Evaluation of the Premier Hospital Quality Incentive Demonstration

Executive Summary: Impacts on Quality of Care, Medicare Reimbursements, and Medicare Beneficiaries' Length of Stay during the First Three Years of the Demonstration

March 3, 2009

Centers for Medicare & Medicaid Services
Office of Research, Development and Information
Research and Evaluation Group
Division of Research on Traditional Medicare

Table of Contents

List of Tables

Section 1: Description of the Premier Hospital Quality Incentive Demonstration

This Executive Summary presents findings of the Abt Associates evaluation of the Premier Hospital Quality Incentive Demonstration (PHQID) during its first three years (from October 2003 through September 2006).[1, 2] The PHQID included financial incentives for participating hospitals, public reporting to improve their quality of care, and penalties for low-scoring hospitals in the third year of the Demonstration. The primary focus of the Demonstration was to test whether paying incentives for high quality care would improve the quality of hospital inpatient care. The evaluation examined the impact of the demonstration on the quality of inpatient hospital care, Medicare reimbursements, and on Medicare beneficiaries' length of stay.

The Demonstration was managed by Premier, Inc., and was open to hospitals that were reporting quality measures through Premier's Perspective™ quality measurement system. Premier, Inc. is a healthcare purchasing and services company. As of 2004, it was owned by a collection of 203 hospitals and systems and had 1,418 members.[3] Perspective™ is a reporting system marketed by Premier and includes data on the quality measures used in the PHQID. Participation in the Demonstration was offered to the approximately 444 hospitals that were reporting under the Perspective™ system as of March 31, 2003. A total of 267 Medicare providers, representing 278 hospitals, elected to participate. Premier also markets the Perspective™ system to non-members, so that Demonstration participants include both members and non-members.[4]

This summary is organized as follows. Section 1 describes the Demonstration. Section 2 describes the total changes in the quality measure scores over the first three years of the Demonstration. Section 3 reports on those portions of the change in the scores that were due to the Demonstration. Sections 4 and 5 report the effects of the Demonstration on Medicare reimbursements and outlays and the Medicare beneficiary average length of stay, respectively. The last section, Section 6, summarizes the findings of the evaluation.

[1] Kennedy, S., Kling, R., Burstein, N., and Patrabansh, S. *Evaluation of the Premier Hospital Quality Incentive Demonstration: Impacts on Quality, Medicare Reimbursements, and Medicare Lengths of Stay: Volumes 1 and 2*, Abt Associates Inc., December 1, 2008.

[2] CMS extended the Demonstration for an additional three years, from October 1, 2006 through September 30, 2009.

[3] Source: http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp#TopOfPage.

[4] Premier estimates that about 15 percent of the hospitals that subscribe to *Perspective*™ are not Premier members. Of the 278 hospitals that participated in the Demonstration, 32 (about 12 percent) subscribe to Perspective™ but were not Premier members.

1.1 Demonstration Quality Measures

Demonstration incentives are based on performance measures of hospital quality in treating patients in each of five clinical areas, including:

- Heart attack (acute myocardial infarction or AMI)
- Heart failure (HF)
- Pneumonia (PN)
- Isolated coronary artery bypass graft (CABG) and
- Hip or knee replacement (HK).

Inpatients are categorized into a clinical area using certain primary and secondary diagnostic and procedure codes.[5]

With the exception of HK, the admissions in each clinical area included all adult inpatients with diagnostic or procedure codes for that clinical area. Demonstration hospitals did not necessarily have patients in each of the five clinical areas. Table 1.1 shows the numbers and percentages of Demonstration hospitals that had admissions in a clinical area in all of the first three years. Only about half of the Demonstration hospitals performed coronary artery bypass grafts in each of the first three years. Most performed HK or had AMI admissions in all three years. Almost all had HF and PN admissions in all three years. Total admissions in the various clinical areas ranged from 92,000 CABG admissions to more than 350,000 PN admissions.

Table 1.1: Demonstration Hospitals with Clinical Area Admissions in All Three Years [a]

	Clinical Area				
	AMI	**CABG**	**HF**	**PN**	**HK**
Number of Hospitals	216	118	232	235	195
Percent of Hospitals [b]	91.5	50.0	98.3	99.6	82.6
Total Number of Cases (FY2004-FY2006)	210,819	92,318	309,294	350,277	118,426

a) For the 236 Demonstration hospitals, excluding 24 dropouts, five hospitals with partial data, and one hospital with anomalous data for Hip and Knee readmission rates in FY2006 (see Section 2.2).
b) As a percent of the 236 hospitals.
Source: Kennedy, et al.

There were a total of 34 Demonstration quality measures for the five clinical areas, with four to nine measures for each area. Demonstration quality measures generally apply to all inpatients 18 years and older. The measures for each clinical area are listed in Table 2.2. The measures were taken from measures developed by CMS and its Quality Improvement Organizations, from the Joint Commission[6] core measures, from the Hospital Quality Alliance, and from the Agency for Healthcare Research and Quality (AHRQ) patient safety indicators, among others. Some

[5] Patients can be categorized into more than one clinical area, though this is unusual.

[6] The Joint Commission was formerly known as the Joint Commission on Accreditation of Healthcare Organizations (JCAHO).

measures were revised over the course of the demonstration (as shown in Table 1.2). The PHQID quality measures include both process and outcome measures.

Table 1.2: Demonstration Measures

Measure	Type of Measure	Description
Acute Myocardial Infarction (AMI)		
AMI-1	Process	Aspirin within 24 hours of arrival (if no contraindications).
AMI-2	Process	Aspirin prescribed at discharge (if no contraindications).
AMI-3	Process	Angeotensin converting enzyme inhibitor (ACEI) or (after 10/01/05) angeotensin receptor blocker (ARB) prescribed at discharge for left ventricular Systolic Dysfunction (LVSD) if no contraindications.
AMI-4	Process	Adult smoking cessation advice or counseling during hospital stay for patients who have smoked in the past year.
AMI-5	Process	Beta blocker prescribed at discharge (if no contraindications).
AMI-6	Process	Beta blocker prescribed within 24 hours of arrival (if no contraindications).
AMI-7	Process	Thrombolytic agent within 30 minutes of arrival (if thrombolysis within six hours).
AMI-8	Process	Percutaneous Coronary Intervention (PCI) within 120 minutes (90 minutes after 7/01/06) of admission (if PCI within 24 hours).
AMI-9	Outcome	Inpatient survival index for adult AMI patients.
Coronary Artery Bypass Graft (CABG)		
CABG-10	Process	Aspirin prescribed at discharge (if no contraindications).
CABG-11	Process	Coronary artery bypass graft using internal mammary artery. (Suppressed for entire Demonstration due to problems in identifying exclusions.)
CABG-12	Process	Prophylactic antibiotic within one or two hours prior to incision.
CABG-13	Process	Appropriate prophylactic antibiotic selection for isolated CABG patients. (Suppressed 10/01/05 through 6/30/06; revised selection guidelines after 7/01/06).
CABG-14	Process	Prophylactic antibiotics discontinued within 24 hours (48 hours after 1/01/05) after the end of surgery.
CABG-15	Outcome	Inpatient survival index for adult isolated CABG patients.
CABG-16	Outcome	Avoidance index for post-operative hemorrhage or hematoma.
CABG-17	Outcome	Avoidance index for post-operative physiologic and metabolic derangement.
Heart Failure (HF)		
HF-18	Process	Left ventricular function (LVF) assessment performed or planned.
HF-19	Process	Written discharge instructions covering activity level, diet, discharge medications, follow-up appointment, weight monitoring, and what to do if symptoms worsen.
HF-20	Process	Angeotensin converting enzyme inhibitor (ACEI) or (after 10/01/05) angeotensin receptor blocker (ARB) prescribed at discharge for left ventricular Systolic Dysfunction (LVSD) at discharge (if no contraindications).
HF-21	Process	Adult smoking cessation advice or counseling during hospital stay for patients who have smoked in the past year.

Table 1.2: Demonstration Measures (cont.)

Measure	Type of Measure	Description
Pneumonia (PN)		
PN-22	Process	Arterial blood gas or pulse oximetry with 24 hours before or after arrival.
PN-23	Process	Initial antibiotic regimen consistent with current guidelines during the first 24 hours (if antibiotics within 36 hours).
PN-24	Process	Initial blood culture prior to the first hospital administration of antibiotics.
PN-25	Process	Influenza vaccination for pneumonia patients aged 50 or older discharged during Oct. through Feb. (Suppressed for FY2005 and the first quarter of FY2006 due to vaccine shortage.)
PN-26	Process	Pneumococcal vaccination for patients aged 65 or older.
PN-27	Process	Initial antibiotic within 4 hours of arrival if received antibiotics within 24 hours.
PN-28	Process	Adult smoking cessation advice or counseling during hospital stay for patients who have smoked in the past year.
Hip and Knee Replacement (HK)		
HK-29	Process	Prophylactic antibiotic within one or two hours prior to incision.
HK-30	Process	Appropriate prophylactic antibiotic selection. (Suppressed 10/01/05 through 6/30/06; revised selection guidelines after 7/01/06).
HK-31	Process	Prophylactic antibiotics were discontinued with 24 hours after the end of surgery.
HK-32	Outcome	Avoidance index for post-operative hemorrhage or hematoma within one day after surgery.
HK-33	Outcome	Avoidance index for post-operative physiologic and metabolic derangement.
HK-34	Outcome	Avoidance index for readmission to an acute care hospital within 30 days of discharge.

Source: Kennedy, et al.

Process Measures: Scores for process measures reflect the proportion of patients whose treatment conforms to certain recommended practices. These recommendations are subject to various exclusions. The process measures are calculated with respect to the number of measure-relevant patients, after exclusions. This number is referred to as the score's denominator. The numerator is the number of the measure-relevant cases whose treatment conformed to the recommended treatment procedure.

Outcome Measures: Three of the five clinical areas also have outcome measures (AMI, CABG, and HK). Outcome measures involve the incidence of certain adverse events – usually during the hospital stay or within a short period after discharge. These include inpatient mortality rates for AMI and CABG, two types of surgical complications for CABG and HK (post-operative hemorrhages or hematomas, and post-operative physiologic and metabolic derangements), and 30-day readmission rates for HK. As with process scores, some patients may be excluded from the calculation of adverse event rates. Scores for the outcome measures are expressed in terms of an "adverse event avoidance index". First, the adverse event rate is converted into an avoidance rate by subtracting it from one (to give the proportion of patients who do *not* experience the adverse event). This creates a measure for which higher values are better,

consistent with the process measures. Second, the measure is adjusted for certain patient risk factors. Differences among hospitals in the rate of adverse events may reflect differences in patient circumstances that are outside of the hospital's control. To adjust for this, the observed avoidance rate is divided by a risk-adjusted avoidance rate[7] to create an avoidance index.

Measure Exclusions**:** Some patients within a clinical area may be excluded from one or more of the measures for that area. Some of these exclusions reflect situations in which the recommended treatment for that particular patient would differ from the general recommendation reflected in the quality measure. Other exclusions reflect situations in which the hospital may not have control over all stages of the treatment (such as patients who were transferred from another facility). In some cases, exclusions were revised over the course of the three years.

Composite Scores: Composite scores are constructed from the individual process and outcome scores for each of the clinical areas.

1.2 Demonstration Incentives

The Demonstration includes financial incentives and benefits from reporting quality scores. Financial incentives involve annual payments to top performers and penalties (in the third year of the Demonstration) for hospitals that do not meet certain benchmarks. Incentive payments and penalties are determined separately for each clinical area. Performance is assessed for each area based on a hospital's annual composite quality score for the area.[8] Hospitals whose composite quality scores for a clinical area are in the top 10 percent of scores for participating hospitals receive an incentive payment equal to 2 percent of the hospital's basic Medicare reimbursements[9] for patients in that clinical area.[10] Hospitals whose composite quality scores are in the second decile (the top 20 percent to 11 percent) for that clinical area receive an incentive payment equal to 1 percent of the hospital's basic Medicare reimbursements for patients in that clinical area. For Year Three, penalties were applied to hospitals whose composite scores fall into the bottom 20 percent of the first year scores. The PHQID financial

[7] The risk-adjustment methodology varies by outcomes measure.

[8] Incentive payments are based on Medicare reimbursements, but quality is measured for all adult patients in each of the five clinical areas, with the exception of HKs. Quality measures for HKs are measured only for Medicare beneficiaries, because data for one of the HK quality measures (readmission rates) are only available for Medicare beneficiaries.

[9] Basic Medicare payments are payments for discharges of Medicare beneficiaries whose principal diagnosis or procedure code places them in that clinical area, after adjustment of the standardized payment amount for local wages and cost of living factors. Incentive payments are paid on all Medicare discharges, including those that were excluded from the calculation of the quality scores.

[10] Composite scores for hospitals with very few patients may not be indicative of the hospital's actual quality. Accordingly, hospitals with fewer than 30 discharges in a clinical area in any year are excluded from the calculation of deciles for that year and are not eligible to receive any incentive payments for that clinical area in that year.

incentives payments to top-performing hospitals increased government outlays by $24.6 million over the three years. Penalties in year three were less than $104,000. Incentive payments and penalties by clinical area and year are displayed in Table 1.3.

Table 1.3: Incentive Payments and Penalties [a]

Payments	AMI	CABG	HF	PN	HK	Total [b]
Year 1 (FY2004)						
Number of Hospitals	49	27	52	52	43	123
Payment Amount	$1,755,902	$2,077,667	$1.817,574	$1,139,353	$2,060,639	$8,851,138
Year 2 (FY2005)						
Number of Hospitals	46	27	50	50	42	115
Payment Amount	$1,706,336	$1,877,534	$1,741,605	$1,069,370	$2,295,602	$8,690,447
Year 3 (FY2006)						
Number of Hospitals	44	25	48	48	41	112
Payment Amount	$1,422,183	$1,267,693	$1,560,325	$796,892	$1,961,687	$7,008,780
Penalties						
Year 3 (FY2006)						
Number of Hospitals	3	1	1	0	6	9
Penalty Amount	$22,759	$6,593	$5,494	0	$69,076	$103,922

a) Numbers from Centers for Medicare & Medicaid Services.
b) Total number of hospitals may be less that the sum of the clinical area numbers, since a hospital may receive payments or be penalized in more than one clinical area.

Source: Kennedy, et al.

CMS also provides public recognition of better performers by publishing lists of the top 50 percent of hospitals in each clinical area with special recognition of the hospitals with scores in the first or second deciles. These lists are published on the CMS website.[11]

Section 2: Changes in Average Quality Scores over the First Three Years of the Demonstration

This section describes the total changes in Demonstration hospital quality scores during the first three years of the Demonstration. Average scores for patients treated in Demonstration hospitals often rose substantially. In addition, differences among individual hospital composite scores were reduced over the three-year period.

This section describes changes in Demonstration quality measure scores. Measure scores are for the aggregate of Demonstration hospitals.[12] They were calculated as weighted averages of hospital scores, where weights are based on each hospital's number of cases relative to the total number of cases for that measure or that composite measure. Standard tests of statistical

[11] See http://www.cms.hhs.gov/HospitalQualityInits/35_hospitalpremier.asp.
[12] This excludes the 24 hospitals that withdrew before the end of FY2006, five other hospitals whose data were suppressed for one or more years because they failed data validation, and one hospital with apparently anomalous data for the hip and knee readmission rate in the third year of the Demonstration.

significance are used to distinguish real changes in the expected average scores for admissions to Demonstration hospitals from random fluctuations due to chance variation.[13] These tests take account of the number of changes being tested in assessing statistical significance; shaded entries in the tables below indicate changes in scores that are statistically significant at these lower p-values. In this section of the Executive Summary, note that finding statistically significant changes only indicates that there were real changes in measure scores for the Demonstration hospitals. Statistical significance does not indicate the extent to which changes are caused by the Demonstration; comparisons of changes in scores for Demonstration hospitals with changes in other hospitals are reported in Section 3.

Individual Measure Scores

Process Measure Scores: Table 2.1 presents data on measure scores for the first and twelfth quarters of the demonstration, p-values for the changes in scores, and the percent of the gap closed by the twelfth quarter for each score (where the gap is defined as the difference between the average score for the first quarter and a perfect score). For the AMI-1 measure, as an example, the first quarter value for the average Demonstration hospital is 93.7 percentage points, about 4 percentage points less than the average score for the twelfth quarter of the Demonstration. The average hospital closed about 63 percent of the difference between the average first quarter score and a perfect score for that measure. All process measure scores show statistically significant improvement, except for AMI-7, which also had the lowest initial value.

Table 2.1: Total Changes in Process Measure Scores during the First Three Years [a]

Measure		Q1	Q12	Q12 − Q1	Percent of Gap [b]	P-value Q12 − Q1
Acute Myocardial Infarction (AMI)						
AMI-1	Mean	0.937	0.977	0.040	63.2	<0.001
	(Std Err)	(0.005)	(0.002)	(0.005)		
AMI-2	Mean	0.948	0.982	0.034	64.9	<0.001
	(Std Err)	(0.004)	(0.002)	(0.004)		
AMI-3	Mean	0.758	0.906	0.148	61.4	<0.001
	(Std Err)	(0.013)	(0.008)	(0.014)		
AMI-4	Mean	0.806	0.990	0.184	94.6	<0.001
	(Std Err)	(0.016)	(0.002)	(0.015)		
AMI-5	Mean	0.904	0.978	0.075	77.6	<0.001
	(Std Err)	(0.007)	(0.002)	(0.006)		
AMI-6	Mean	0.878	0.953	0.074	61.0	<0.001

[13] Estimated standard errors for weighted average rates take into account both patient-level and hospital-level random effects.

Table 2.1: Total Changes in Process Measure Scores during the First Three Years [a]

Measure		Q1	Q12	Q12 − Q1	Percent of Gap [b]	P-value Q12 − Q1
	(Std Err)	(0.008)	(0.004)	(0.008)		
AMI-7 [c]	Mean	0.334	0.394	0.060	9.0	0.359
	(Std Err)	(0.032)	(0.068)	(0.064)		
AMI-8 [c]	Mean	0.540	0.796	0.256	55.7	<0.001
	(Std Err)	(0.023)	(0.015)	(0.026)		
Coronary Artery Bypass Graft (CABG)						
CABG-10	Mean	0.941	0.985	0.044	74.4	<0.001
	(Std Err)	(0.008)	(0.003)	(0.008)		
CABG-12	Mean	0.692	0.948	0.256	83.2	<0.001
	(Std Err)	(0.026)	(0.008)	(0.025)		
CABG-13	Mean	0.946	0.983	0.038	69.4	0.010
	(Std Err)	(0.013)	(0.006)	(0.014)		
CABG-14 [d]	Mean	0.479	0.821	0.342	65.7	<0.001
	(Std Err)	(0.038)	(0.027)	(0.035)		
Heart Failure (HF)						
HF-18	Mean	0.857	0.960	0.103	72.1	<0.001
	(Std Err)	(0.007)	(0.003)	(0.007)		
HF-19	Mean	0.423	0.774	0.351	60.8	<0.001
	(Std Err)	(0.025)	(0.014)	(0.023)		
HF-20	Mean	0.762	0.891	0.128	54.0	<0.001
	(Std Err)	(0.010)	(0.007)	(0.012)		
HF-21	Mean	0.603	0.976	0.373	93.9	<0.001
	(Std Err)	(0.025)	(0.004)	(0.026)		

Table 2.1: Total Changes in Process Measure Scores during the First Three Years [a]

Measure		Q1	Q12	Q12 − Q1	Percent of Gap [b]	P-value Q12 − Q1
Pneumonia (PN)						
PN-22	Mean	0.977	0.998	0.022	93.3	<0.001
	(Std Err)	(0.003)	(0.000)	(0.003)		
PN-23	Mean	0.762	0.878	0.116	48.7	<0.001
	(Std Err)	(0.007)	(0.005)	(0.008)		
PN-24	Mean	0.815	0.926	0.111	60.1	<0.001
	(Std Err)	(0.006)	(0.005)	(0.006)		
PN-25 [e]	Mean	0.400	0.788	0.388	64.7	<0.001
	(Std Err)	(0.018)	(0.013)	(0.017)		
PN-26	Mean	0.418	0.829	0.411	70.6	<0.001
	(Std Err)	(0.018)	(0.010)	(0.019)		
PN-27	Mean	0.640	0.811	0.171	47.4	<0.001
	(Std Err)	(0.009)	(0.007)	(0.009)		
PN-28	Mean	0.574	0.955	0.381	89.4	<0.001
	(Std Err)	(0.021)	(0.006)	(0.021)		
Hip/Knee Replacement (HK)						
HK-29	Mean	0.696	0.948	0.252	82.8	<0.001
	(Std Err)	(0.024)	(0.004)	(0.023)		
HK-30	Mean	0.973	0.987	0.014	52.2	0.001
	(Std Err)	(0.004)	(0.003)	(0.004)		
HK-31	Mean	0.489	0.889	0.401	78.4	<0.001
	(Std Err)	(0.029)	(0.012)	(0.030)		

Shading is used to indicate p-values that are statistically significant at an overall 0.05 level after a Bonferroni-Hochberg adjustment to take account of the number of tests in each column of p-values (26).

a) For the 236 Demonstration hospitals, excluding 24 dropouts, five hospitals with partial data, and one hospital with anomalous data for Hip and Knee readmission rates in FY2006.
b) Percent of Gap = the change in score as a percent of the possible improvement = 100(Q12 − Q1)/(1 − Q1)
c) Figures for AMI-7 and AMI-8 are for Q1 and Q11, because of measure definition and/or exclusion changes in the last quarter of the Demonstration.
d) Figures for CABG-14 are for Q1 and Q9, because of measure definition changes in Q10.
e) Figures for PN-25 (flu vaccination) are for Q1 to Q10, because the measure only applies to winter months.

Source: Kennedy, et al.

The maximum possible change is given by the gap between that actual Q1 scores and a perfect score of one. In general, the changes in individual measure scores were positively related to the size of their gaps, with the exception of AMI-7. Changes in scores for these 25 process measures (excluding AMI-7) covered from 47 to just under 95 percent of the gap (Table 2.1), with an average of just under 70 percent of the gap. Analysis indicates that changes in scores during the first year were usually considerably larger than changes during the third year.

Outcome Measure Scores: Quality scores were very high in Q1 for all of the outcomes measures. None of the changes in outcomes measures scores was statistically significant after adjusting for multiple comparisons.

Composite Quality Scores

All five composite quality scores for admissions in each clinical area increased during the first 12 quarters of the Demonstration, and all increases were statistically significant (Table 2.2). Scores for three areas (AMI, CABG, and HK) were greater than or equal to 0.86 in the first quarter of the Demonstration, and increased by 6.8 to 11.4 percentage points. Scores in the other two areas (HF and PN) were initially lower (66.9 and 69.8, respectively) and increased by 21.7 and 20.3 points, respectively. In addition, changes by the twelfth quarter of the Demonstration were significantly smaller than the changes at the start of the Demonstration.

Table 2.2: Weighted Average Composite Quality Scores during the First Three Years [a, b]

Condition		Composite Score		Q12 − Q1	P-value For Q12 − Q1
		Q1	Q12		
Heart Attack (AMI)	Mean	0.899	0.967	0.068	<0.001
	(Std Err)	(0.005)	(0.002)	(0.004)	
Coronary Artery Bypass Graft (CABG)	Mean	0.866	0.978	0.113	<0.001
	(Std Err)	(0.008)	(0.002)	(0.007)	
Heart Failure (HF)	Mean	0.669	0.886	0.217	<0.001
	(Std Err)	(0.012)	(0.006)	(0.011)	
Pneumonia (PN)	Mean	0.698	0.901	0.203	<0.001
	(Std Err)	(0.006)	(0.003)	(0.007)	
Hip/Knee Replacement (HK)	Mean	0.860	0.974	0.114	<0.001
	(Std Err)	(0.007)	(0.002)	(0.007)	

Shading is used to indicate p-values that are statistically significant at an overall 0.05 level after a Bonferroni-Hochberg adjustment to take account of the number of tests (5).

a) For the 236 Demonstration hospitals, excluding 24 dropouts, five hospitals with partial data, and one hospital with anomalous data for Hip and Knee readmission rates in FY2006.

b) Each hospital is weighted in proportion to its clinical area population (in the quarter).

Source: Kennedy, et al.

The changes in composite quality scores are almost entirely driven by changes in process scores. On average, the 26 weighted-average process scores started more than 28 percentage points below a perfect score of one in the first quarter and increased by more than 18 percentage points by the twelfth quarter. In contrast, the average difference between first-quarter outcome index scores and their maxima was less than 3 percentage points in the first quarter, and the average change between first and twelfth quarter was less than one percentage point.

As noted above, incentive payments are based on hospitals' relative scores in each Demonstration year, with payments going to hospitals in the top two deciles. The improvements in average performance and reductions in inter-hospital variation noted above do not necessarily imply any change in the rank ordering of hospitals. Analysis indicated, however, that there was some movement of hospitals' relative ranks over the three years of the Demonstration.

Section 3: Demonstration Impacts on Hospital Compare Quality Scores

Only a portion of the changes in participating hospitals' quality scores described in Section 2 can be attributed to the Demonstration. To estimate the effects of the Demonstration, Demonstration hospital outcomes are compared with estimates of what they would have been in the absence of the Demonstration based on the experiences of non-Demonstration hospitals.

3.1 Methods and Data

The Reporting Hospital Quality Data for the Annual Payment Update (RHQDAPU) program provides comparable data for Demonstration and non-Demonstration hospitals on 18 of the 34 Demonstration quality measures. Data on these 18 measures can be used to estimate Demonstration specific impacts. We estimate Demonstration program effects by contrasting Demonstration and comparison hospital scores for the RHQDAPU measures. The 18 RHQDAPU measures include all but one of the 19 process measures for three of the five clinical areas covered by the Demonstration – AMI HF, and PN (Table 3.1). Note that the measure indicator numbers used by RHQDAPU differ from those used by the Demonstration.[14] Results reported here use the better-known RHQDAPU numbers, with descriptive labels in tables to reduce confusion. The third column of Table 3.1 indicates which of the 18 measures are included in the 10 measure "starter set" included under the RHQDAPU program.

[14] RHQDAPU numbering starts anew with each clinical area, and sometimes uses a different ordering.

Table 3.1: Corresponding Hospital Compare and Demonstration Measure Numbers[a]

RHQDAPU Measure Number	Demonstration Measure Number	Starter Set Measure	Short Description
AMI			
AMI _1	AMI_1	S	Aspirin At Arrival
AMI _2	AMI_2	S	Aspirin At Discharge
AMI _3	AMI_3	S	Rx ACEI or ARB For LVSD at Discharge
AMI _4	AMI_4		Smoking Cessation Counseling
AMI _5	AMI_5	S	Rx Beta Blocker at Discharge
AMI _6	AMI_6	S	Beta Blocker at Arrival
AMI_7a	AMI_7		Thrombolytic Agent w/in 30 minutes
AMI_8a	AMI_8		PCI within 120 (90) minutes
HF			
HF _1	HF_19		Discharge Instructions
HF _2	HF_18	S	LVF Assessment
HF _3	HF_20	S	Rx ACEI or ARB For LVSD at Discharge
HF _4	HF_21		Smoking Cessation Counseling
PN			
PN _1	PN_22	S	O_2 Assessment
PN _2	PN_26	S	Pneumococcal Vaccine
PN _3	PN_24		Blood Culture
PN _4	PN_28		Smoking Cessation Counseling
PN_5b	PN_27	S	Initial Antibiotics w/in 4 hrs.
PN _6	PN_23		Antibiotic Selection

a) Includes all Demonstration AMI measures except the outcome measure, AMI_9, survival index, all Demonstration HF measures, and all Demonstration PN measures except PN_25, influenza screening.

Source: Kennedy, et al.

Analyses of Demonstration effects using Hospital Compare scores are based on data for patients in each of the three clinical areas from those Demonstration and comparison hospitals with complete data for the clinical area measures.[15] These data are reported under the RHQDAPU and published on the Hospital Compare website. The analysis uses quarterly hospital-level data.

Estimation of Demonstration effects were derived from a logistic regression specification, used to estimate the probability that the treatment of a given (measure-relevant) patient meets the measure requirements. As the Hospital Compare data do not include any individual patient descriptors, the probability is expressed as a function of hospital and area characteristics. A combination of propensity and regression analyses was used here to take account of potentially confounding differences between Demonstration and comparison hospitals.[16]

[15] Data include calendar quarters covered by Hospital Compare for at least the last two years of the Demonstration. Data for the starter-set measures are richer because they also include data for an additional three quarters of data, starting with the first quarter of 2004.

[16] The scores reported on the Hospital Compare website are four-quarter moving averages, and scores are omitted if a hospital has fewer than 30 cases in the four-quarter period. This analysis, in contrast, is based on the underlying quarter-by-quarter results for all hospitals, including those with fewer than 30 cases. The analysis is weighted to reflect the number of cases in each hospital. Since data on pre-Demonstration scores are not available, estimates of Demonstration effects are necessarily cross-sectional. For details about the models and procedures used to estimate

3.2 Findings

Various data summarizing Demonstration impacts are displayed in Table 3.2. Data in the first two rows are presented as context for depicting the size of Demonstration effects. Average composite scores for Demonstration hospitals during the last quarter of the demonstration were 96.7 for AMI, 88.6 for HF, and 90.1 for PN. During the Demonstration period, these scores increased by 6.8, 21.7, and 20.3 percentage points for AMI, HF, and PN, respectively (row 2). Estimated changes in scores for the Demonstration hospitals that can be attributed to the Demonstration are smaller, however.

Estimated Demonstration impacts for AMI, HF, and PN (row 3) were 0.7 percentage points, 3.8 percentage points, and 2.4 percentage points, respectively, all statistically significant. Impacts for AMI were primarily attributable to improvements in AMI_1, Aspirin at Arrival, and AMI_4, Smoking Cessation Counseling. Two-thirds of the HF impact was attributable to effects on HF_1, Discharge Instructions. The PN impact was mostly due to effects on PN_2, Pneumococcal Vaccine, PN_3, Blood Culture, and PN_5b, Initial Antibiotics.

The size of these impacts can be put into context using data from rows 1 and 2 of Table 3.2. Consider impacts relative to quality scores these hospitals would have achieved absent the Demonstration impact, calculated as the difference between the score in row 1 and the Demonstration impact in row 3 (96.7-0.7=96.0 for AMI, 84.8 for HF, and 87.7 for PN). Demonstration impacts account for portions of Demonstration facility scores, absent Demonstration effects, of from less than 1 percent to less than 5 percent (row 4).

Alternatively, Demonstration impacts can be compared to total changes in composite scores of the Demonstration hospitals (row 2). The percent of score change accounted for by the Demonstration ranges from 10 to 18 percent, i.e. from a tenth to less than a fifth of the observed improvement in quality scores (line 5 in Table 3.2).

the Demonstration impacts on hospital quality scores see *Kennedy, et al, Volume 1 (Chapters 3 and 4) and Volume 2 (Appendices C and D)*, Steven Kennedy, et al, Abt Associates, Inc., December 1, 2008.

Table 3.2: Summary of Demonstration Impacts

	AMI	HF	PN
(1) Mean composite score, twelfth quarter of Demonstration for Demonstration hospitals [a]	96.7	88.6	90.1
(2) Total change in composite score during Demonstration for Demonstration hospitals (percentage points) [a]	6.8	21.7	20.3
(3) Demonstration impact (percentage points)	0.7	3.8	2.4
(4) Demonstration impact, relative to mean score [b]	0.7%	4.5%	2.8%
(5) Demonstration impact, relative to total change in composite score (row 2)	10.3%	17.5%	11.9%

Data in this table derived from Tables 2.6 above and 4.16 in Kennedy, ibid. Shading indicates statistically significant effects. The statistical significance of each of the three estimated effects on composite scores was areas was assessed using a Bonferroni–Hochberg sequential adjustment, with a starting two–tailed critical p–value of 0.0167 (=0.05/3). More stringent test levels for sets of tests were set so that the probability of having any false positive in the set was less than 0.05.

[a] Mean values based on Demonstration data, for the analytic sample of Demonstration hospitals with complete Hospital Compare data.

[b] Measured as Demonstration mean (row 1), net of estimated Demonstration impact for Year 3 (row 3b).

Source: Kennedy, et al.

The large increases in quality scores not attributable to the Demonstration may reflect a variety of factors, including normal improvements in medical practice, the advent of widespread public reporting under Hospital Compare, and the influence of other quality improvement initiatives, including other pay-for-performance programs.

Section 4: Demonstration Impacts on Medicare Reimbursements and Payments

Under the Medicare inpatient prospective payment system, the hospital stay's Diagnostic Related Group (DRG) and certain area or hospital-specific cost factors determine reimbursements for beneficiary inpatient stays. Quality changes could affect the hospital's total Medicare reimbursements either by affecting the reimbursement per stay or the number of hospital stays. Quality changes could affect reimbursements per stay by changing the diagnosis or the incidence of certain complications in ways that would change the stay's DRG. Quality changes could affect the number of stays by affecting the incidence of readmissions.

While improvements in quality could affect Medicare reimbursements, it is not a guarantee. Improved treatment need not change a stay's DRG. Even if changes in quality measures such as prescribing aspirin for AMI patients have material effects on patients' long-term health, they may not have an appreciable effect on short-term readmissions. Most importantly, the

Demonstration's immediate effects on Medicare inpatient *reimbursements* to hospitals may be quite different from its effects on participating hospitals' *costs* of care. Improvements in the PHQID quality scores could affect the costs of care in ways that would not be reflected in current Medicare inpatient hospital reimbursements. For example, improved care might reduce hospitals' costs by reducing lengths of stay (discussed in Section 5) or the incidence of complications. Such cost reductions would not change current hospital reimbursements unless they changed DRG codes or the incidence of readmissions.

4.1 Methods and Data

The Demonstration's effect on Medicare reimbursements was estimated by modeling the reimbursement per episode of care. An episode of care begins when a physician admits a patient to the hospital for one of the five clinical conditions included in the demonstration (AMI, HF, CABG, PN, and HK); this first admission is referred to as the index admission. The episode also includes any subsequent admissions for any conditions that begin within 90 days after discharge for the index admission.[17] The dependent variable for the model was defined as the DRG weight for the episode of care; it includes the DRG weights for the index admission and any subsequent readmissions that occur within 90 days after discharge for the index admission. A combination of propensity and regression analyses was used to estimate Demonstration effects. The estimated DRG weight was later converted into a dollar effect. Effects associated with episodes in each of the five clinical areas covered by the Premier Demonstration measures were obtained along with the estimated total effect of the all five clinical areas taken together.[18] Demonstration effects on total government outlays under the Demonstration were calculated by combining reimbursements with the PHQID incentive payments (net of penalty payments in the third year of the Demonstration).

Medicare Provider Analysis and Review files (MEDPAR) data were used in the analyses for the five fiscal years, FY2002 through FY2006. These years include the two fiscal years before the Demonstration began and the first three years of the Demonstration. Information on reimbursements and patient characteristics was abstracted from the November 2007 update of the MEDPAR. The major source of data on hospital characteristics was the 2005 American Hospital Association (AHA) survey, which collected data for calendar year 2003. Data from the

[17]Readmissions for the five clinical conditions included in the Demonstration do not create new episodes of care if they are included as readmissions in the episode associated with a previous admission.

[18] Medicare discharges in these five clinical areas accounted for about 20 percent of total Medicare discharges from all acute care hospitals. The PHQID participants account for about 11 percent of total Medicare discharges in the five clinical areas. Effects of the Demonstration on the DRG weight per episode were estimated, and then converted into estimates of Medicare reimbursements by multiplying the DRG weight per episode by the standardized payment amount for 2006, which was also adjusted to take account of local area wages. For more detail about the methods used to estimate reimbursement effects see Kennedy, S. et al, Vol. 1 (Chapters 3 and 5) and Vol. 2 (Appendix E).

Healthcare Cost Report Information System (HCRIS) and the Provider of Service (POS) data file, maintained and updated periodically by CMS, were used to perform the analyses.

4.2 Findings

Analysis revealed no evidence of any material Demonstration effect on Medicare reimbursements (Table 5.1). The average episode reimbursement per index admission in Demonstration hospitals across all five clinical areas was $13,128.[19] The estimated Demonstration effect on reimbursements per episode admission was an increase of $11, or one-tenth of 1 percent of the average total reimbursement per episode. This estimate was not statistically significantly different from zero. In addition, the estimated effect on Medicare reimbursements for each of the individual clinical areas was not statistically significant.

[19] As noted above, these are average per episode reimbursements. Amounts and effects include reimbursements for any subsequent readmissions as well as reimbursements for the index admission.

Table 4.1: Estimated Demonstration Effects on Medicare Reimbursements per Episode for the First Three Years of the Demonstration [a]

	Mean Reimbursement Per Episode [b]	Effect on Reimbursements Per Episode		
		Dollars	Percent of Mean [d]	P–value (two–tailed)
COMBINED AREAS				
Estimate	$13,128	$11	0.1%	0.850
95% Interval		[–$104, $126]	[–0.8%, 1.0%]	
AMI				
Estimate	$15,764	–$73	–0.5%	0.457
99% Interval [c]		[–$327, $180]	[–2.1%, 1.1%]	
CABG				
Estimate	$27,419	–$253	–0.9%	0.301
99% Interval [c]		[–$882, $377]	[–3.2%, 1.4%]	
HF				
Estimate	$12,260	$27	0.2%	0.820
99% Interval [c]		[–$283, $337]	[–2.3%, 2.8%]	
PN				
Estimate	$10,527	$64	0.6%	0.535
99% Interval [c]		[–$201, $329]	[–1.9%, 3.1%]	
HK				
Estimate	$11,361	$66	0.6%	0.509
99% Interval [c]		[–$191, $322]	[–1.7%, 2.8%]	

Shading would indicate any statistically significant effects on reimbursements. The statistical significance of the estimated effect for the combined areas was assessed using a two–tailed test at the 0.05 level. The statistical significant of the estimated effects for the individual clinical areas was assessed using a Bonferroni–Hochberg sequential adjustment, with a starting two–tailed critical p–value of 0.01. More stringent test levels for sets of tests were set so that the probability of having any false positive in the set was less than 0.05.

a) Estimated dollar effects are equal to the estimated Demonstration effects on reimbursement DRG weights per episode times the standardized payment amount for FY06.
b) The mean reimbursements per episode are equal to the mean reimbursement DRG weight per episode in Demonstration hospitals, times the standardized payment amount for FY06.
c) 99% intervals are used for the clinical area estimates in order to provide 95% intervals for the entire set – that is, so that the probability is at least 0.95 that every one of the clinical intervals covers the true value.
d) Details of dollar and percentage changes may not agree due to rounding. Confidence intervals for percentage effects are simply the confidence intervals for the estimated effect on reimbursements expressed as percentages of mean reimbursements per episode.

Source: Kennedy, et al.

Since the PHQID does not appear to have reduced Medicare reimbursements, there is no evidence that the Demonstration is budget neutral. Estimated demonstration effects are not significantly different from zero,[20] and payment of incentives under the Demonstration increases the magnitude of payments per episode (Table 4.2). The estimated effect on reimbursements plus incentive payments (net of penalty payments) was a net increase in costs of $41 (three-tenths of 1 percent) per episode over all clinical areas.

Table 4.2: Estimated Demonstration Effects on Total Medicare Payments per

[20] A 90 percent confidence interval was used to assess statistical significance, reflecting the fact that the principal concern was whether the total might be less than or equal to zero. Therefore, a one-tailed test was used to test this hypothesis.

Episode for the First Three Years of the Demonstration [a]				
	Mean Reimbursement Per Episode [b]	Effect on Payments Per Episode		
		Dollars	Percent of Mean Reimbursements [c]	P–value (one–tailed)
COMBINED AREAS				
Estimate	$13,159	$41	0.3%	0.240
90% Interval [d]		[$-55, $137]	[-0.4%, 1.0%]	
AMI				
Estimate	$15,803	$-34	-0.2%	0.364
98% Interval [e]		[$-263, $195]	[-1.7%, 1.2%]	
CABG				
Estimate	$27,513	$-158	-0.6%	0.258
98% Interval [e]		[$-727, $410]	[-2.6%, 1.5%]	
HF				
Estimate	$12,282	$49	0.4%	0.342
98% Interval [e]		[$-231, $329]	[-1.9%, 2.7%]	
PN				
Estimate	$10,538	$76	0.7%	0.231
98% Interval [e]		[$-163, $315]	[-1.6%, 3.0%]	
HK				
Estimate	$11,406	$110	1.0%	0.134
98% Interval [e]		[$-122, $342]	[-1.1%, 3.0%]	

Shading would indicate statistically significant effects on reimbursements. The statistical significance of the estimated effect for the combined areas was assessed using a one–tailed test at the 0.05 level. The statistical significance of the estimated effects for the individual clinical areas was assessed using a Bonferroni–Hochberg sequential adjustment, with a starting one–tailed critical p–value of 0.01. More stringent test levels for sets of tests were used so that the probability of having any false positive in the set was less than 0.05.

a) Estimated dollar effects are equal to the sum of the estimated Demonstration effects on reimbursement DRG weights per episode and the actual incentive payment DRG weights per episode (net of penalty payments) times the standardized payment amount for FY06.
b) The mean reimbursement per episode is equal to the mean reimbursement DRG weight per episode in Demonstration hospitals, times the standardized payment amount for FY06.
c) Details of dollar and percentage changes may not agree due to rounding. Confidence intervals for percentage effects are simply the confidence intervals for the estimated effect on reimbursements expressed as percentages of mean reimbursements per episode.
d) A 90% interval is reported because the primary interest is in the one–tailed test of the null hypothesis of no increase in total payments.
e) 98% intervals are used for the clinical area estimates in order to provide 90% intervals for the entire set – that is, so that the probability is at least 0.90 that every one of the clinical intervals covers the true value.

Source: Kennedy, et al.

Analyses revealed that reported findings are not sensitive to various changes in data definitions and model assumptions. Estimated effects using episodes based on 30-day and 60-day windows of care are smaller than reported 90-day demonstration effects. Second, it was hypothesized that Demonstration effects on reimbursements might increase over time. Separate Demonstration effects for each of the first three years of the Demonstration were estimated. Findings offered no evidence of differences in effects for these three years. Finally, analysis revealed that estimates were robust to modest variations in methods used to estimate the underlying statistical models.

Findings are subject to three caveats. First, the Demonstration hospitals consisted of hospitals that were already participating in the Premier Perspective™ reporting system. Consequently, it is not certain that the Demonstration results can be generalized to other hospitals. Second, while estimated statistical models appeared to perform reasonably well, it is not certain that all

confounding influences on inpatient hospital reimbursements (in the absence of random assignment) have been accounted for. Finally, it should be noted that the estimated Demonstration effects should capture the effects of the PHQID financial incentives as well as the effects of public reporting for the PHQID measures that were not included in the RHQDAPU public reporting initiative.

Section 5: Demonstration Impacts on Medicare Beneficiary Episode Length of Stay

While improved quality of care might reduce lengths of stay, Demonstration effects on length of stay would not affect Medicare reimbursements as noted in the previous section. However, effects on lengths of stay would be expected to affect hospital costs (other things equal), which might eventually indirectly affect Medicare reimbursements to hospitals.

Effects of the Demonstration on episode length of stay were analyzed using the same empirical specifications used to estimate Demonstration effects on Medicare reimbursements.[21] The episode is the unit of analysis (as in Section 4) to capture effects on total inpatient days associated with both the index admission and any subsequent readmissions during the episode.

Table 5.1 shows the average episode lengths of stay for Medicare beneficiary episodes in Demonstration and comparison hospitals. There are material reductions in episode lengths of stay in both Demonstration and comparison hospitals over the period FY 2002 through FY 2006. The average episode length of stay for all five areas declined by 8.2 percent in Demonstration hospitals and by 6.6 percent in comparison hospitals. These reductions included large reductions for HK, moderate reductions for AMI, and smaller reductions for CABG, HF, and PN.

[21]They include readmissions to acute-care inpatient hospitals only (see p. 16). For a detailed description of the procedures used to estimate the average length of stay demonstration effects, see Kennedy, S., et al, Vol. 1 (Chapter 6).

Table 5.1: Average Medicare Beneficiary Episode Lengths of Stay in Demonstration and Comparison Hospitals [a, b]

	All Areas	AMI	CABG	HF	PN	HK
Demonstration Hospitals						
FY2002	10.434	9.620	13.290	10.879	11.036	7.457
FY2006	9.581	8.812	13.211	10.418	10.687	5.845
FY06 − FY02	–0.854	–0.808	–0.078	–0.461	–0.349	–1.613
Percent Change [c]	-8.2%	-8.4%	-0.6%	-4.2%	-3.2%	-21.6%
Comparison Hospitals [d]						
FY2002	10.752	10.208	13.817	11.165	11.151	7.595
FY2006	10.046	9.438	13.579	10.687	10.931	6.248
FY06 − FY02	–0.707	–0.771	–0.238	–0.478	–0.220	–1.348
Percent Change [c]	-6.6%	-7.6%	-1.7%	-4.3%	-2.0%	-17.7%

a) The sample of hospitals is the sample described in the previous section. It excludes hospitals that are not reimbursed under PPS and hospitals that were not recorded as participating in Medicare in the Provider of Service (POS) files for FY2002 through FY2005. In addition, the samples for individual clinical areas exclude hospitals with fewer than ten admissions in that area in any of the five fiscal years (FY2002 through FY2006). The sample of episodes includes all episodes associated with Medicare beneficiary admissions to sample hospitals in the five clinical areas covered by the Demonstration quality measures.
b) Lengths of stay are for all episodes associated with admissions of Medicare beneficiaries in any of the five clinical areas covered by the Demonstration quality measures. Lengths of stay are measured in days and they include both the initial admission and any subsequent readmissions within 90 days of the index admission discharge.
c) FY02 to FY06 change as a percent of the FY02 length of stay.
d) Comparison hospitals exclude hospitals that had no counterpart in the analysis of any clinical area.

Source: Kennedy, et al.

The somewhat larger decline in average lengths of stay in Demonstration hospitals reflects differences in hospital and patient characteristics rather than the effects of the Demonstration. Table 5.2 shows the mean episode lengths of stay and the estimated Demonstration effects on episode lengths of stay during the first three years of the Demonstration. The first row reports effects for episodes in all five clinical areas, and subsequent rows present estimates for each of the five clinical areas. The estimated Demonstration effect for all five clinical areas is an estimated increase of 24 thousandths of 1 day in the average Medicare beneficiary episode length of stay, or two-tenths of 1 percent of the average 9.85 day episode length of stay in Demonstration hospitals during the first three years of the Demonstration. The estimate is not statistically significant. In addition, none of the estimates for the five clinical conditions is statistically significant.

Table 5.2: Estimated Demonstration Effects on Episode Lengths of Stay [a]

	Mean Episode Length of Stay [b]	Effect on Episode Lengths of Stay		
		Days	Percent of Mean [d]	P–value (two–tailed)
COMBINED AREAS				
Estimate	9.854	0.024	0.2%	0.722
95% Interval		[–0.108, 0.156]	[–1.1%, 1.6%]	
AMI				
Estimate	9.152	–0.073	–0.8%	0.427
99% Interval [c]		[–0.310, 0.164]	[–3.4%, 1.8%]	
CABG				
Estimate	13.238	–0.060	–0.5%	0.827
99% Interval [c]		[–0.766, 0.646]	[–5.8%, 4.9%]	
HF				
Estimate	10.613	0.094	0.9%	0.410
99% Interval [c]		[–0.200, 0.388]	[–1.9%, 3.7%]	
PN				
Estimate	10.708	0.003	0.0%	0.982
99% Interval [c]		[–0.292, 0.297]	[–2.7%, 2.8%]	
HK				
Estimate	6.380	0.062	1.0%	0.641
99% Interval [c]		[–0.282, 0.406]	[–4.4%, 6.4%]	

Shading would indicate any statistically significant effects on lengths of stay. The statistical significance of the estimated effect for the combined areas was assessed using a two–tailed test at the 0.05 level. The statistical significant of the estimated effects for the individual clinical areas was assessed using a Bonferroni–Hochberg sequential adjustment, with a starting two–tailed critical p–value of 0.01. We set more stringent test levels for sets of tests so that the probability of having any false positive in the set was less than 0.05.

a) Estimated effects are for Medicare beneficiary episodes in Demonstration Hospitals during the first three years of the Demonstration (FY04 through FY06).

b) The mean episode length of stay is equal to the average number of inpatients days for Medicare beneficiary episodes in Demonstration Hospitals during the first three years of the Demonstration (FY04 through FY06). It includes inpatient days for both the initial admissions and any readmissions within 90 days of the initial admission discharge date.

c) 99% intervals are used for the clinical area estimates in order to provide 95% intervals for the entire set – that is, so that the probability is at least 0.95 that every one of the clinical intervals covers the true value.

d) Details of day and percentage changes may not agree due to rounding. Confidence intervals for percentage effects are simply the confidence intervals for the estimated effect on lengths of stay expressed as percentages of mean episode length of stay.

Source: Kennedy, et al.

The estimated combined effects for all three Demonstration years could be misleading if Demonstration effects on Medicare beneficiary episode lengths of stay changed over the course of the Demonstration. To explore this possibility, Demonstration effects in each of the three Demonstration years, for the combined clinical areas and for each area separately, were estimated. There were no statistically significant differences over the three-year period. Analysis also indicates that there is no Demonstration effect on average length of stay for the index admission or for readmissions when estimated separately for all five clinical areas combined and for each clinical area.

Section 6: Conclusion

Results of the Demonstration evaluation suggest that the Demonstration contributed to quality increases. The average composite quality scores for each of the five clinical conditions increased substantially for hospitals participating in the Demonstration during the first three years. However, quality also increased substantially for hospitals that posted their quality scores on the

Hospital Compare website. Only a small portion of the increase in quality for participating hospitals can be attributed to the Demonstration—between 11 percent and 18 percent of the total change in quality--for the 3 conditions for which data were available (AMI, HF, and PN). At the same time, no statistically significant effect on acute care hospital Medicare reimbursements for the five clinical conditions taken together or for any of the individual conditions was found. This is not surprising, given that Medicare reimburses for inpatient care on a DRG basis and that the Demonstration effect on quality was found to be small. Evaluation results suggest that Demonstration hospitals did not incur savings associated with reductions in length of stay that could be attributed to the Demonstration. In general, length of stay was observed to have decreased over time for both demonstration and comparison hospitals.

Since the PHQID did not appear to reduce Medicare reimbursements, it is unlikely that the demonstration was budget-neutral – in the aggregate, payments were made to high-performing hospitals, net of penalty payments and in the absence of savings that could be attributed to the Demonstration. While evidence from the Demonstration does not indicate that the pay-for-performance experiment has resulted in program savings, participating hospitals may have experienced savings that were not realized by the Medicare program and not identified in this analysis. More study is needed to help determine whether pay-for-performance arrangements are cost-saving to facilities.