annually (15).

Using clinical audits for financial reward or punishment, rather than a collegial and reflective effort to upgrade care, amplifies these challenges to performance measurement. Payment incentives may mutate honesty and goodwill into legal trickery, leaving the clinical data needed for real quality improvement as accurate as a tax return (16).

Cheating may so thoroughly distort reality that rewards become uncoupled from actual performance. If gaming were limited to a few bad apples, policing P4P might be workable. But experiments and experience indicate that good people frequently cheat—while deceiving themselves into believing that they are not really cheating. Rewards make us view the world differently.

Such cognitive distortion—"motivated reasoning"—was described in a classic 1954 study of football fans' perceptions of which team was the aggressor and which the victim in a particularly bloody game between Dartmouth and Princeton (17). More recently, experiments documented widespread cheating among subjects (Massachusetts Institute of Technology (MIT) and Yale students) by offering payments for scoring high on a self-graded math task; subjects consistently added a few points when they thought no one could check (18). Strikingly, despite incentives to predict accurately, cheaters inflated their predictions of how they would score on a future, monitored test, suggesting self-deception.

In sum, neither current nor foreseeable P4P metrics can reliably ascertain global quality and differentiate variations due to providers' performance from those due to their gaming efforts, their patients' characteristics, or random variations. The low signal:noise ratio may negate P4P's educational value; when measured performance differs from total contribution, rewarding a subset of seemingly desirable behaviors may mis-educate practitioners, distort rather than improve performance, and worsen overall quality.

2. The Current Payment System is Too Simple

P4P replaces looser, more general payment contracts with ones specifying the "deliverables" in greater detail. For instance, contracts for Accountable Care Organizations mandated under the 2010 Affordable Care Act in the United States incentivize 33 quality standards, while Britain's primary care P4P program initially tabulated 146 parameters, with more to come (19). Yet when it comes to contractual detail, more may not be better.

The optimal specificity of contracts has interested economists at least since Coase's 1937 paper on the nature of firms (20)—work that was recognized with a Nobel economics prize and laid the foundation for Oliver Hart's pioneering work on incomplete contracts (21). Incomplete contracts are akin to handshake deals (22). They lay out only the general parameters of the exchange (e.g., spend 30 minutes with the patient), while details and unexpected contingencies are covered by social and professional norms. In contrast, complete contracts cement the deal with an airtight agreement specifying every detail and contingency in advance.

Coase and Hart noted the exorbitant administrative and legal costs of spelling out and enforcing complete contracts (20, 21)—a phenomenon familiar to clinicians (23). (Indeed, they posited that these transactional inefficiencies drive entrepreneurs to form firms rather than outsourcing all tasks.)

Costly administration is not the only downside of complete contracts. If something is omitted from an exquisitely detailed agreement, there is no presumption of default to goodwill—it is happy hunting season. When one of the authors (D.A.) asked the dean of Duke University's Law School about its honor code, he replied that it amounted to little more than "Don't do anything dishonorable." Lists of rules ("Don't raise chickens in your dorm room"; "Don't smoke hashish") implicitly permit everything else.

Moreover (as detailed below), more prescriptive contracts may be perceived as controlling, and thus undermine the intrinsic motivation critical to maintaining quality when no one is looking. When specifying every detail and contingency is not possible, as is clearly the case in medicine, it may be better to rely on professional and social norms (24).

3. Variation in Performance is Caused by Variation in Motivation

Studies have pinpointed many causes of quality breeches in medical care: fatigue; poorly designed workflow and care systems; undue commercial influence; knowledge gaps; memory lapses; reliance on inappropriate heuristics; poor interpersonal skills and insufficient teamwork, to name just a few. But "not trying" is rarely cited.

We are unaware of systematic evidence to support or reject the notion that doctors' lack of motivation is a common cause of poor medical performance.

4. Financial Incentives Will Add to Doctors' Intrinsic Motivation, Not Undermine It

The simple model of reward-induced performance ignores the complexity of human drive, particularly the role of intrinsic motivation—the desire to perform an activity for its own inherent rewards. Financial incentives may "crowd out" intrinsic motivation. Offering your dinner party host a $10 reward for cooking a fine meal is unlikely to motivate future invitations.

Deci's 1971 experiment (25) first identified negative effects of financial incentives. Students would spontaneously play with interesting puzzles, but once they had been paid to solve them, they lost interest in playing for free. The introduction of payment diminished the internal drive to explore an intellectual challenge.

A 1973 randomized controlled trial (RCT) documented motivational crowd-out among blood donors. Among frequent (presumably highly motivated) donors, an incentive payment (about $55 in today's dollars) decreased donations (26). In contrast, incentive payments increased donations among those who had not

donated for years. A Swiss study of volunteer work reached a similar conclusion: unpaid volunteers worked, on average, four hours more monthly than those offered a small payment (27).

Financial incentives also backfired in an RCT in Israeli day care centers (28). When centers began fining parents for picking up their children late, tardiness increased—a phenomenon not observed at the control centers. Strikingly, the late pickup rate remained high even after the fines were rescinded. The fine had transformed promptness from a moral duty to a market transaction governed by price.

Moreover, upping performance-based payments may not overcome motivational crowd-out; RCTs have found that even very large financial incentives can undermine performance on complex cognitive tasks. Massachusetts Institute of Technology students offered up to $300 for solving mathematical puzzles performed much worse than students offered smaller amounts (29). (In contrast, the highly incentivized students did better on simple tasks requiring only manual effort.) Huge incentives offered to rural villagers in India—about half their annual money income—worsened performance on complex memory and puzzle-solving tasks (29). High-stakes incentives may be distracting, interfering with cognitive focus and creativity.

A meta-analysis summarizing 128 studies indicates that such findings are representative of a consistent body of experimental data (30). The conclusions that emerge from the extensive literature on motivational crowd-out include (31):

- Tangible rewards—particularly monetary ones—undermine motivation for tasks that are intrinsically interesting or rewarding—an effect that is quite large.
- Symbolic rewards (e.g., praise or flowers) do not crowd out intrinsic motivation and may augment it.
- The negative effects of monetary rewards are strongest for complex cognitive tasks.
- Crowding-out effects tend to reduce reciprocity and augment selfish behaviors.
- Crowding-out may spread (to both other tasks and coworkers), decreasing intrinsic motivation for work not directly incentivized by the monetary rewards.
- Crowding-out is strongest when external rewards are large; perceived as controlling; contingent on very specific task performance; or associated with surveillance, deadlines, or threats.

Finally, motivational crowd-out works in the opposite direction to economists' standard supply curve, where performance rises with price. The net effect of financial rewards depends on the relative size of the price effect and the crowding-out effect. When crowding-out is modest, the classic economic model underlying P4P holds; you get what you pay for. However, if intrinsic motivation is high and crowding-out is strong, payment may worsen performance.

5. Hospitals and Physicians Currently Delivering Poor-Quality Care Should Get Fewer Resources

P4P assumes that financial incentives will goad substandard providers to upgrade care or risk seeing their patients migrate to higher-quality options. Yet when poor performance is due to financial distress (as is sometimes the case [32])—and is hence beyond the provider's control—penalizing low scorers can make matters worse. Safety-net providers are often cash-strapped, score low on P4P metrics (14, 33), and serve many patients without other health care options (34). P4P may push under-resourced providers into a downward spiral, punishing their patients and exacerbating quality disparities.

Hospitals and physician practices delivering irremediably deficient care should be closed—although closures in underserved communities may have grave consequences (34), necessitating clear plans for alternative care options. It makes little sense to put already quality-challenged providers on a starvation diet.

HAS PAY-FOR-PERFORMANCE WORKED SO FAR?

Despite the widespread embrace of P4P, evidence of benefit is slim. Reviews of early, mostly small P4P studies found mixed evidence on improvement on incentivized, process-based measures; virtually no evidence of global quality improvement; and occasional unintended harms (35, 36).

More recent systematic reviews reach similarly agnostic conclusions. A 2011 Cochrane Collaborative overview concluded that "financial incentives may be effective in changing health care professional practice," but found "no evidence that financial incentives can improve patient outcomes" (37). Another 2011 Cochrane review focused on primary care found "insufficient evidence to support or not support the use of financial incentives" (38).

Although recapitulation of these reviews is beyond the scope of this article, we highlight below some key individual P4P studies. Three observational studies of outpatient care in HMOs yielded largely null results. Quality indicators improved at the same rate among Pacificare providers with and without P4P incentives (39). When Kaiser offered financial incentives for cervical cancer and diabetic retinopathy screening, rates initially improved, only to fall below baseline when incentives ended (40). When California's Medicaid program rewarded high-performing HMOs with lucrative automatic enrollment expansions, scores did not improve for any quality measure, and actually worsened for some (41).

England's extensive experience with P4P is also mixed. In 2004, the National Health Service implemented the largest P4P initiative, offering family physicians bonuses that could augment their incomes by 25 percent for meeting 146 quality standards (19). Early results were encouraging; in the first year, physicians achieved 96.7 percent of all possible bonus points (19). However, by 2007, improvement had plateaued for incentivized measures, and quality actually

deteriorated for two measures not linked to incentives (42). Moreover, although doctors met virtually all P4P hypertension treatment targets, neither population blood pressures nor hypertension complications decreased (43).

The lone P4P success, a recent British hospital program that appeared to result in improved outcomes, allocated all P4P funds for further quality improvement programs and prohibited paying anyone a bonus (44). In contrast, a previous British study of financial incentives to hospitals found worrisome side effects. Incentives to shorten surgical queues worsened heart attack mortality (45); focusing resources and attention on elective surgery cases may have distracted from emergency care.

P4P has been even less successful in the more market-oriented U.S. context. In Medicare's Premier Hospital Quality Incentive Demonstration, the largest U.S. P4P program, the 200 participating hospitals' process-based quality indicators improved more rapidly than control hospitals' over the first two years, according to an oft-cited study (46). But differences evaporated by five years (47) and patient outcomes did not improve at all (48, 49). Incentives targeted to low-performing hospitals were also ineffective (50).

Finally, the impact of P4P on professional performance has been tested in two large RCTs in public schools. A $75 million study involving more than 200 high-needs New York City schools employing more than 20,000 teachers offered incentives of up to $3,000 per teacher based on students' test scores, graduation and attendance rates, and results of learning environment surveys. Notably, most sites opted to pool bonuses among their teachers—the type of institution-level incentives that some P4P proponents advocate. Yet, ". . . incentives . . . did not increase student achievement in any meaningful way. If anything, student achievement declined" (51). In a Tennessee RCT, middle school students whose mathematics teachers were offered P4P bonuses of up to $15,000 based on standardized test results scored no higher than students of control teachers (52).

These disappointing studies have scarcely cooled payers' and policymakers' ardor for P4P (53). Moreover, even scholars who remain agnostic regarding the benefits of P4P have focused mostly on technical specification problems (e.g., identifying better yardsticks for performance and the right mix of incentives, or fine-tuning risk adjustment) (54). Few have countenanced the possibility that P4P may simply not work.

CONCLUSIONS

Paying for quality has strong intuitive appeal; at first blush it seems an obvious path to better performance in medicine, as well as in education, finance, and many other fields. But this belief rests on scant evidence. Rats in mazes predictably follow a pattern of incentive-induced performance; human responses are more complex and context-specific. P4P might work for simple, manual tasks such as pushing a lever or checking a box. But does it really work for the complex array of

tasks that constitute good doctoring? Will incentive payments increase motivation, or alienate doctors and nurses from their work? Here, as is often true in medicine, evidence is required because intuition may mislead.

Thus far, studies have unearthed a variety of bad ways to pay doctors, and no particularly good one. Financing shortcuts cannot circumvent the hard work and commitment needed for quality improvement, and may corrode the indispensable tools of progress: conscientious data collection, honest self-reflection, altruism, and creativity.

None can doubt medicine's grave quality deficits and cost excesses. As a remedy, P4P suggests manipulating greed, a fuel that has powered exponential growth in productivity in the overall economy (55). But Adam Smith, who first recognized greed's awesome power, was also a moral philosopher who believed that commodity production required a parallel public service economy driven by social duty (55, 56).

Sadly, greed has caused many of the worst abuses within the current health care system. Injecting different monetary incentives into health care can certainly change it, but not necessarily in the ways that we would plan, much less hope for.

REFERENCES

1. Skinner, B. F. *Science and Human Behavior*. Macmillan, New York, 1953.
2. Lazear, E. P. Performance Pay and Productivity. NBER Working Paper No, 5672. National Bureau of Economic Research, Cambridge, MA, July, 1996.
3. Berwick, D. Appraising appraisal. *Qual. Connect.* 1(1):1–4, 1991.
4. Schiff, G. D., and Goldfield, N. I. Deming meets Braverman: Toward a progressive analysis of the continuous quality improvement paradigm. *Int. J. Health Serv.* 24:655–673, 1994.
5. Shahian, D. M., et al. Variability in the measurement of hospital-wide mortality rates. *N. Engl. J. Med.* 363:2530–2539, 2010.
6. Song, Y., et al. Regional variations in diagnostic practices. *N. Engl. J. Med.* 363:45–53, 2010.
7. Welch, H. G., et al. Geographic variation in diagnosis frequency and risk of death among Medicare beneficiaries. *JAMA* 305:1113–1118, 2011.
8. Administrative Consultant Services. Reporting secondary diagnoses that impact MS-DRG payment. *MS-DRG News*, October 1, 2011. http://www.acsteam.net/sites/acs/uploads/documents/newsletters/secondarydx2011.pdf (accessed May 15, 2012).
9. Brown, J., et al. How Does Risk Selection Respond to Risk Adjustment? Evidence from the Medicare Advantage Program. National Bureau of Economic Research Working Paper No. 16977. NBER, Cambridge, MA, April, 2011.
10. Silverman, E., and Skinner, J. Medicare upcoding and hospital ownership. *J. Health Econ.* 23:369–389, 2004.
11. Anonymous. Maryland hospital accused of inflating treatment claims. *Mod. Healthc.* 41(43):4, 2011.

12. van Walraven, C., et al. Incidence of potentially avoidable urgent readmissions and their relation to all-cause urgent readmissions. *Can. Med. Assoc. J.* 183:E1067–1072, 2011. Epub August 22, 2011.
13. Kanwar, M., et al. Misdiagnosis of community acquired pneumonia and inappropriate utilization of antibiotics: Side effects of the 4-h antibiotic administration rule. *Chest* 131:1865–1869, 2007.
14. Hong, C. S., et al. Relationship between patient panel characteristics and primary care physician clinical performance rankings. *JAMA* 304:1107–1113, 2010.
15. Friedberg, M. W., et al. Paying for performance in primary care: Potential impact on practices and disparities. *Health Aff.* 29:926–932, 2010.
16. Hart, J. T. Two paths for medical practice. *Lancet* 340:772–775, 1992.
17. Hastorf, A. H., and Kantril, H. They saw a game: A case study. *J. Abnorm. Psychol.* 49:129–134, 1954.
18. Mazar, N., Amir, O., and Ariely, D. The dishonesty of honest people: A theory of self-concept maintenance. *J. Mark. Res.* 45:633–644, 2008.
19. Doran, T., et al. Pay-for-performance programs in family practices in the United Kingdom. *N. Engl. J. Med.* 355:373–384, 2006.
20. Coase, R. H. The nature of the firm. *Economica* 4(16):386–405, 1937.
21. Hart, O., and Moore, J. Incomplete contracts and renegotiation. *Economica* 56: 755–785, 1988.
22. Ariely, D. In praise of the handshake. *Harv. Bus. Rev.* 89(3):40, 2011.
23. Woolhandler, S., and Himmelstein, D. U. The deteriorating administrative efficiency of U.S. health care. *N. Engl. J. Med.* 324:1253–1258, 1991.
24. Vladeck, B. Ineffective approach. *Health Aff.* 23:285–286, 2004.
25. Deci, E. L. Effects of externally mediated rewards on intrinsic motivation. *J. Pers. Soc. Psychol.* 18(1):105–115, 1971.
26. Upton, W. E., III. Altruism, Attribution and Intrinsic Motivation in the Recruitment of Blood Donors. Doctoral thesis, Cornell University Department of Social Psychology, 1973.
27. Frey, B. S., and Goette, L. Does Pay Motivate Volunteers? Institute for Empirical Research in Economic Working Paper No. 7. May 1999. http://www.iew.uzh.ch/wp/iewwp007.pdf (accessed November 21, 2011).
28. Gneezy, U., and Rustichini, A. A fine is a price. *J. Legal Stud.* 29(1):1–17, 2000.
29. Ariely, D., et al. Large stakes and big mistakes. *Rev. Econ. Stud.* 76:451–469, 2009.
30. Deci, E. L., Koestner, R., and Ryan, R. M. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol. Bull.* 125:627–668, 1999.
31. Frey, B. S. Motivational crowding theory – a new approach to behavior. In *Behavioral Economics and Public Policy, Roundtable Proceedings, Melbourne, August 8–9, 2007*. Commonwealth of Australia, Melbourne, 2008. http://www.pc.gov.au/__data/assets/pdf_file/0005/79250/behavioural-economics.pdf (accessed August 20, 2011).
32. Ly, D. P., Jha, A. K., and Epstein, A. M. The association between hospital margins, quality of care, and closure or other change in operating status. *J. Gen. Intern. Med.* 26:1291–1296, 2011.
33. Werner, R. M., Goldman, E., and Dudley, R. A. Comparison of change in quality of care between safety-net and non-safety net hospitals. *JAMA* 299:2180–2187, 2008.

34. Walker, K. O., et al. Effect of closure of a local safety-net hospital on primary care physicians' perceptions of their role in patient care. *Ann. Fam. Med.* 9:496–503, 2011.
35. Rosenthal, M. B., and Frank, R. J. What is the empirical basis for paying for quality in health care? *Med. Care Res. Rev.* 63(2):135–157, 2006.
36. Petersen, L. A., et al. Does pay-for-performance improve the quality of care? *Ann. Intern. Med.* 145:265–272, 2006.
37. Flodgren, G., et al. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviors and patient outcomes. Cochrane Database of Systematic Reviews 2011, Issue 7, Art. No. CD009255.
38. Scott, A., et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database of Systematic Reviews 2011, Issue 9, Art. No. CD008451.
39. Mullen, K. J., Frank, R. G., and 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.
40. Lester, H., et al. The impact of removing financial incentives from clinical quality indicators: Longitudinal analysis of four Kaiser Permanente indicators. *BMJ* 340:c1898, 2010.
41. Guthrie, B., Auerback, G., and Bindman, A. B. Health plan competition for Medicaid enrollees based on performance does not improve quality of care. *Health Aff.* 29: 1507–1516, 2010.
42. Campbell, S. M., et al. Effects of pay for performance on the quality of primary care in England. *N. Engl. J. Med.* 361:368–378, 2009.
43. Serumaga, B., et al. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom. *BMJ* 342:D108, 2011.
44. Sutton, M., et al. Reduced mortality with hospital pay for performance in England. *N. Engl. J. Med.* 367:1821–1828, 2012.
45. Propper, C., Burgess, S., and Gossage, D. Competition and quality: Evidence from the NHS internal market 1991–1999. *The Econ. J.* 118(1):138–170, 2008.
46. Lindenauer, P. K., et al. Public reporting and pay for performance in hospital quality improvement. *N. Engl. J. Med.* 356:486–496, 2007.
47. Werner, R. M., et al. The effect of pay-for-performance in hospitals: Lessons for quality improvement. *Health Aff.* 30:690–698, 2011.
48. Jha, A. K., et al. The long-term effect of premier pay for performance on patient outcomes. *N. Engl. J. Med.* On-line first, March 28, 2012. http://www.nejm.org/doi/pdf/10.1056/NEJMsa1112351 (accessed April 15, 2012).
49. Glickman, S. W., et al. Pay for performance, quality of care and outcomes in acute myocardial infarction. *JAMA* 297:2373–2380, 2007.
50. Ryan, A. M., Blustein, J., and Casalino, L. P. Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Aff. (Millwood)* 31:797–805, 2012.
51. Fryer, R. G. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. National Bureau of Economic Research Working Paper No 16850. NBER, Cambridge, MA, March, 2011. http://www.economics.harvard.edu/faculty/fryer/files/teacher%2Bincentives.pdf (accessed August 14, 2011).

52. Springer, N. G., et al. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives at Vanderbilt University, Nashville, TN, 2010. http://www.rand.org/pubs/reprints/2010/RAND_RP1416.pdf (accessed August 14, 2011).
53. McKinney, M. Paying for performance: Insurers expected to follow WellPoint's lead. *Mod. Healthc.* May 23, 2011.
54. Rosenthal, M. B., et al. Climbing up the pay-for-performance learning curve: Where are the early adopters now? *Health Aff.* 26:1674–1682, 2007.
55. Hart, J. T. Worlds of difference. *Lancet* 371:1883–1885, 2008.
56. Smith, A. *An Enquiry into the Nature and Causes of the Wealth of Nations*. Oxford University Press, Oxford, 1993.

Direct reprint requests to:

David U. Himmelstein, MD
255 West 90th Street
New York, NY 10024

dhimmels@hunter.cuny.edu