
MARKET WATCH

Taking Stock Of Pay-For-Performance: A Candid Assessment From The Front Lines

Data from the nation's largest P4P program show some positive changes but no breakthrough improvements in the quality of care.

by Cheryl L. Damberg, Kristiana Raube, Stephanie S. Teleki, and Erin dela Cruz

ABSTRACT: Pay-for-performance (P4P) has been widely adopted, but it remains unclear how providers are responding and whether results are meeting expectations. Physician organizations involved in the California Integrated Healthcare Association's (IHA) P4P program reported having increased physician-level performance feedback and accountability, speeded up information technology adoption, and sharpened their organizational focus and support for improvement in response to P4P; however, after three years of investment, these changes had not translated into breakthrough quality improvements. Continued monitoring is required to determine whether early investments made by physician organizations provide a basis for greater improvements in the future. [*Health Affairs* 28, no. 2 (2009): 517-525; 10.1377/hlthaff.28.2.517]

PAY-FOR-PERFORMANCE is being deployed at all levels of the U.S. health system to stimulate improvements in quality and, more recently, the cost-efficiency with which care is delivered.¹ Despite the widespread adoption of such programs and national policy recommendations for expanding the use of pay-for-performance (P4P) across all health care settings, these efforts remain experimental, given limited empirical evidence about whether P4P works and what mix of strategies work best under which conditions.²

Economic theory posits that tying payment to performance will induce providers of care to change their behavior.³ Although payment and performance increasingly are being linked in

health care, little is known about what behavior changes have occurred as a result of P4P and whether the changes made by providers have led to desired improvements. Absent well-designed research studies on P4P or, in most cases, any type of formal evaluation, the challenge for decisionmakers, moving forward, is how to extract information from existing P4P experiments to inform better policy making and program design. Recent efforts by the Centers for Medicare and Medicaid Services (CMS) to implement national value-based purchasing for physicians and hospitals underscores the importance of extracting lessons being learned from existing P4P programs.⁴

The views and experiences of those on the front lines of implementing P4P payment re-

Cheryl Damberg (damberg@rand.org) is a senior researcher, Health Program, at RAND in Santa Monica, California. Kristiana Raube is an adjunct professor in the Haas School of Business at the University of California, Berkeley. Stephanie Teleki is a policy analyst and Erin dela Cruz is a research assistant at RAND.

forms offer a rich source of information to fill this knowledge void. We interviewed key stakeholders involved in the California Integrated Healthcare Association's (IHA's) state-wide P4P program. Since 2003 the IHA has operated the largest P4P experiment in the United States. The program targets 225 capitated integrated medical groups and independent practice associations (IPAs), which contract with the seven largest health maintenance organization (HMO) plans in California.⁵ The physician organizations represent approximately 35,000 physicians who care for more than 6.2 million patients enrolled in commercial HMO and point-of-service (POS) plans. The IHA program is unique because of the alignment of P4P measures across seven major payers whose enrollees constitute roughly 60 percent of the revenue stream for the physician organizations.⁶

The physician organizations receive capitated payments from health plans, through either shared- or full-risk contracts. IHA-related P4P bonus payments are made by the plans to physician organizations as the contracting unit. Individual physicians within each organization are paid on a salary, capitated, or fee-for-service (FFS) basis and may be subject to performance-based incentives established by their organization.

Annually, the IHA produces performance scores for the physician organizations, using a uniform measure set. Many of the clinical measures mimic Healthcare Effectiveness Data and Information Set (HEDIS) measures, and the program seeks to align the accountability of physician organizations on HEDIS measures with measures for which the health plans are accountable. Scores are generated from combined health plan administrative data or audited data reported by the physician organizations. Annual performance-based payments for clinical quality, patient experience, and information technology (IT) capabilities have been made to the physician organizations since 2004. Between 2004 and 2007, health plans collectively paid \$203 million in IHA-related incentive payments.

Our study addressed four questions: (1)

Has P4P affected the behavior of physician organizations and, if so, how? (2) What challenges do physician organizations face in making quality improvements? (3) Is P4P helping the stakeholders meet their specified goals? (4) What challenges need to be addressed to sustain stakeholders' engagement moving forward? We solicited the perspectives of physician organizations, health plans, and purchasers three years into the P4P experiment, to allow sufficient time for the physician organizations to engage in the program and make changes to their systems and processes and for the program to demonstrate impact.

Study Data And Methods

■ **Sample and data collection.** From the 182 physician organizations that contracted with the seven HMOs as of the 2006 payout and had performance scores, we selected a stratified sample of thirty-five physician organizations (19.2 percent of the total).⁷ We sampled physician organizations along the dimensions of (1) size, to account for differences in resources available for systems investments and quality improvement (QI); (2) geographic region, to capture variation in market competition and capitation rates; (3) organizational model (medical group/IPA), to reflect varying levels of integration; and (4) performance, to capture the experiences of "winners" and "losers."⁸

Our sample also included all seven health plans and the two leading purchaser representatives involved in the program (one of which represents fifty major purchasers of care). We conducted semistructured one-hour telephone interviews between January and July 2007 with the chief executive officer (CEO) or medical director of each selected physician organization, the P4P medical director from each health plan, and the purchaser representatives.

■ **Survey content.** The interview protocols were developed in consultation with P4P experts and the IHA P4P Steering Committee, and included closed-ended 1-5 Likert rating scales, categorical response items, and open-ended response items. Physician organizations were asked about the following topics: organi-

zational culture for quality and support for QI; ability to monitor quality performance at the physician organization and physician levels; incentive payments and presence or absence of a return on investment (ROI); level of engagement and commitment to the IHA P4P program; actions taken in response to P4P; observed unintended consequences of the program; the value of public reporting of performance results; challenges faced in driving improvements within their organizations; and recommendations for future program modifications. Health plan and purchaser leaders were asked about the following topics: level of engagement in P4P; metrics used to gauge program success and ratings of performance against those metrics; strengths and weaknesses of the P4P program; percentage of total capitation the incentive should represent and how to fund the incentive; and recommendations for future program modifications.

We computed means and proportions for the scaled and categorical response items. We summarized the qualitative responses by identifying and organizing the information into themes that emerged from the interviews, and we generated counts of the number of respondents reporting a view or having taken a type of action.

■ **Respondents.** We completed interviews with thirty-five unique physician organizations, seven health plans, and two purchasers. Of the initially selected thirty-five physician organizations, fourteen did not respond or were ineligible; these were replaced with similar physician organizations matched on region, group size, performance, and medical group/IPA model. Also, in some cases, groups refused to answer certain questions. Our final sample (Exhibit 1) included a somewhat greater proportion of organizations that were larger, were based in Northern California, and performed better as compared to the general population of physician organizations.

Study Results

■ **A shift in behavior.** Twenty-five of the physician organizations (N = 31) believed that P4P has directly affected organizational be-

havior by increasing accountability for quality, influencing the speed of IT adoption, improving data collection for quality management, and creating greater organizational focus and support for quality programs and goals. Twenty organizations also thought that the IHA P4P program had affected the behavior of their individual physicians, leading physicians to implement patient management systems, collaborate with administration and staff to strengthen QI efforts, and perform more intensive outreach to patients.

Exhibit 2 shows the most common actions physician organizations reported taking in response to P4P. The twenty-one physician organizations (N = 35) that hired additional staff did so mostly to strengthen IT capabilities and capture data, observing that a constraint to doing well in the P4P program was the inability to fully capture data to demonstrate positive performance. Seventeen organizations invested substantial resources in self-reporting data to the IHA instead of relying on aggregated plan data. Twenty-three organizations reported that their investments affected their 2006 performance and payout, particularly investing in the capacity to give feedback to individual physicians and investments made in IT for self-reporting and data mining.

Twenty-nine of the physician organizations (N = 35) had either modified or developed physician-level incentives to align with the IHA measures to strengthen signals to the individual providers. Quality incentives were tied more heavily to primary care physicians' (PCPs') performance than to that of specialists, and incentives were relatively small, ranging between 1 and 5 percent of an average PCP's salary (approximately \$1,700–\$8,500)—an amount some physician organization leaders asserted "isn't worth the trouble." Twenty-one organizations felt that 10 percent or more of a physician's income needs to be tied to quality to affect his or her behavior.

■ **QI efforts hindered by challenges in monitoring performance.** Despite reported investments and good organizational support for addressing quality issues (mean of 4.1, Exhibit 3), the physician organizations felt chal-

EXHIBIT 1
Descriptive Characteristics Of California Physician Organizations (POs) In Study Sample, Integrated Healthcare Association (IHA) Pay-For-Performance (P4P) Program, 2007

	POs in the IHA P4P program (N = 182)	Number	POs Interviewed (n = 35)	Number	p value ^a
Mean no. of members enrolled	38,703		63,066		0.002
<30,000 (small)	57.7%	105	31.4%	11	0.005
30,001-100,000 (medium)	35.7	65	54.3	19	
>100,000 (large)	6.6	12	14.3	5	
PO model					0.307
Medical group	40.1%	73	48.6%	17	
IPA	59.9	109	51.4	18	
Region					0.012
Northern rural	6.6%	12	17.1%	6	
Central Valley	10.4	19	11.4	4	
Bay Area	8.8	16	20.0	7	
Santa Barbara, Ventura, Orange	14.3	26	5.7	2	
Los Angeles	37.4	68	34.3	12	
Riverside, San Bernardino, San Diego	22.5	41	11.4	4	
Performance on P4P measures					0.007
High performance, most clinical measures fall in 4th quartile	19.8%	36	40.0%	14	
Medium performance, most clinical measures fall in 2nd or 3rd quartile	55.5	101	34.3	12	
Low performance, most clinical measures fall in the 1st quartile	24.7	45	25.7	9	

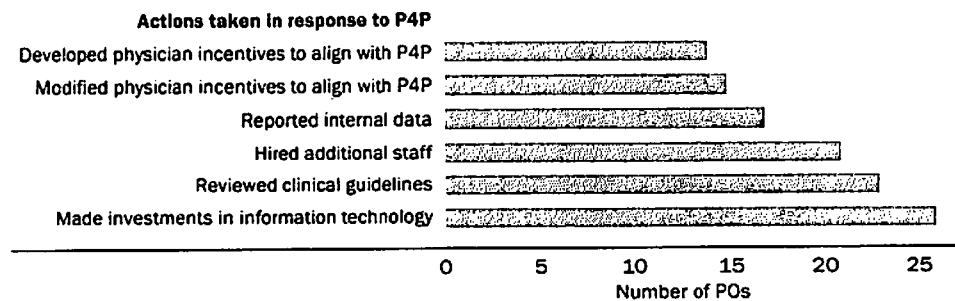
SOURCE: Data collected as part of the IHA P4P program and by the authors as part of the program evaluation.

NOTE: IPA is independent practice association.

^aBased on chi-square goodness-of-fit test comparing the sample of 35 POs to the universe of 182 POs participating in the IHA program.

lenged in monitoring quality performance, especially at the physician level (mean of 3.9). Most physician organizations indicated that they report aggregated HEDIS or IHA measure results and utilization data to their physicians, but only a handful report physician-specific

EXHIBIT 2
California Physician Organizations' (POs') Responses To Pay-For-Performance (P4P), 2007



SOURCE: Data collected by the authors during interviews with thirty-five physician group leaders, 2007.

**EXHIBIT 3
Perspectives Of California Physician Organizations (POs) In Study Of Integrated Health Association (IHA) Pay-For-Performance (P4P) Program, 2007**

	Rating scale (1-5), mean by performance level				
	Overall	High	Medium	Low	p value ^a
Views on P4P design elements					
Importance of the IHA P4P program to the PO (1 = not important, 5 = very important)	4.0	4.2	3.8	4.0	0.636
Importance of public reporting (1 = little/no importance, 5 = very important)	3.0	3.3	2.9	2.5	0.427
Increasing the incentive as percent of total capitation (1 = lower, 3 = about right, 5 = higher)	4.4	4.7	4.4	3.8	0.030
Measurement domains (1 = little/no importance, 5 = very important)					
Clinical measures	4.6	4.7	4.5	4.7	0.583
Patient experience	3.9	4.1	3.7	3.8	0.560
Information technology capability	4.1	4.4	4.4	3.3	0.018
Views on quality environment and support for quality					
Organizational culture of quality (1 = weak, 5 = strong)	4.7	4.8	4.6	4.7	0.500
Support organization dedicates to addressing quality issues (1 = very little/no support, 5 = strong support)	4.1	4.4	4.2	3.6	0.081
Success in monitoring POs' quality performance (1 = not successful, 5 = very successful)	3.9	4.1	3.7	3.7	0.246

SOURCE: Data collected by the authors during interviews with thirty-five physician group leaders, 2007.

^aBased on the Kruskal-Wallis nonparametric test adjusted for ties.

performance results.

Physician organizations face data and IT challenges as obstacles to routine monitoring and feedback for QI and physician-level incentive programs. Large organizations (more than 100,000 enrollees) indicated greater ability to monitor quality compared with smaller organizations. Thus far, the IHA program's incentives for IT capability have been of greater importance to physician organizations that perform well than to those that perform less well (Exhibit 3).

■ **Benefits outweigh adverse consequences.** Although no physician organization was systematically tracking or could quantify the frequency of occurrence, nine gave examples of negative consequences associated with the P4P program, such as dropping noncompliant patients and diverting resources to activities with no health benefit. However, more than two-thirds of the physician organizations reported that the program had resulted in more positives than negatives. Several cau-

tioned that as performance incentives constitute a larger fraction of income or as performance scores compress at the top end, this will increase the likelihood for negative consequences such as gaming the data, ignoring other aspects of quality, or refusing to treat complex patients.

■ **Whole of the program is stronger than the sum of its parts.** Regardless of performance level, twenty-three physician organizations rated the P4P program as being important or very important, regardless of performance level (mean score 4.0, Exhibit 3). All but two cited the alignment of measure sets across participating health plans as very important and "IHA's greatest achievement."

Fifteen physician organizations commented that public reporting creates a positive competitive incentive among physician organizations and focuses management's attention. Yet they rated its importance 2.7 on a scale of 1 to 5, citing that consumers are not using the information.

Roughly half of the organizations reported a positive ROI, meaning that the amount the physician group received in IHA P4P bonus dollars was greater than what they had spent to comply with the program. Six organizations opined that the bonus amount “was barely enough to cover investments the group had to make to measure and manage the processes.”

Physician organizations expressed concern about the weak incentive signal. Bonuses paid in 2006 represented a small portion of the organizations’ total payments, equaling on average 2 percent or less of total capitation. The organizations observed that the broader set of health plan financial incentives place greater emphasis on utilization metrics, leading these organizations to “chase the utilization dollars rather than the quality dollars.” The organizations indicated that they could obtain a far greater financial upside associated with more complete coding of risk factors in the context of risk-adjusted Medicare capitation rates.

There was widespread support (twenty-eight organizations) for increasing incentives at the organization level to 5–10 percent of capitation payments, which would increase physicians’ attention and provide a positive ROI to the organization by defraying setup and compliance costs. Support for increasing the incentive amount was more concentrated among better-performing physician organizations than among the others.

■ **Continued engagement, despite modest gains.** Although physician organizations had responded to P4P, these changes had not translated into success based on health plans’ and purchasers’ self-defined metrics. All seven health plans indicated that the P4P program had not achieved the stated goal of breakthrough quality improvements. Plans rated the success of the program as 2.5 on a scale of 1 to 5 (1 = not successful, 5 = very successful), with one plan commenting, “P4P is an appendage, not a new way of doing business or managing care.” Several plans questioned whether observed improvements were attributable to better data capture versus actual changes in quality, a suspicion echoed by some physician organizations.⁹

Plans indicated that success would translate into improvements in the care of their members relative to national HEDIS benchmarks. Because of alignment with P4P clinical measures, plans anticipated improvements in plan-level HEDIS scores. However, California plans’ HEDIS scores continue to lag behind national HEDIS ninetieth-percentile benchmarks, and P4P has not helped close this gap. Although it was not an explicit goal of the program in its first three years, health plans were unanimous that another success marker—net savings accrued—was not met, because there has been “no evidence of any savings or moderation in cost trends to justify increased investment in the program.” Purchasers provided comparable success ratings (mean score 2.5) and echoed health plans’ sentiments, flagging as weaknesses the lack of efficiency measures, the small incentive pool, and the limited number of clinical effectiveness measures. Although shortcomings were noted, purchasers and plans acknowledged multistakeholder collaboration, a push toward QI and IT investments, and greater focus on accountability and transparency as program strengths.

■ **Changes required moving forward.** Engagement in the IHA P4P program remains high, yet all stakeholders felt that modifications were required to make the program more effective in closing the quality gap. Health plan and purchaser leaders suggested (1) withholding a portion of any growth in capitation rates at risk for performance to fund a larger incentive pool and create increased commitment for physician organizations; (2) changing the imbalance of large groups relative to smaller groups with lower performance in influencing program design; and (3) shifting away from the current relative-performance payment structures, which creates disparities in payouts and leaves physician organizations that make the greatest changes or need QI resources the most with little or no rewards. Although the lowest-performing physician organizations made some of the largest year-to-year gains, health plans emphasized that P4P had not enabled the lowest performers to close the quality gap and recommended reallocating

resources to achieve a more sizable payoff among low performers, such as directing a portion of incentive funds to targeted QI interventions or incorporating year-to-year improvement in a more substantial way into the compensation formula.¹⁰ The issues identified by the IHA stakeholders are consistent with findings from other studies regarding the implications of various P4P design elements.¹¹

Exhibit 4 reveals sharp differences in stakeholders' views about future priorities. Physician organizations and purchasers gave strong support for increasing the incentive amount, but plans rated this as a low priority because of the marginal results attained thus far and questions about what other types of investments might yield larger improvements. Health plans and purchasers are proposing to expand the measures to assess care more comprehensively, drive improvements more broadly, avoid "teaching to the test," and include cost-efficiency measures to address the issue of rising health care costs. There is also a desire to reexamine payment structures to gain more traction, potentially by reallocating incentive resources and directing support to-

ward physician organizations that are lagging. In contrast, physician organizations did not support measure expansion, feeling that moving beyond the range of ten to fifteen measures would dilute current QI efforts and increase organizations' administrative costs.

Discussion

Three years into the experiment, California physician organizations had made changes in response to P4P; however, these changes did not translate into the breakthrough improvements in quality desired by plans and purchasers. The initial structure of the IHA P4P program did not result in all groups' raising their performance to high levels, and a considerable gap persists between groups in the bottom and top quartiles of performance on most measures.

The inability to close the performance gap highlights the need to adjust the current P4P program design to maximize its effectiveness and to consider additional supporting strategies. A great deal of resources and energy have been directed at "chasing down the data" to demonstrate performance and investing in IT for care management, with only marginal ben-

EXHIBIT 4
Stakeholders' Ratings Regarding The Integrated Healthcare Association (IHA) Pay-For-Performance (P4P) Program Priority Areas Moving Forward, 2007

Priority areas	Mean score (1-5)		
	POs (n = 35)	Health plans (n = 7)	Purchasers (n = 2)
Increasing incentive amount	4.2	2.0	5.0
Expanding clinical measures set	2.5	3.9	4.5
Providing technical assistance on how to improve	3.3	3.7	2.0
Retiring measures that have topped out	2.6	4.2	1.0
Adding specialty care measures	3.2	4.2	3.5
Setting uniform measures for all health plans	4.6	3.8	3.0
Aligning IHA measures with national measures	3.7	3.8	2.0
Adding efficiency measures	2.9	4.7	4.5
Adding in other measures used by plans	2.5	3.0	2.0
Expanding to include Medicaid	2.1	2.0	2.0
Expanding to include Medicare risk	4.3	2.8	4.5
Expanding to include PPO business	2.9	2.8	2.5

SOURCE: Data collected by the authors during interviews with stakeholder group leaders, 2007.

NOTES: On the 1-5 rating scale, 1 = low priority and 5 = high priority. PO is physician organization. PPO is preferred provider organization.

efits; time will reveal whether organizations' IT investment will translate into larger improvements in care delivery.

■ **Challenges.** A lingering question for the IHA program sponsors, which also applies to most current P4P experiments, is whether the size of the incentive was insufficient to fully garner the engagement of the physician organizations and their physicians and to stimulate fundamental changes in the way they provide care.¹² Current incentive levels affect the average physician organization's reimbursement at the margins. The near-term challenge for P4P program design is how to create a sufficiently strong incentive within current payment systems to achieve a tipping point in behavior change without creating adverse behavioral responses. Absent more fundamental payment reforms that foster wholesale reengineering of the care delivery process, the toxicity of the existing payment system may prevent weakly powered incentives from overcoming the Sisyphean task of QI.

Physician organizations face a number of challenges that impede their ability to move forward on quality, including difficulties in engaging and changing the behavior of front-line physicians (particularly if an organization doesn't "own" a substantial fraction of the physician's practice) and lack of knowledge of what specific improvement actions physicians should take. Meanwhile, purchasers and providers are challenged regarding how to gauge the opportunity costs of investing in P4P versus elsewhere, and how to expand performance accountabilities given limitations of current performance measure sets and the data needed to generate measures. Greater investment in developing a more robust set of performance measures is a linchpin for broader accountability efforts.

■ **Caveats.** The strength of this study is that it elicits the experiences of those engaged in the largest P4P experiment in the United States and sheds light on critical issues that must be addressed if P4P is to prove a useful tool in the larger QI toolbox. Because our study is based on a small, nonrandom sample of systematically identified physician organi-

zations, their experiences reflect the market context and structure of the IHA P4P program and thus might not generalize to other settings. The data also are based on subjective self-assessments.

■ **Policy implications.** As the CMS positions itself to engage in paying for performance through its value-based purchasing initiatives, and as other early adopters of P4P consider next steps in their program evolution, this study's findings reveal important lessons and challenges that must be considered.¹³

Paramount is the need to evaluate the impact of programmatic refinements in demonstrating the value of P4P, as we move into second-generation P4P program designs. If the strength of incentives is increased, will this lead to better results or adverse consequences? Does paying for improvement lead to greater improvements among the lowest performers? How does the number of measures affect physician engagement or influence the likelihood of broader system improvements? Collectively, do these efforts to strengthen P4P program design lead to larger performance gains?

The true potential of P4P has not been fully addressed by the first round of experiments, which generally had weak incentives and a limited number and scope of measures, and focused on areas that had been previously the target of measurement and QI interventions.¹⁴ Although Premier reported gains of 15.8 percent in the average composite clinical quality score during the first three years of the CMS/Premier Hospital Quality Incentive Demonstration, P4P programs targeting physicians have seen more modest and sometimes mixed results on performance.¹⁵ The authors of these studies point to an array of design issues that require more careful consideration to address initial shortcomings of programs, including the target of the incentive, the type and mix of measures, and the structure and size of the incentive.

.....
An earlier version of this paper was presented at the annual meeting of AcademyHealth, June 2007, in Washington, D.C.. The authors are grateful for the

generous financial support provided by the California HealthCare Foundation under the Rewarding Results Initiative, which made this work possible. In addition to her RAND work, Cheryl Damberg serves as director of research for the Pacific Business Group on Health and is responsible for the analytic work related to the California Cooperative Healthcare Reporting Initiative's annual Patient Assessment Survey and the California Physician Performance Initiative. The authors thank Thomas Williams and Dolores Yanagihara at the Integrated Healthcare Association (IHA) for their assistance; the IHA physician organization, plan, and purchaser stakeholders for their participation in the independent evaluation of the IHA program; and Lily Bradley and Magdalen Paskell at RAND for their assistance with data entry, analysis, and interview scheduling.

NOTES

1. S. Trude, M. Au, and J.B. Christianson, "Health Plan Pay-for-Performance Strategies," *American Journal of Managed Care* 12, no. 9 (2006): 537-542; G. Baker and B. Carter, *Provider Pay-for-Performance Programs: 2004 National Study Results* (San Francisco: Medvantage, 2005); and M.B. Rosenthal et al., "Climbing Up the Pay-for-Performance Learning Curve: Where Are the Early Adopters Now?" *Health Affairs* 26, no. 6 (2007): 1674-1682.
2. Institute of Medicine, *Rewarding Provider Performance: Aligning Incentives in Medicare* (Washington: National Academies Press, 2007); Medicare Payment Advisory Commission, *Report to the Congress: Medicare Payment Policy* (Washington: MedPAC, 2005); C.L. Damberg et al., *An Environmental Scan of Pay for Performance in the Hospital Setting: Final Report*, November 2007, <http://aspe.hhs.gov/health/reports/08/payperform/PayPerform07.pdf> (accessed 12 December 2008); M.B. Rosenthal et al., "Early Experience with Pay-for-Performance: From Concept to Practice," *Journal of the American Medical Association* 294, no. 14 (2005): 1788-1793; and L.A. Petersen et al., "Does Pay-for-Performance Improve the Quality of Health Care?" *Annals of Internal Medicine* 145, no. 4 (2006): 265-272.
3. C. Predergast, "The Provision of Incentives in Firms," *Journal of Economic Literature* 35, no. 1 (1999): 7-63; and B. Asch and J. Warner, "Incentive Systems: Theory and Evidence," in *Handbook of Human Resource Management*, ed. D. Lewin, D. Mitchell, and M. Zaidi (Stamford, Conn.: JAI Press, 1996), 175-215.
4. *Medicare Improvements for Patients and Providers Act (MIPPA)* of 2008, PL 110-275, sec. 131, 15 July 2008; and U.S. Department of Health and Human Services, *Report to Congress: Plan to Implement a Medicare Hospital Value-Based Purchasing Program* (Washington: DHHS, 21 November 2007).
5. C.L. Damberg et al., "Paying for Performance: Implementing a Statewide Project in California," *Quality Management in Health Care* 14, no. 2 (2005): 66-79.
6. Wells Shoemaker, medical director, California Association of Physician Groups, personal communication, June 2008.
7. A total of twenty-six physician organizations that contracted with one or more of the participating HMOs did not meet minimum encounter data submission thresholds or minimum denominator sizes for clinical measures to qualify for scoring.
8. Competition is greater in some regions of California, which results in lower capitation rates paid in those regions as compared to regions exhibiting more consolidation of physician organizations.
9. Based on internal IHA data, administrative-only rates increased from 38.06 in 2002 to 57.42 in 2005, an increase of 19.36 percentage points.
10. Starting with measurement year 2006 (payout year 2007), the IHA recommended that plans include year-to-year improvement in their payout formulas; however, for measurement year 2006, less than 4 percent of the total payout was for improvement.
11. M.B. Rosenthal and R.A. Dudley, "Pay-for-Performance: Will the Latest Payment Trend Improve Care?" *Journal of the American Medical Association* 297, no. 7 (2007): 740-744. See Petersen, "Does Pay-for-Performance Improve the Quality of Health Care?"; G.J. Young et al., "Effects of Paying Physicians Based on Their Relative Performance for Quality," *Journal of General Internal Medicine* 22, no. 6 (2007): 872-876; and S.D. Pearson et al., "The Impact of Pay-for-Performance on Health Care Quality in Massachusetts, 2001-2003," *Health Affairs* 27, no. 4 (2008): 1167-1176.
12. M.B. Rosenthal, "Beyond Pay for Performance—Emerging Models of Provider Payment Reform," *New England Journal of Medicine* 359, no. 12 (2008): 1197-1200; and Young, "Effects of Paying Physicians."
13. See Note 4.
14. See Note 11.
15. Premier, "Patient Lives Saved as Performance Continues to Improve in CMS, Premier Healthcare Alliance Pay-for-Performance Project," Press Release, 17 June 2008, <http://www.premierinc.com/about/news/june08/p4pProject061708.jsp> (accessed 22 September 2008).

MARKET WATCH

Choosing The Best Hospital: The Limitations Of Public Quality Reporting

Despite multiple and often conflicting ratings, public reporting on hospital quality seems to have gained a permanent foothold.

by **Michael B. Rothberg, Elizabeth Morsi, Evan M. Benjamin, Penelope S. Pekow, and Peter K. Lindenauer**

ABSTRACT: The call for accountability in health care quality has fueled the development of consumer-oriented Web sites that provide hospital ratings. Taking the consumer perspective, we compared five Web sites to assess the level of agreement in their rankings of local hospitals for four diagnoses. The sites assessed different measures of structure, process, and outcomes and did not use consistent patient definitions or reporting periods. Consequently, they failed to agree on hospital rankings within any diagnosis, even when using the same metric (such as mortality). In their current state, rating services appear likely to confuse, rather than inform, consumers. [*Health Affairs* 27, no. 6 (2008): 1680–1687; 10.1377/hlthaff.27.6.1680]

DRIVEN BY A DESIRE TO increase the transparency and accountability of the U.S. health care system, hospitals face growing requirements to release performance data to the public. For more than a century, however, public reporting has been controversial, frequently criticized for being unreliable and for failing to capture the true complexity of patients' conditions.¹

Despite these concerns, in the face of rapidly rising costs and striking gaps in the quality of health care, consumers are being asked to make better-informed health care choices, with the expectation that market forces will drive down prices and improve quality. High-deductible insurance plans and health savings

accounts (HSAs) are predicated on this idea, despite a lack of evidence that health care consumers make choices based on performance data.² Even if consumers do not use the performance data, their public release appears to stimulate quality improvement initiatives by hospitals, perhaps by appealing to the professional ethos of clinicians, and out of fear that poor performance would threaten the reputation of the organization.³

Over the past decade, a number of hospital rating services have emerged, offering free Web sites aimed at health care consumers. The services report on multiple aspects of health care quality, including those reflecting structural aspects of care, processes, and outcomes.

Michael Rothberg (Michael.Rothberg@bhs.org) is an assistant professor of medicine at Tufts University School of Medicine, Baystate Medical Center, in Springfield, Massachusetts. Elizabeth Morsi is a research assistant in the medical center, and Evan Benjamin is vice president for quality. Penelope Pekow is director of the Center for Research and Education in Women's Health at the University of Massachusetts School of Public Health in Amherst. Peter Lindenauer is an associate professor at the Tufts University School of Medicine.

For consumers to make good decisions, however, publicly reported information must be accessible, interpretable, and consistent.⁴ In this context, it remains unclear whether consumers can correctly interpret the available data or whether such data would lead to consistent health care decisions. Moreover, the spectrum of performance measurement has been limited to few diseases, so most patients will not find quality information about their condition. Information about related conditions might be helpful, but only if performance in one disease area predicts performance in another. Taking the perspective of a patient seeking to choose the best hospital, we examined five reporting services and evaluated their level of agreement across hospitals for four conditions.

Study Data And Methods

In August 2007 we identified all hospitals within a thirty-mile radius of Boston, Massachusetts. We focused our analysis on hospitals competing for consumers who would be willing to travel up to one hour to receive high-quality care. To create a manageable sample, we limited our analysis to nonspecialty hospitals with at least 250 beds, because we did not believe that consumers would travel to receive care in a community hospital.⁵

We identified five well-known public reporting services: Hospital Compare, HealthGrades, Leapfrog Group, *U.S. News and World Report*, and Massachusetts Healthcare Quality and Cost (Mass QC, a state-run service sponsored by the state's executive office of health and human services). First, we compared the services based on the number of diagnoses or procedures reported; the use of structure, process, and outcome measures; the reporting time period; the patient population used to derive the ratings; and the method of presentation.

We then selected four nonemergent conditions to compare across the databases: community-acquired pneumonia, total hip replacement, percutaneous coronary intervention (PCI), and coronary artery bypass grafting (CABG). In a separate analysis, we com-

pared adjusted mortality for acute myocardial infarction, because three of the five services reported this measure. For each condition, we ranked all hospitals within each database, using the measures provided by that database. We evaluated association among pairs of raters for the same diagnosis via Spearman rank correlations of hospital scores; Spearman correlations were also used to evaluate the association of ratings for pairs of diagnoses within each hospital.

Study Results

■ Overview of hospital rating services.

The number of conditions reported on by the rating services ranged from four to thirty-two, with the mean and median both fourteen (Exhibit 1). Of the three types of quality measures—structure, process, and outcomes—most of the rating services focused on only one, although some reported measures in two spheres. Only Leapfrog reported measures in all three. Three of the five rating services reported outcome measures, all of which included risk-adjusted mortality. Although we examined all services on the same date, the time period on which the rankings were based varied in both duration and start year. There was even variation in time periods within the same rating system: *U.S. News and World Report* based its 2007 ratings on mortality data for fiscal years 2003–2005 and physician survey data for calendar years 2005–2007. Four of the services used at least some data that were two to four years old. Three services limited their data to Medicare or Medicaid patients. Finally, three services incorporated reports from other services: HealthGrades reported some Leapfrog measures, Mass QC reported Hospital Compare measures, and Leapfrog linked to Hospital Compare.

■ **Hospital rankings.** Of twenty-five general hospitals within thirty miles of Boston, nine had at least 250 beds. For any given diagnosis, there was little overall agreement among the rating services.⁶ For example, the two hospitals ranked first for CABG by at least one service were also ranked fourth and last by another (Exhibit 2). Conversely, the two hospi-

EXHIBIT 1
Comparisons Of Five Hospital Quality Rating Services: Summary

Rating service	No. of conditions scored	Structure measures	Process measures	Outcome measures	Time period	Patient population
HealthGrades	32	Presence of newborn ICU	Leapfrog safe-practices score	Mortality, major complications, volume of vaginal and c-section deliveries, newborn mortality, women's health rating	2003–2005	Medicaid, Medicare
Leapfrog	8	CPOE, ICU staffing	Safe-practices score, procedure volumes	PCI, CABG, and AAA mortality	2007 (mortality: 2006)	All patients
Hospital Compare	4	None	Antibiotic use in surgery, 5 measures for pneumonia, 8 measures for heart attack, 4 measures for heart failure	Adjusted mortality	2005–2006	Medicaid, Medicare
Mass QC	14	None	Provided recommended care	Cost, risk-adjusted mortality, length-of-stay, total c-sections and VBAC utilization	FY 2004 and FY 2005	All patients
U.S. News	16	Advanced services, trauma center, magnet hospital, epilepsy center certification, FACT accreditation, nursing ratio, patient service index, NCI cancer center, total volume	Reputation based on physicians surveyed	Risk-adjusted mortality, total volume	2005–2007 survey (mortality: FY 2003–2005)	Medicare

SOURCE: Results derived by authors from providers' Web sites.

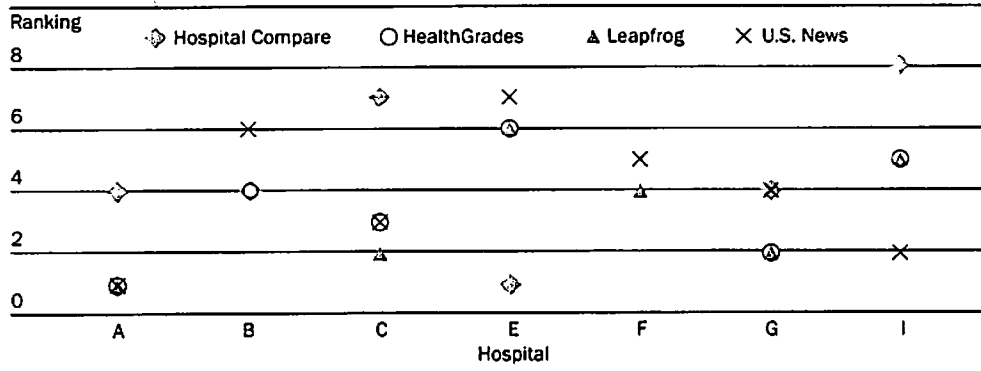
NOTES: ICU is intensive care unit. CPOE is computerized physician order entry. PCI is percutaneous coronary intervention. CABG is coronary artery bypass graft. AAA is abdominal aortic aneurysm. VBAC is vaginal birth after caesarean. FACT is Foundation for the Accreditation of Cellular Therapy. NCI is National Cancer Institute. Additional information and more detailed definitions are available in Appendix Exhibit 1, online at <http://content.healthaffairs.org/cgi/content/full/27/6/1680/DC1>.

tals that were ranked last for CABG by at least one service were ranked first or second by another. Spearman rank correlation among pairs of services within a diagnosis was lowest for Mass QC and Leapfrog rating of PCI ($r = -0.67$, $p = 0.14$), and best for HealthGrades and Leapfrog ratings of both PCI and CABG ($r = 0.97$, $p < 0.005$). The strong agreement between Leapfrog and HealthGrades was not seen for total hip replacement ($r = -0.41$, $p = 0.42$).

Within a given rating service, there was little

agreement among the ratings for a single hospital across pairs of diagnoses, except for Leapfrog and Hospital Compare. Leapfrog's mostly structural measures tended to be the same for each condition ($r = 0.97$ to $r = 1.00$, $p < 0.001$); Hospital Compare offered a single measure—preoperative antibiotic timing—for all surgical procedures, so its surgical rankings were perfectly correlated ($r = 1.00$, $p < 0.0001$), but none of these was correlated with community-acquired pneumonia ($r = 0.14$, $p = 0.73$).

EXHIBIT 2
Seven Hospitals' Rankings For Coronary Artery Bypass Grafting (CABG) Based On Publicly Reported Ratings Using Four Different Hospital Quality Rating Systems



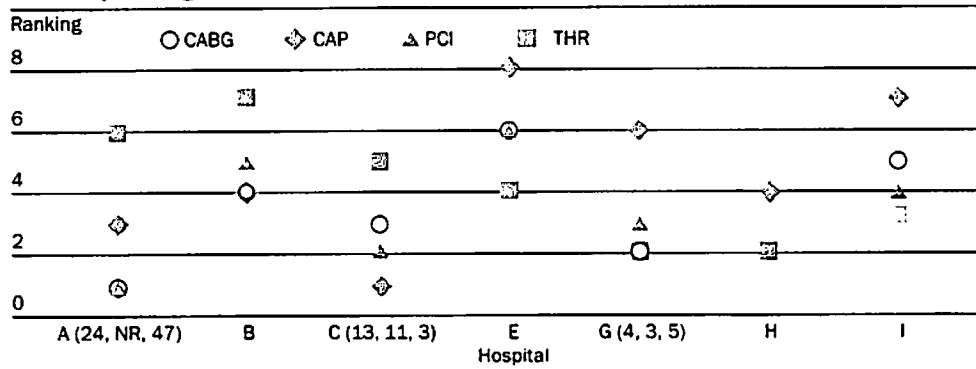
SOURCE: Authors' computations based on providers' Web sites.

HealthGrades showed a strong mortality correlation between CABG and PCI ($r = 0.89, p = 0.019$) but not between CABG and total hip replacement ($r = -0.14, p = 0.79$). No hospital scored in either the top or bottom half on all four diagnoses (Exhibit 3).

■ **Mortality rates.** Two services provided mortality rates for all of the diagnoses, and three services provided mortality rates for AMI (Exhibit 4). Neither the observed mortality rates nor the observed/predicted rates were

consistent across services. The hospital ranked first by HealthGrades had the second-highest observed mortality; it ranked seventh according to Mass QC. Conversely, HealthGrades' seventh-ranked hospital (the only hospital statistically worse than average) was ranked first by Mass QC. This same hospital was ranked fifth in the nation by *U.S. News and World Report* for cardiology.

EXHIBIT 3
Seven Hospitals' Rankings For Four Different Conditions Based On Publicly Reported Mortality Ratings From HealthGrades



SOURCE: Authors' computations based on providers' Web sites.

NOTES: Hip replacement rankings are based on complication rates instead of mortality. Numbers indicated the hospital rank for respiratory disorders, orthopedics, and heart by *US News and World Report*. CABG is coronary artery bypass grafting. CAP is community-acquired pneumonia. PCI is percutaneous coronary intervention. THR is total hip replacement. NR is not ranked.

EXHIBIT 4
Acute Myocardial Infarction Mortality And Ratings From Three Hospital Quality Rating Services

Hospital	Rating service								Hospital Compare Rating
	HealthGrades				Mass QC				
	Obs.	Pred.	Ranking ^a	Rating ^b	Obs.	Pred.	Ranking	Rating	
A	6.61	7.16	2	**	5.45	8.11	2	**	**
B	7.95	7.14	5	**	6.64	9.54	3	**	**
C	6.53	6.35	3	**	5.57	7.93	4	**	**
D	20.51	18.85	4	**	15.56	14.65	9	**	**
E	9.26	7.81	8	**	10.47	12.06	5	**	**
F	-c	-c	-d	-c	9.5	8.99	8	**	**
G	7.83	6.82	7	*	5.57	9.18	1	**	**
H	19.46	22.27	1	**	17.65	20.25	7	**	**
I	7.56	6.79	5	**	6.54	7.53	6	**	**

SOURCE: Providers' Web sites.

^aRanking is based on observed/predicted mortality.

^bRating compares the adjusted mortality to the mean. Two stars mean not statistically different from average, and one star means worse than average. Three stars (not pictured) means better than average.

^cNot available.

^dNot applicable.

Discussion

■ **Lack of consistent agreement.** In an environment of increasing transparency, in which patients are taking an active role in choosing health care providers and assuming more financial responsibility, accurate ratings of hospitals are essential. In this study of five leading health care rating services, we found that these services failed to consistently agree on either top- or bottom-performing hospitals in a single metropolitan area. Among the nine institutions studied, hospitals ranked first or second by one system were often ranked seventh or eighth by another. These findings can be explained by variations in the methods used by the rating systems, which measured different processes and outcomes, examined widely different time periods, and included different patient populations. Even when two services simultaneously measured the same outcome (mortality), agreement among the services was poor. Most rating systems did not perform statistical tests, but when they did, all nine hospitals were indistinguishable. In fact, only one hospital (out of seventy-one) in the state

had cardiac mortality that was statistically better than the mean.

■ **Previous studies of rating services.** Studies of rating services have generally been limited to two-way comparisons and have included hospitals nationwide. *U.S. News and World Report's* "Best Hospitals" has been compared to Hospital Compare, the Cooperative Cardiovascular Project, and risk-standardized mortality rates.⁷ "Best Hospitals" as a group usually performed better on both process measures and mortality, although some individual "Best Hospitals" performed poorly, so a "Best Hospital" rating does not ensure quality. Our sample included several "Best Hospitals," none of which performed consistently well. Comparing HealthGrades to the Cooperative Cardiovascular Project produced similar conclusions.⁸ The process measures from Hospital Compare are inversely related, albeit weakly, to standardized mortality rates.⁹ We found that not only were mortality rankings poorly correlated with process measures, but they were poorly correlated with other rating services' mortality rankings.

■ **Study limitations.** Our study has a

number of limitations. First, data were collected from large hospitals located in a single metropolitan area with a high concentration of hospitals. Most health care consumers do not have as much choice, although today patients are encouraged to move beyond traditional geographic boundaries in search of better care. In this context, ratings are even more essential. Second, our data are limited to a single point in time, while health care rating systems are in rapid continuous evolution.

However, our data are representative of the information available to the public in 2007. Third, we used unweighted rankings instead of absolute scores. Adding weights might produce different rankings, but the rating services themselves did not identify any measure as more important than another, nor does the current science support using any particular measure over another. Finally, we did not account for the magnitude of differences between hospitals, and we do not know whether patients would take this into account. For this reason, we also examined mortality measures for which there was statistical testing and found that the systems generally could not discriminate among hospitals, demonstrating the limitations of such measures.

■ **Barriers patients face.** Public reporting has the potential to be a powerful agent for quality improvement. Patients maintain that they would use mortality data in making decisions; yet despite the wide availability of mortality reports, few patients actually use them, and even the best public reports do not seem to affect market share or consumers' choices.¹⁰ This may change as consumers become more aware of the rating services and as high-deductible plans drive patients to seek care beyond their local hospital. Nevertheless, the existence of multiple and often conflicting reports can only further complicate the relationship between public reporting and consumer choice.

We identified several additional barriers patients face. First, to rank the hospitals, it is

necessary to interpret the data, but the services offer no guidance as to the relative benefit of specific measures. For example, when considering pneumonia, is it more important to receive antibiotics within four hours or to receive the correct antibiotic? In cardiac surgery, is it more important to choose a high-volume hospital or one that has an electronic health record and a high safe-practices score? Second, even seemingly straightforward measures such as mortality rates vary by rating service, because there is no standardized risk-adjustment methodology.¹¹

■ **Recommendations.**

Despite these limitations, public reporting seems to have gained a permanent foothold; it stimulates quality improvement activities, although its effects on outcomes are mixed.¹² Unfortunately, proliferation of conflicting hospital rating services may dilute the impact of low scores, because poor performers on one service's measures can still claim to be rated as top performers by another.¹³

How, then, might hospitals and reporting sites join forces to offer information that is useful, accurate, and consistent? First, instead of passively accepting reporting mandates, hospitals must embrace public reporting and help design measures that can be efficiently collected and that represent the true quality of care offered. The Hospital Quality Alliance (HQA) represents a step in this direction, but its early measures focus on processes that, although laudable, are not of particular interest to patients and have relatively weak links to outcomes.¹⁴ The National Surgical Quality Improvement Program (NSQIP) is another example of hospitals' taking the initiative to measure and share outcome data for the purposes of improving quality.¹⁵ Although these measures are not currently available to the public, in the future they could be.

Patients, however, may want different information. Qualitative studies suggest that patients want to know about other patients' experiences and that they are willing to choose

"Even the best public reports do not seem to affect market share or consumers' choices."

physicians based on that information.¹⁶ Patients also care about patient satisfaction, being cared for, and having their physical and informational needs met.¹⁷ Some of these concerns are addressed through the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, which was recently added to Hospital Compare. Alternatively, rapidly growing social networking Web sites may eclipse traditional public reporting for both physician and hospital rating.¹⁸ More research is needed, on both what patients want and what they can understand. In the future, reporting mandates should incorporate new quality assessments that are meaningful to patients; these should focus on both disease-specific measures (for example, functional status following stroke) and hospital-specific measures (for example, rates of hospital-acquired infections). Such efforts would benefit from the participation of providers and consumer advocacy groups.

Finally, rating services must solve the problems of risk adjustment and random variation. Standardizing risk-adjustment methodology and reporting periods would aid in transparency and level the playing field for providers; also, allowing consumers to compare two hospitals directly might reveal statistical differences that are obscured when comparing each hospital to the national or state mean.

OUR STUDY demonstrates that sometimes more is less. Providing more measures may actually obscure comparisons, because consumers may be overwhelmed by confusing and conflicting information. If public reporting is to succeed in improving health care, hospitals should join with raters to provide stakeholders with data that are consistent and easily interpretable. As a first step, many of the rating services now include other services' data on their Web sites, providing a range of measures of structure, process, and outcomes. Still missing is the means to prioritize and synthesize these measures. Until this is achieved, public reporting may remain of limited value to those whom it was designed to help.

.....
 Michael Rothberg is the recipient of a Doris Duke Clinical Scientist Development Award.

NOTES

1. F. Nightingale, *Notes on Hospitals*, 3d ed. (London: Longman Green, 1863); L.K. Altman, "Report on Mortality: Guarded Praise," *New York Times*, 18 December 1987; and E.C. Schneider and A.M. Epstein, "Use of Public Performance Reports: A Survey of Patients Undergoing Cardiac Surgery," *Journal of the American Medical Association* 279, no. 20 (1998): 1638-1642.
2. J.H. Hibbard, J. Stockard, and M. Tusler, "Hospital Performance Reports: Impact on Quality, Market Share, and Reputation," *Health Affairs* 24, no. 4 (2005): 1150-1160; and M.N. Marshall et al., "The Public Release of Performance Data: What Do We Expect to Gain? A Review of the Evidence," *Journal of the American Medical Association* 283, no. 14 (2000): 1866-1874.
3. *Ibid.*
4. Hibbard et al., "Hospital Performance Reports"; and L.M. Schwartz, S. Woloshin, and J.D. Birkmeyer, "How Do Elderly Patients Decide Where to Go for Major Surgery? Telephone Interview Survey," *British Medical Journal* 331, no. 7520 (2005): 821.
5. A more detailed description of the methods appears in the online Technical Appendix, <http://content.healthaffairs.org/cgi/content/full/27/6/1680/DC1>.
6. The rankings of the hospitals for each of the four conditions studied appear in Appendix Exhibit 2; *ibid.*
7. L.K. Halasyamani and M.M. Davis, "Conflicting Measures of Hospital Quality: Ratings from 'Hospital Compare' versus 'Best Hospitals,'" *Journal of Hospital Medicine* 2, no. 3 (2007): 128-134; J. Chen et al., "Do 'America's Best Hospitals' Perform Better for Acute Myocardial Infarction?" *New England Journal of Medicine* 340, no. 4 (1999): 286-292; and O.J. Wang et al., "'America's Best Hospitals' in the Treatment of Acute Myocardial Infarction," *Archives of Internal Medicine* 167, no. 13 (2007): 1345-1351.
8. H.M. Krumholz et al., "Evaluation of a Consumer-Oriented Internet Health Care Report Card: The Risk of Quality Ratings Based on Mortality Data," *Journal of the American Medical Association* 287, no. 10 (2002): 1277-1287.
9. R.M. Werner and E.T. Bradlow, "Relationship between Medicare's Hospital Compare Performance Measures and Mortality Rates," *Journal of the American Medical Association* 296, no. 22 (2006): 2694-2702; and A.K. Jha et al., "The Inverse Rela-

- tionship between Mortality Rates and Performance in the Hospital Quality Alliance Measures," *Health Affairs* 26, no. 4 (2007): 1104-1110.
10. Schneider et al., "Use of Public Performance Reports"; Hibbard et al., "Hospital Performance Reports"; Romano et al., "Do Well-Publicized Risk-Adjusted Outcomes Reports Affect Hospital Volume?"; Schwartz et al., "How Do Elderly Patients Decide Where to Go?"; and D.W. Baker et al., "The Effect of Publicly Reporting Hospital Performance on Market Share and Risk-Adjusted Mortality at High-Mortality Hospitals," *Medical Care* 41, no. 6 (2003): 729-740.
 11. L.I. Iezzoni, "The Risks of Risk Adjustment," *Journal of the American Medical Association* 278, no. 19 (1997): 1600-1607; and R. Behal, "The Lake Wobegon Effect: When All the Patients Are Sicker," *American Journal of Medical Quality* 21, no. 6 (2006): 365-366.
 12. C.H. Fung et al., "Systematic Review. The Evidence That Publishing Patient Care Performance Data Improves Quality of Care," *Annals of Internal Medicine* 148, no. 2 (2008): 111-123.
 13. P.J. Pronovost, M. Miller, and R.M. Wachter, "The GAAP in Quality Measurement and Reporting," *Journal of the American Medical Association* 298, no. 15 (2007): 1800-1802.
 14. Werner and Bradlow, "Relationship between Medicare's Hospital Compare Performance Measures and Mortality Rates"; and E.H. Bradley et al., "Hospital Quality for Acute Myocardial Infarction: Correlation among Process Measures and Relationship with Short-Term Mortality," *Journal of the American Medical Association* 296, no. 1 (2006): 72-78.
 15. K.S. Rowell et al., "Use of National Surgical Quality Improvement Data as a Catalyst for Quality Improvement," *Journal of the American College of Surgeons* 204, no. 6 (2007): 1293-1300.
 16. S.A. Richard, S. Rawal, and D.K. Martin, "Patients' Views about Cardiac Report Cards: A Qualitative Study," *Canadian Journal of Cardiology* 21, no. 11 (2005): 943-947; and G. Fanjiang et al., "Providing Patients Web-Based Data to Inform Physician Choice: If You Build It, Will They Come?" *Journal of General Internal Medicine* 22, no. 10 (2007): 1463-1466.
 17. L.V. Doering, A.W. McGuire, and D. Rourke, "Recovering from Cardiac Surgery: What Patients Want You to Know," *American Journal of Critical Care* 11, no. 4 (2002): 333-343.
 18. L. Landro, "Social Networking Comes to Health Care," *Wall Street Journal*, 27 December 2006.