NIH Public Access
Author Manuscript
Health Aff (Millwood). Author manuscript; available in PMC 2010 December 20.

Published in final edited form as:
Health Aff (Millwood). 2009 ; 28(2): 526–532. doi:10.1377/hlthaff.28.2.526.

Measuring The Report Card: The Validity Of Pay-For-Performance Metrics In Orthopedic Surgery:

The current generation of P4P measures based on process is inadequate to measure quality in orthopedic surgery

Timothy Bhattacharyya, **Andrew A. Freiberg**, **Priyesh Mehta**, **Jeffrey Neil Katz**, and **Timothy Ferris**
Timothy Bhattacharyya (tbhattacharyya@bcc-ortho.com) is chief of orthopedic trauma at Suburban Hospital in Bethesda, Maryland. Andrew Freiberg is chief of arthroplasty at the Massachusetts General Hospital in Boston. Priyesh Mehta is a medical student at the University of New England College of Osteopathic Medicine in Biddeford, Maine. Jeffrey Katz is an associate professor in rheumatology at Brigham and Women's Hospital in Boston. Timothy Ferris is medical director at the Mass General Physicians Organization.

Abstract

To assess the validity of performance measures used in a nationwide pay-for-performance (P4P) project on hip and knee replacement, we analyzed hospital performance Data from a medicare P4P initiative and compared them to publicly available outcomes Data. Overall, the ability to measure hospital quality was poor. A hospital's ranking On the composite score was primarily determined by process measures. A higher composite Quality score was not associated with lower rates of complications or mortality. The current Medicare P4P quality measure has limited validity because of poor discrimination, lack of Measure balance, and lack of correlation with important clinical outcomes.

Pay-for-performance (P4P) programs Have been advocated as a method To improve the quality of health care in the United States.[1] The success of P4P programs hinges upon the development of performance measures that accurately reflect the quality of clinical care delivered. Ideally, quality measures should be easily obtainable, discriminate between high- and low-quality Providers, be adjusted for the severity of case mix, and correlate with important external measures such as mortality and complication rates.2 Although numerous P4P programs have been implemented, little is known about how well the quality measures used by these Programs actually perform, especially in the resource-intensive surgical fields. 3 Rigorous assessment of performance measures should Precede expansion of P4P programs.

Suitability of arthroplasty for P4P

P4P programs target total hip and knee replacement because arthroplasty is one of the few well-defined procedures in medicine Where performance can be measured. The surgical procedure is relatively standardized. Existing treatment care guidelines and clinical pathways facilitate the development of clinical performance measures.[4] Outcomes after total

© 2009 Project HOPE–The People-to-People Health Foundation, Inc.

Timothy Bhattacharyya is a consultant for Stryker Orthopaedics and Best Doctors Inc. Andrew Freiberg receives royalties and research support from Zimmer and is a consultant for Biomet.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

hip and total knee arthroplasty are well studied, and the relationship between greater hospital Case volume and improved outcomes has been described. Total joint replacement is common and costly—more than $9 billion per Year in the United States.[5] Together, these factors make arthroplasty an attractive candidate for P4P.

Medicare P4P project

The Centers for Medicare and Medicaid Services (CMS)premier Hospital Quality Initiative Demonstration (HQID) was a voluntary demonstration Project in which hospitals reported data on quality and outcomes.[6] The CMS and Premier Inc., an organization owned by not-for-profit hospitals, collaborated in July 2003 to launch HQID. This became the first national P4P demonstration to examine the relationship between quality and cost.

The HQID included 260 hospitals in 38 states, focusing on five primary clinical areas: acute myocardial infarction (AMI), heart failure, community-acquired pneumonia, coronary artery bypass graft (CABG), and hip and knee surgery. The CMS began collecting data in October 2003 and released its data from Year one in 2005. Through the Deficit Reduction Act (DRA) of 2005, Congress authorized the development of a Medicare hospital value–based purchasing program. On 21 November 2007 the CMS released its data from year two and submitted a plan to Congress to implement P4P on a national scale in 2009. Congress approved the plan and will be responsible for creating a final program.

In the first year of the demonstration project, hospitals that scored in the top 10 percent on the composite quality measure received a Performance bonus consisting of 2 percent of diagnosis-related group (DRG) payments for total hip and knee arthroplasty for the study Year. Hospitals in the second decile of the composite quality measure received a 1 percent DRG bonus. All hospitals scoring in the top 50 Percent of performance were publicly recognized On the HQID Web site.

Medicare quality measure

The CMS has proposed one method to measure the quality Of hip and knee replacement at the hospital Level.[7] It includes a composite score created from three measures of surgical process quality and three measures of surgical outcome. The composite score was used in the HQID.

Study Methods

Hospital performance assessment

Using publicly available data, we performed a Cross-sectional analysis of hospitals participating in the hip and knee segment of the HQID and assessed the validity of the CMS quality scores with clinically important outcomes.[8] We assessed hospital performance in two ways: hospital performance tier and composite quality index. First, we identified four Ordinal tiers of hospitals. Tier 1 hospitals were in the top 10 percent of performance, tier 2 hospitals were in the second decile, tier 3 hospitals were in the top 50 percent but not in the top two deciles, and tier 4 hospitals were in the bottom 50 percent. Second, we calculated the hip and knee composite quality index according to CMS guidelines.

Correlating quality measures with outcomes

We obtained three external measures of hospital performance. First, we collected data on inpatient mortality after hip and knee arthroplasty. Second, as a measure of complications, we used “iatrogenic complications” (physician-caused complications) and “urinary tract infection” (UTI), which are risk-adjusted, Validated measures in surgical patients.[9] Third, we compared the performance measures to surgical volume (total number of total hip and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

knee arthroplasties performed at a hospital in the study year), because volume is known to correlate with quality.10

Study Results

The sample included both teaching and nonteaching hospitals from all U.S. geographic areas, with a range of procedure volumes.[11]

Analysis of quality measures

Poor performance was noted for the three outcome measures (metabolic derangement index, hematoma index, and readmission avoidance index). The measures showed no variance and thus contributed little to interhospital differences in composite quality scores (Exhibit 1). The standard deviations ranged from 0.000036 to 0.0272. The composite quality score was primarily determined by performance on the three surgical process measures (related to antibiotics). Thus, it was not balanced; it ignored outcome and measured only process.

Distribution of composite quality scores

Seventy-four percent of hospitals were within 10 percent of the mean composite quality score.[12] We noted that classifying the hospitals into performance tiers required calculation Of the composite quality score to the fourth decimal place because numerous hospitals had essentially the same score. The low variance of composite quality measure demonstrates that the measure has poor ability to discriminate among hospitals.

A strong ceiling effect was observed—that is, a large chunk of the class received “A” grades. For the metabolic derangement index and the hematoma index, 100 percent of hospitals were at 99 percent or above. For the readmission index, 81 percent of hospitals were at 99 percent or above (Exhibit 2). Performance on the antibiotic process measures primarily differentiated hospitals in the top two deciles from the remaining hospitals.

Quality scores, volume, and outcomes

Both hospital quality scores and hospital performance tiers correlated moderately with surgical volume ($r = 0.268$; $p < 0.001$). Thus, higher-volume hospitals tended to have higher quality scores, but the correlation was not strong. No hospitals with more than 400 cases per year were in the lower half of hospital Performance measures (Exhibit 3).

Higher-tier hospitals did not have lower complications (Exhibit 4).[13] This is true for complication measures such as iatrogenic complications and UTI. Although there was no significant difference in mortality associated with hip and knee arthroplasty across the hospital tiers, there was a trend toward a higher rate of mortality in tier 4 hospitals ($r = 0.116$; $p = 0.088$). All hospitals with mortality greater than 2.0 percent were in tiers 3 and 4.

Discussion

Limitations of the quality measures

We observed that the quality measures adopted by the CMS have marked limitations. The “top-performing hospitals” were entirely distinguished by their performance on the process measures for antibiotic administration. The limited distribution of the scores requires calculation of the scores to the level of the fourth decimal place to separate the hospitals into deciles. Furthermore, the quality measure did not correlate with important outcomes such as complication rates and mortality.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

P4P programs have a laudable goal of linking reimbursement to quality of care, yet surgical quality is difficult to measure. Proper performance Of a total hip replacement requires satisfactory completion of several disparate steps, including proper patient selection and preoperative planning, safe anesthesia, surgical preparation to prevent infection, surgical approach with minimal trauma, proper technical positioning of acetabular and femoral components, controlled bleeding, appropriate rehabilitation, and prevention of complications such as thromboembolic disease. Because the numerous technical steps are difficult to evaluate and centrally report, attempts have been made to judge quality based on adherence to several discrete agreed-upon measures.

The rates of administration of antibiotics within one hour before total hip or knee replacement and the discontinuation of antibiotics twenty-four hours after surgery are clinical measures that are accepted by the National Quality Forum (NQF). As process measures, they have face validity. However, other performance characteristics for measures have not been determined. For example, it is unknown whether the “antibiotics administered within one hour before surgery” measure has good distribution and minimal ceiling effects, which would allow accurate measurement of high and low performance. These data are only now becoming available as P4P programs evolve from concept to practice.[14] The HQID does have some methodology to identify reasonable clinical variation from the standard of care and to identify patients for whom these measures might not be clinically appropriate. For example, patients who were admitted with preexisting UTI are allowed to receive antibiotics for longer than twenty-four hours. Finally, the evidence correlating improved antibiotic delivery compliance to improved clinical outcomes is exceedingly limited.15

Growing literature

There is a growing literature concerning practical and statistical problems in the nascent field of hospital performance measurement.[16] Elizabeth Bradley and colleagues found that the process measures used to assess treatment of AMI were only moderately correlated with the standardized thirty-day mortality rates and explained a small amount of hospital variation in mortality rates.[17] Furthermore, they found that mortality rates—arguably a more valid measure than Process measures of quality of care—varied by only a few percentage points between the highest- and lowest-quality hospitals. Peter Lindenaur noted that P4P resulted in modest gains in quality measures compared to public reporting alone.[18]

Care process measures versus outcome measures

The HQID was designed to measure a broad concept (the quality of hip and knee replacement), yet we found that the outcome measures used by the CMS did not perform well. Because postoperative hematoma or readmission after total joint replacement are rare events, and severity adjustment is necessary, these measures ultimately reveal little variability across hospitals. Because of Ceiling effects and small measure variances, the HQID ignores outcome measures and ends up measuring a narrow number of care process measures that pertain only to antibiotic administration. Although there is some range in hospital performance in these measures, 74 percent of hospitals are within 10 percent of the mean performance. Subsequently, the antibiotic measures are not good at truly differentiating high performers from the average. The administrative cost of collecting, analyzing, and reporting these data has not been reported.

Measuring a hospital’s ability to follow national guidelines on antibiotic Administration is not the Same as measuring the quality Of total joint replacement. Patients and payers desire joint replacement surgery that is done in a technically optimal manner with the lowest possible rate of major complications such as infection and hip dislocation. Antibiotic administration rates simply do not relate strongly to the quality of total joint replacement.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Subsequently, the finding that the top tier of hospitals does not have lower mortality or complication rates is not surprising. The situation is analogous to trying to measure the quality of a restaurant by only measuring how fast they take your order—a valid process measure that ignores dozens of important aspects such as service and food quality.

A good measure for identifying low-quality hospitals

Instead of being used to reward "top-performing hospitals," the composite quality measure may be better used to identify low-quality hospitals. The measures have low variance and strong ceiling effects and thus can be used to identify outliers. As shown in Exhibit 3, the composite measure clusters hospitals near the top of the scale, thus making identification of low-performing hospitals possible. Hospitals in the lowest tier tended to be low volume and exhibited a trend toward higher mortality. Low volume alone cannot identify low-quality hospitals because several hospitals were both low volume and high quality. The current high degree of adherence to published guidelines demonstrates that most hospitals are meeting or exceeding process guidelines.

Limitations of the data

Our data are limited by a number of factors. First, we did not have complete data on the lower 50 percent Of hospitals. Although we can identify hospitals in the lowest tier, we gathered the data on performance measures in this tier from external data sources. Second, measures of mortality and complications were not available for all hospitals, which limits the statistical power of our analysis. Furthermore, complications were limited to those sustained in the inpatient stay, whereas evaluation Of the first thirty days after surgery would be more beneficial. Hospitals in tier 4 (which tended to be lower-volume hospitals) were more likely than tiers 1, 2, or 3 to have missing data, which tends to underestimate any significant correlations.

Third, the outcomes data that are publicly available (see methods section) have not been formally validated. Yet they are highly regarded by consumers and purchased by insurers to judge hospital quality.[19] The data represent a "black box" to some extent, but they are the best-available data on quality for hip and knee replacement until national outcome registries take effect. Our main conclusion (that outcomes do not correlate with quality measures) is based on the data using the ordinal hospital tiers; these analyses do not involve data from outside the CMS. Fourth, UTI is not a very robust outcome measure because, although preventable, it is easily treatable and frequently underreported. Finally, the voluntary nature of the HQID project introduces selection bias. Hospitals that thought they were performing well on these measures would be more likely to join, thus pushing the average measure scores up.

Our analysis of the HQID demonstrates that the current generation of P4P measures based on process is inadequate. Hospital quality measures did not correlate with complications or mortality. Further research must be performed to create quality measures that assess clinically meaningful outcomes. Patients and payers need indices that accurately measure, in a risk-adjusted fashion, important outcomes such as hip dislocation, thirty-day mortality, and one-year reoperation. Only after such indices are developed and validated should they be tied to payments. However, public reporting of such validated quality and outcome measures, coupled with consumer-driven health care choices, might make P4P measures obsolete.

NOTES

1. See, for example, Epstein AM. Pay for Performance at the Tipping Point. New England Journal of Medicine 2007;356(no. 5):515–517. [PubMed: 17259445]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2. Landon BE, et al. Physician Clinical Performance Assessment: Prospects and Barriers. Journal Of the American Medical Association 2003;290(no. 9):1183–1189. [PubMed: 12953001]
3. Naylor CD. Public Profiling of Clinical Performance. Journal of the American Medical Association 2002;287(no. 10):1323–1325. [PubMed: 11886325]
4. See, for example, Colwell CW. Evidence-Based Guidelines for Venous Thromboembolism Prophylaxis in Orthopedic Surgery. Orthopedics 2007;30(no. 2):129–135. [PubMed: 17323635]
5. Kim S. Changes in Surgical Loads and Economic Burden of Hip and Knee Replacements in the U.S.: 1997–2004. Arthritis and Rheumatism 2008;59(no. 4):481–488. [PubMed: 18383407]
6. Glickman SW, et al. Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction. Journal of the American Medical Association 2007;297(no. 21):2373–2380. [PubMed: 17551130]
7. Premier Inc. Results. [accessed 9 December 2008]. http:www.premierinc.com/p4p/hqi/results/index.jsp
8. An expanded discussion of the methods is available in the Methods Appendix, online at http://Content.healthaffairs.org/cgi/content/full/28/2/526/DC1.
9. Weingart SN, et al. Use of Administrative Data to Find Substandard Care: Validation of the Complications Screening Program. Medical Care 2000;38(no. 8):796–806. [PubMed: 10929992]
10. Katz JN, et al. Association of Hospital and Surgeon Procedure Volume with Patient-Centered Outcomes of Total Knee Replacement in a Population-Based Cohort of Patients Age Sixty-five Years and Older. Arthritis and Rheumatism 2007;56(no. 2):568–574. [PubMed: 17265491]
11. See Appendix Exhibit 1 online, as in Note 8.
12. The distribution of composite quality scores is Shown in Appendix Exhibit 2 online; ibid.
13. See online Appendix Exhibit 3; ibid.
14. Rosenthal MB, et al. Pay for Performance in Commercial HMOs. New England Journal of Medicine 2006;355(no. 18):1895–1902. [PubMed: 17079763]
15. Bhattacharyya T, Hooper DC. Antibiotic Dosing before Primary Hip and knee Replacement as a Pay-for-Performance Measure. Journal Of Bone and Joint Surgery (American Volume) 2007;89(no. 2):287–291.
16. See, for example, R.M. Werner and E.T. Bradlow, Relationship between Medicare's Hospital Compare Performance Measures and Mortality Rates. Journal of the American Medical Association 2006;296(no. 22):2694–2702. [PubMed: 17164455]
17. Bradley EH, et al. Hospital Quality for Acute Myocardial Infarction: Correlation among Process Measures and Relationship with Short-Term Mortality. Journal of the American Medical Association 2006;296(no. 1):72–78. [PubMed: 16820549]
18. Lindenauer PK, et al. Public Reporting and Pay for Performance in Hospital Quality Improvement. New England Journal of Medicine 2007;356(no. 5):486–496. [PubMed: 17259444]
19. For consumer feedback, see Hospital Report Cards: Making the Grade. Harvard Health Letter. 2004 June;

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

EXHIBIT 1

Mean Hospital Performance Scores On Measures Of Quality, CMS/Premier Hospital Quality Initiative Demonstration (HQID) Segment For Hip And Knee Replacement

	Mean hospital score (SD)		
	Top 20 percent	**Remaining hospitals**	***p* value**
Proportion of patients who received prophylactic antibiotics within 1 hour before surgical incision	90% (5.79)	83% (10.56)	<0.0001
Proportion of patients who had appropriate prophylactic antibiotic selection for surgical patients	99% (2.16)	97% (3.43)	<0.030
Proportion of patients whose prophylactic antibiotics were discontinued within 24 hours after surgery end time	85% (9.22)	73% (14.96)	<0.030
Metabolic derangement avoidance index[a]	1.00 (0.000138)	1.00 (0.000161)	0.913
Hematoma avoidance index[a]	1.00 (0.000037)	1.00 (0.0000358)	0.864
Readmission (30-day) avoidance index[a]	1.00 (0.0333)	1.00 (0.0272)	0.488
Composite quality score	0.9570 (0.0169)	0.8466 (0.0984)	<0.0001

SOURCE: Authors' analysis; see methods section in text.

NOTES: CMS is Centers fo Medicare and Medicaid Services. SD is standard deviation.

[a]Top 50 percent performing hospitals only.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

EXHIBIT 2

Ceiling Effects In Performance Measures, CMS/Premier Hospital Quality Initiative Demonstration (HQID) Segment For Hip And Knee Replacement

Measure	Hospitals at 99% or above
Metabolic complication avoidance index[a]	101/101 (100%)
Hematoma avoidance index[a]	107/107 (100%)
Readmission avoidance index[a]	81/107 (81%)
Antibiotics administered within 1 hour before incision	1/212 (0.5%)
Antibiotics discontinued within 24 hours of surgery	4/212 (1.9%)
Antibiotic selection appropriate[a]	3/107 (2.8%)

SOURE: Authors' analysis; see methods section in text.

NOTE: CMS is Centers for Medicare and Medicaid Services.

[a]Data from these variables are available for top 50 percent performing hospitals only.

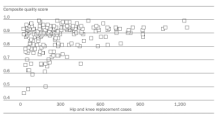
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

EXHIBIT 3

Relationship Between Composite Quality Score And Hip And Knee Surgical Volume Among Hospitals Participating In The CMS/Premier Hospital Quality Initiative Demonstration, 2004/05



SOURCE: Authors' analysis; see methods section in text.

NOTE: CMS is Centers for Medicare and Medicaid Services.

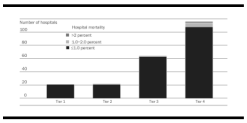
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

EXHIBIT 4

Mortality From Hip And Knee Replacements Among Medicare Beneficiaries, BY Hospital Tier, Among Hospitals Participating In The CMS/Premier Hospital Quality Initiative Demonstration, 2004/05



SOURCE: Authors' analysis; see methods section in text.

NOTES: For explanation of hospital tiers, see text. CMS is Centers for Medicare and Medicaid Services.