January 28, 2011

Evaluation of the Premier Hospital Quality Incentive Demonstration

Phase I Final Report

Prepared for

Linda Radey
Centers for Medicare & Medicaid Services
Office of Research, Development, and Information
Mail Stop C3-19-26
7500 Security Boulevard
Baltimore, MD 21244-1850

Prepared by

Pamela Spain
Shulamit Bernard
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709

RTI Project Number 0209853.019.001.004

Evaluation of the Premier Hospital Quality Demonstration
Phase I Final Report

Shulamit Bernard, Project Director
Pamela C. Spain

Federal Project Officer: Linda Radey

RTI International

CMS Contract No. HHSM-500-2005-00029I

January 28, 2011

This project was funded by the Centers for Medicare & Medicaid Services under contract no. HHSM-500-2005-00029I. The statements contained in this report are solely those of the authors and do not necessarily reflect the views or policies of the Centers for Medicare & Medicaid Services. RTI assumes responsibility for the accuracy and completeness of the information contained in this report.

RTI International is a trade name of Research Triangle Institute.

ACKNOWLEDGMENTS

We would like to express our gratitude to RTI staff members Alton Wright, Valentina Akhmerova, and Aleksandra Petrovic for their programming support for this report.

CONTENTS

List of Figures

List of Tables

EXECUTIVE SUMMARY

Overview

In July 2003, the Centers for Medicare & Medicaid Services (CMS) contracted with Premier, Inc., to launch a hospital pay-for-performance Demonstration whereby hospitals would receive a financial bonus for high-quality care and be at risk for paying penalties for poor-quality care that has not improved. Called the Premier Hospital Quality Incentive Demonstration (PHQID or Demonstration), this national pay-for-performance project was designed to determine whether economic incentives are effective at improving the quality of inpatient hospital care. The initial Demonstration ran for 3 fiscal years, beginning on October 1, 2003, and ending on September 30, 2006; the Demonstration was then extended for an additional 3 years beginning October 1, 2006, and ending September 30, 2009.

PHQID was designed to test the effects of pay-for-performance incentives on the quality of hospital care. To motivate improvements in quality, the Demonstration included an evolving array of annual payments beginning with payments to top performers in each of five clinical areas:

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

Incentives for the first 3 years of the Demonstration were paid to hospitals whose composite quality score (CQS) for a given clinical area was in the first or second decile of scores for participating hospitals. An incentive payment was equal to 2% of the hospital's basic Medicare reimbursements for patients in that clinical area for hospitals scoring in the first decile. Hospitals whose CQSs were in the second decile (the top 11% to 20%) for that clinical area received an incentive payment equal to 1% of the hospital's basic Medicare reimbursements for patients in that clinical area. Beginning in Year 3, penalties were applied to hospitals whose composite scores fell below thresholds established in the first year based on the scores for the hospitals scoring in the bottom two deciles.

Whereas during the first 3 years of the Demonstration hospitals received incentive payments when their composite scores were in the top 20% for a given fiscal year, beginning in Year 4, the incentive system was changed substantially. Incentives were awarded based on threshold attainment, top performance, and significant improvement:

- **Median-Level Attainment Award:** Hospitals that attain or exceed the median level as measured by the 2 years prior CQS in a given clinical area, will receive an incentive payment.

- **Top Performer Award:** Hospitals that have a CQS for a given clinical area that falls in the top 20% of scores will receive an additional payment. These hospitals will receive the median-level attainment award as well.

- **Top Improvement Award:** Hospitals that attain median-level performance and are among the top 20% of hospitals with the largest percentage quality improvement in each clinical area will receive an additional incentive payment. Improvement will be calculated based on the performance year compared to 2 years prior. This group will receive the median-level attainment award as well. Top performers are not eligible for the improvement award.

In a CMS-funded evaluation of the first three years of the PHQID, Kennedy and colleagues (2008) at Abt Associates performed an evaluation to describe the changes in the quality measure scores over the first 3 years of the Demonstration. They reported on the portion of the changes in the scores that could be attributed to the Demonstration as well as the effects of the Demonstration on Medicare reimbursement and outlays and Medicare beneficiary average length of stay. Results of the evaluation suggest that quality increased considerably among hospitals participating in the Demonstration. However, quality also increased substantially among hospitals that posted their scores on the Hospital Compare Web site and who were not part of the Demonstration. The evaluation could only examine the Demonstration effect for three conditions for which comparison data were available (AMI, HF, and PN). Among these measures, participation in the Demonstration contributed to a small portion of the increase in quality (between 10% and 18% of the total change in quality).

Kennedy and colleagues' evaluation results also suggested that the Demonstration hospitals did not experience savings associated with reductions in length of stay that could be attributed to the Demonstration. In general, during the first 3 years of the Demonstration, length of stay decreased over time for both Demonstration and comparison hospitals. Finally, the Abt report concluded that since the PHQID did not appear to reduce Medicare reimbursement, it is unlikely that the Demonstration was budget-neutral; in the aggregate, payments were made to high-performing hospitals, net of penalty payment and in the absence of savings that could be attributed to the Demonstration (Kennedy et al., 2008). Complete budget neutrality for the Medicare program was not explored since differences in post-acute care were not examined.

Current Evaluation of the Demonstration

The current evaluation of the Demonstration, conducted by RTI International, includes all 6 years of the Demonstration and is conducted in two Phases. In Phase I, we report on results of the analysis of Demonstration data for Years 1 to 4; in Phase II we will extend the analysis to all 6 years of the Demonstration.

In this Phase I Report, we present our findings for Years 1 to 4 of the Demonstration focusing on trends in quality within the Premier Demonstration hospitals that were in the Demonstration Year 1 to Year 4. We describe the trends in the quality measures for each quarter of Year 1 through Year 4, including those hospitals that reported measures in all 16 quarters. We also describe the results of the Pay for Performance in Year 4, including analyses of the hospital decile thresholds and the payments that were made to hospitals in Year 4. This phase of the

evaluation does not address the proportion of improvement in quality that is attributable to the Demonstration; that analysis, using comparison hospitals, will be conducted in Phase II.

Phase I Results

Results described in this report are grouped into two broad categories: (1) Participating Hospitals' Trends in Quality for Demonstration Years 1 to 4 and (2) Incentive Payment Analyses.

Participating Hospitals' Trends in Quality: Demonstration Years 1 to 4

We performed trend analyses among hospitals that reported measures in all 16 quarters of the Demonstration between Year 1 and Year 4. **Table ES-1** below shows the number of hospitals analyzed within each clinical focus area: 194 for AMI; 109 for CABG; 215 for HF; 164 for HK; and 220 for PN.

Table ES-1
Hospitals with CQS measures in all 16 quarters of PHQID

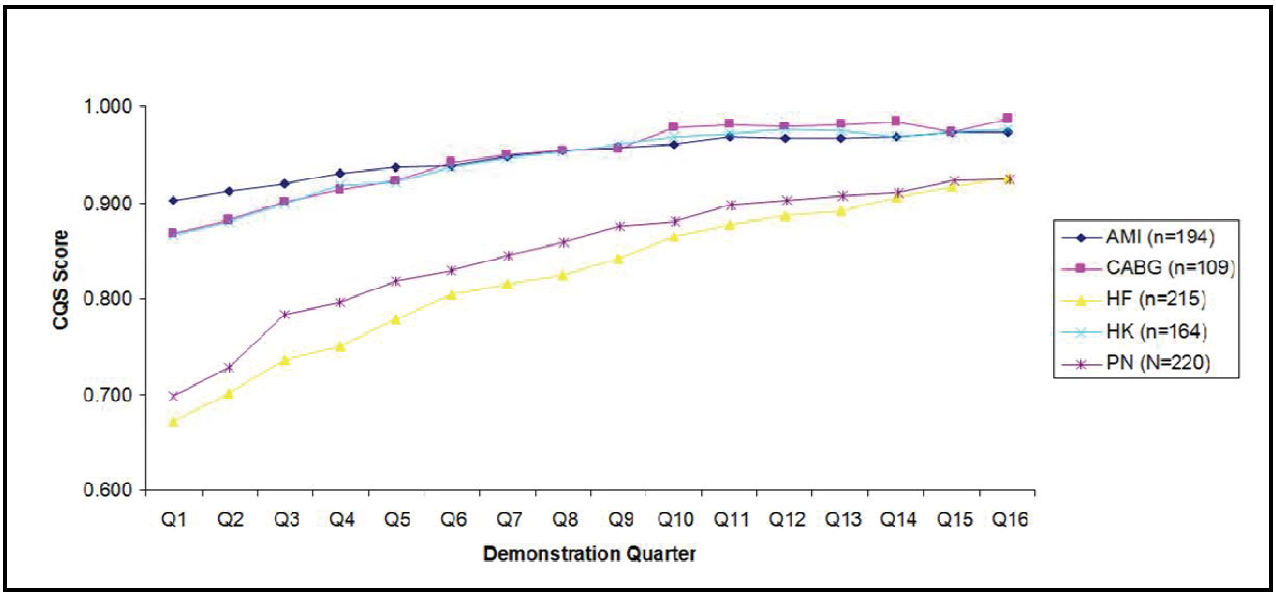
Clinical focus area	N hospitals with CQS In Year 4	N (%) hospitals with CQS In all 16 quarters
AMI	207	194 (93.7%)
CABG	116	109 (94.0%)
HF	224	215 (96.0%)
HK [a]	215	164 (76.3%)
PN	222	220 (99.1%)

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] The difference in percentage of hospitals reporting all 16 quarters of data for HK versus other clinical focus areas may be explained by the change in the population included in HK. Beginning in Y4 of the Demonstration, all patients, regardless of payer, were included in this clinical focus; this compares with only Medicare patients during the Years 1–3.

Trends over Time. We examined trends for all clinical areas and all process and outcome measures looking for differences in measure means between Q1 and Q16. We performed the Bonferonni correction for multiple testing. For most measures, the trend showed increasing quality over time between Q1 and Q16. However, there were several measures where we saw ceiling effects; some measures were high at the outset and remained high throughout the 4 years. An example is shown below in **Figure ES-1**.

Figure ES-1
CQS measures for Demonstration Years 1–4, [a] means



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

Percentage Change over Time. In addition to examining trend lines we also quantified the amount of change that occurred during the first 4 Demonstration years. We computed, for every measure, the percentage change between Q1 and Q16 as:

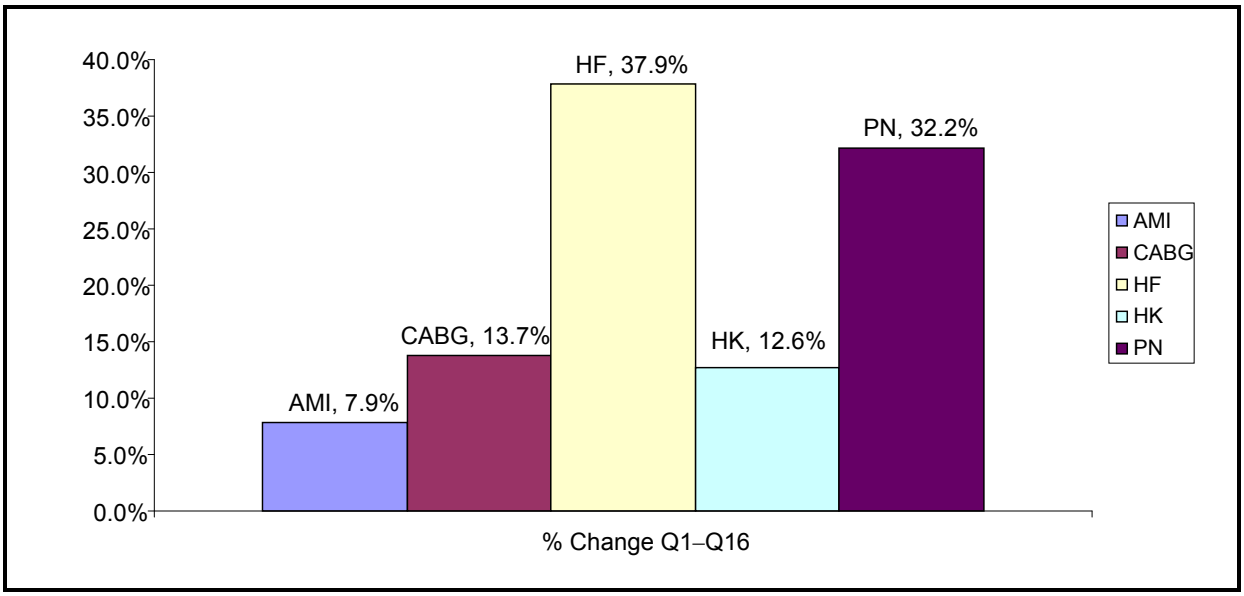
[(Year 4 score – Year 1 score)/Year 1 score]*100

Nearly all process, outcome, and CQS scores had positive percent changes between Q1 and Q16. An example of this is shown below in **Figure ES-2**.

Rates of Change over Time. We also examined whether there were statistical differences in *rates* of change for Year 1(Q1 to Q4) versus Year 2(Q5 to Q8) versus Year 3(Q9 to Q12) versus Year 4(Q13 to Q16). This analysis was designed to examine whether improvements in quality were more notable during the earlier years of the Demonstration and whether quality improvement was slowing down over time.

For most measures we found that the rate of change during Year 1 (between Q1 and Q4) was much greater than the rate of change during the other years. The rate of change between Q1 and Q4 for most measures was statistically significant; however, this was not the case for the other years. Increases in quality did slow down during the course of the Demonstration and by Year 4, quality scores increased very little, if at all. An example is shown below in **Figure ES-3**.

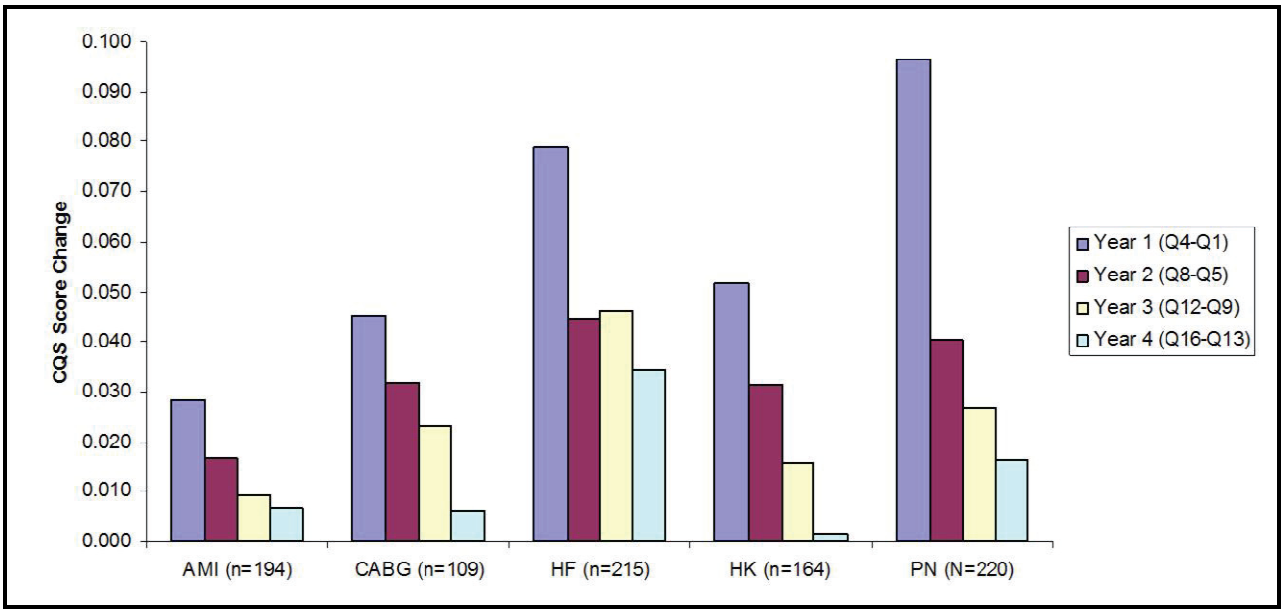
Figure ES-2
Percentage Change in CQS Measure between Q1 and Q16 [a]



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

Figure ES-3
Yearly Percentage Change in CQS Measures: Demonstration Years 1-4 [a]

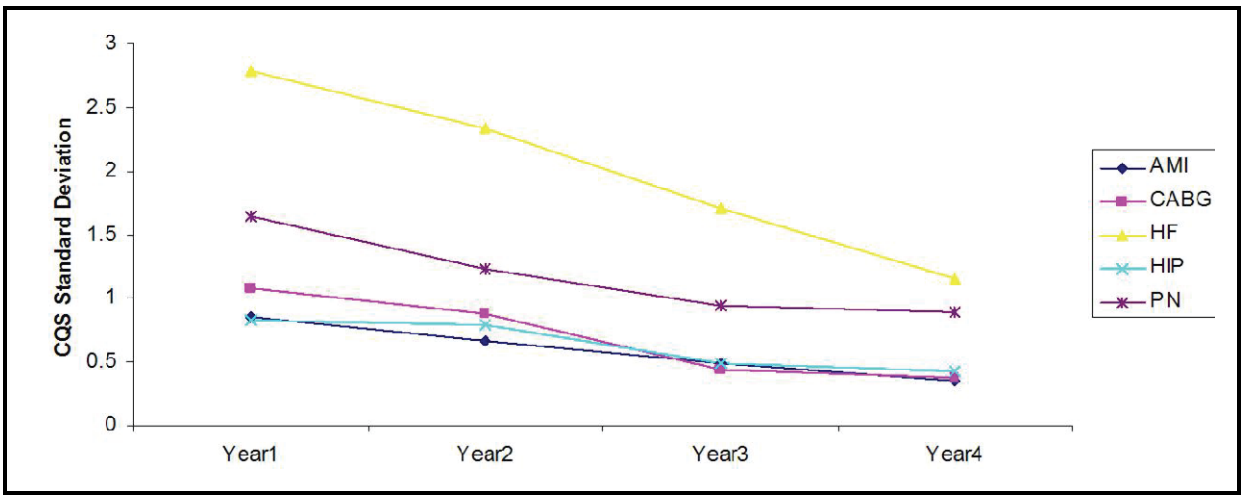


AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

Measure Variation over Time. The standard deviation for a measure provides an indication of the amount of variation in performance among Demonstration hospitals. Larger standard deviations indicate more measure variation across participating hospitals, and smaller standard deviations indicate less variation. We computed the average standard deviation for each measure between Year 1 and Year 4 to examine how measure variation changed over the course of the Demonstration. For the majority of measures, we found a negative percent change between Year 1 and Year 4, indicating that the standard deviation decreased over time and hospital performance was clustering around a mean value, without a wide range between high and low performing hospitals. **Figure ES-4** below displays the standard deviation in the CQS measures for each clinical focus area during Years 1 to 4 of the Demonstration.

Figure ES-4
Trends in Standard Deviation in CQS Measures: Demonstration Years 1-4



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

Incentive Payment Analyses

In addition to examining trends in the quality measures among Premier participants during Years 1 to 4 of the Demonstration, we also performed analyses related to the incentive payments. This analysis included the examination of how hospitals moved among deciles between Year 1 and Year 4, and how the decile thresholds changed over the course of the Demonstration.

Movement among Deciles. This analysis included all hospitals that had a clinical quality score in Year 1. High performers in Year 1 were hospitals in the top 2 deciles for a given clinical area, and low performers were hospitals in the bottom two deciles. This analysis was designed to address the following research questions:

- Are high performers in Year 1 likely to remain high performers in Year 4?

- Are low performers in Year 1 likely to remain low performers in Year 4?

- How does high performance or low performance during Year 1 affect Demonstration participation in Year 4?

Among the top performers in Year 1, we found that between 28% (HK) and 39% (PN) of hospitals in the top two deciles in Year 1 remained in the top two deciles in Year 4 (**Table ES-2**). In addition, hospitals in the top 20% in Year 1 were highly *unlikely* to drop to the bottom 20% by Year 4. Only 2% to 8% of hospitals dropped from the top two deciles in Year 1 to the bottom two deciles by Year 4. Few hospitals in the top 20% in Year 1 were *missing* a CQS score in Year 4. For example, only 4% of hospitals in the top two deciles for PN CQS in Year 1 were missing a Pneumonia CQS measure in Year 4. However, 15% of hospitals with in the top two deciles for CABG CQS in Year 1 no longer had a CABG CQS by Year 4. It must be noted that the number of hospitals reporting in the CABG clinical area is smaller than the other clinical areas (e.g., 134 for CABG vs. 243 for AMI), so that, although the number of hospitals missing a Year 4 CABG CQS score is consistent with the number missing CQS scores by Year 4 for other clinical areas (see Table ES-2), the overall percent of CABG hospitals missing a CABG year score by Year 4 is high relative to the other clinical areas (e.g., 14.8% for CABG vs. 8.2% for AMI).

Table ES-2
Hospitals in deciles 1 and 2 in Year 1 (top 20%): Comparison of decile ranking in Year 1 and Year 4

Clinical focus area	N hospitals with decile rankings Year 1	N hospitals in deciles 1 or 2 Year 1	Deciles 1 or 2 (top 20 %) Year 4	Deciles 3 to 8 (middle 60%) Year 4	Deciles 9 or 10 (bottom 20%) Year 4	Missing Year 4[a]
AMI	243	49	18 (36.7%)	24 (49.0%)	3 (6.1%)	4 (8.2%)
CABG	134	27	10 (37.0%)	12 (44.4%)	1 (3.7%)	4 (14.8%)
HF	259	52	15 (28.9%)	28 (53.9%)	4 (7.7%)	5 (9.6%)
HK	214	43	12 (27.9%)	26 (60.5%)	1 (2.3%)	4 (9.3%)
PN	261	52	20 (38.5%)	28 (53.9%)	2 (3.9%)	2 (3.9%)

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Missing data may be due to a variety of issues including: hospital withdrawal from the Demonstration, insufficient volume for the timeframe (<30 cases for the year), and/or hospitals failing validation.

Among the bottom performers in Year 1 (**Table ES-3**), we found that between 23% (HK) and 40% (CABG) remained in the bottom 20% of hospitals in Year 4. Relatively few hospitals improved enough to be in the top 2 deciles in Year 4; however 11% of both HK and CABG hospitals improved enough to move from the bottom 20% in Year 1 to the top 20% in Year 4. In general, for all clinical areas except HK, hospitals that were in the bottom 20% in Year 1 were much more likely to be missing from the Demonstration in Year 4, as compared with hospitals in the top 20% in Year 1. For example, between 25% (PN) and 35% (AMI) of hospitals that were in the bottom two deciles in Year 1 were missing CQS scores for those clinical focus areas by Year 4. We conclude from these results that hospitals that were the bottom performers early in the Demonstration seem more likely not to continue participating in the Demonstration by Year 4, as compared with hospitals that were top performers early in the Demonstration.

Table ES-3
Hospitals in deciles 9 and 10 in Year 1 (top 20%): Comparison of decile ranking in Year 1 and Year 4

Clinical focus area	N hospitals with decile rankings Year 1	N hospitals in deciles 9 or 10 Year 1	Deciles 1 or 2 (top 20 %) Year 4	Deciles 3 to 8 (middle 60%) Year 4	Deciles 9 or 10 (bottom 20%) Year 4	Not in Demonstration Year 4[a]
AMI	243	49	2 (4.1%)	15 (30.6%)	15 (30.6%)	17 (34.7%)
CABG	134	27	3 (11.1%)	4 (14.8%)	11 (40.7%)	9 (33.3%)
HF	259	52	4 (7.7%)	20 (38.5%)	14 (26.9%)	14 (26.9%)
HK	214	43	5 (11.6%)	15 (34.9%)	10 (23.3%)	1 (2.3%)
PN	261	52	3 (5.8%)	23 (44.2%)	13 (25.0%)	13 (25.0%)

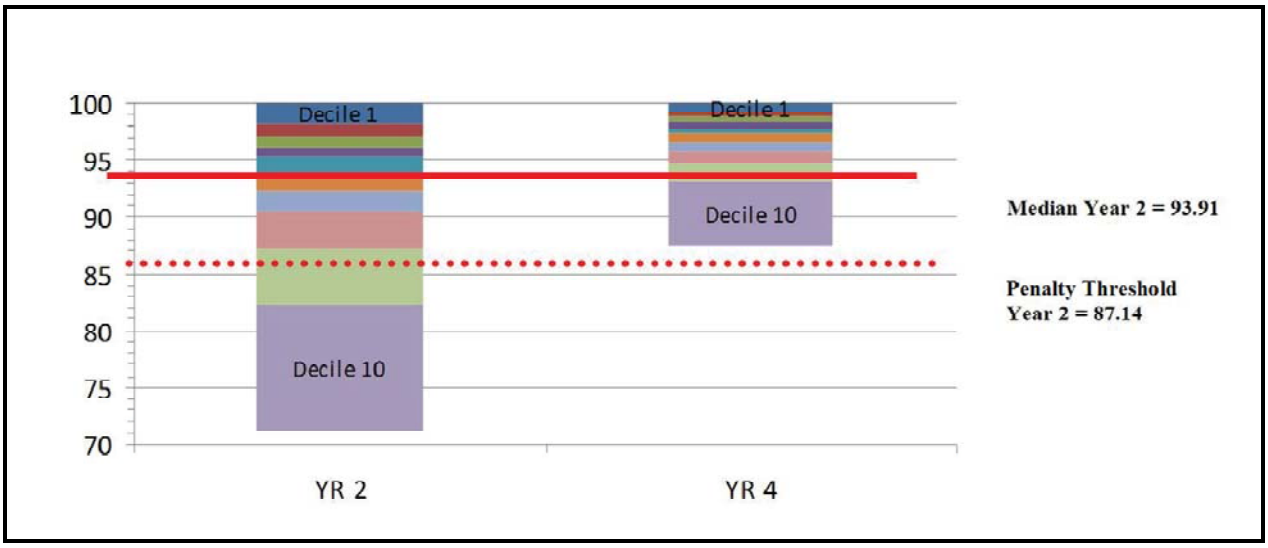
AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Missing data may be due to a variety of issues including: hospital withdrawal from the Demonstration, insufficient volume for the timeframe (<30 cases for the year), and/or hospitals failing validation.

Variation in Decile Thresholds. In this analysis, we compared the decile threshold ranges for Years 2 and 4. This analysis was performed using the decile thresholds that were provided for Years 1 to 4 on the Premier Web site: (http://www.premierinc.com/p4p/hqi/year4/decile-threshold-year-4.pdf). The decile threshold is the CQS that defines the upper limit of that particular decile. This analysis was designed to examine how the upper and lower values in each threshold changed over time.

We found that for all clinical focus areas, decile thresholds became increasingly compressed over the years of the Demonstration. An example of how the results were displayed is shown below for the Hip/Knee CQS Measure (**Figure ES-5**). In the graph we show stacked bars representing the deciles for Year 2 and Year 4. We also show a solid red line that represents the median for Year 2, because this value is used to determine which hospitals received the attainment award in Year 4. Finally, we also show a dotted red line that represents the value below which hospitals in Year 4 were penalized. Penalties in Year 4 were assessed based on the value of the ninth and tenth deciles in Year 2.

Figure ES-5
Demonstration: Decile ranges Year 2 and Year 4: HK [a]



[a] The figure depicts the reference deciles based on the Year 2 CQS scores. Beginning in Year 4, penalties are not assessed for hospitals scoring at least 85%.

In the Hip/Knee example, results indicate that decile thresholds compressed between Year 2 and Year 4. In Year 2, the upper limit for decile 10 was 82.36, and by Year 4 it increased to 93.04. The lower limit for receiving a high performer award (decile 2) was 97.13 in Year 2 and increased only to 98.81 in Year 4. The difference in threshold between decile 2 and decile 3 for HK (or the difference between receiving a high performer award and not receiving one) decreased dramatically between Year 2 (1.10) and Year 4 (0.34). By Year 4 there was very little difference between deciles 2 and 3 (in other words, very little difference between receiving a high performer award and not receiving one). This finding is consistent with the earlier finding showing that the standard deviations became smaller over time; performance among hospitals became less variable over time.

This figure also shows that the median value in Year 2 was 93.91. As the solid red line shows, this value in Year 4 indicates that virtually all hospitals received the attainment award, except for those in the tenth decile. The dotted red line indicates that hospitals in Year 4 that had a HK CQS score of 87.14 or below paid a penalty (no hospitals for HK in Year 4). However, beginning in Year 4, CMS set a cap of 85% so that hospitals that had a CQS score of at least 85% would not pay a penalty.

When examining the variation in decile scores across time, it is important to note that when measures change, historical thresholds are recalculated so that longitudinal comparisons are more accurate. For example, when the AMI PCI measure changed from 120 to 90 minutes, the threshold was recalculated to reflect the change.

Summary and Phase II Steps

Overall, our results show that quality measures showed significant increases between Q1 and Q16 of the Demonstration. Quality is improving across all clinical focus areas for process

and outcome measures as well as for composite quality scores. However, most of the gains in quality were seen during the first 2 years of the Demonstration while Year 4, in general, showed that early gains were sustained but not markedly improved.

There were a few measures that did not show increases between Q1 and Q16; these were measures that were very high at the start of the Demonstration (including mostly the outcome measures such as Survival index and Post-op physical/metabolic derangement avoidance index). For measures that were at or near the ceiling during Year 1, there was little room for improvement; however, in addition to improving, performance can be sustained or experience a decrease. These measures did not show decreases in performance over time. Measures that start out at such a high level of performance create a ceiling effect by influencing the mean as well as masking improvements in other measures that make up the composite.

For the majority of process, outcome, and CQS measures, improvements happened at a relatively fast rate during the early years (Years 1 or 2) of the Demonstration, and began to level off by Year 4. We found very few significant measure increases between Q13 and Q16. However, the CQS for both heart failure (HF) and pneumonia (PN) increased significantly during all years, including Year 4; although the rate of change during Year 4 was significantly smaller than the rate of change during Year 1, indicating a slowing in the rate of improvement. Because the Demonstration continued for 2 additional years after Year 4, our Phase II analysis will focus on whether hospitals are able to sustain this high level of performance and whether the leveling off in quality improvement continued for the remainder of the Demonstration.

Measure variation among hospitals is decreasing and hospital performance for a given measure is beginning to "cluster" with less of a difference between the high and low performers; everyone is getting better. This is a positive finding, in that less variation in measures over time indicates that participating hospitals are improving. Phase II analyses will continue to examine the trends in measure standard deviations over the course of the Demonstration to assess whether Demonstration hospitals reach a minimum standard deviation such that further decrease in variation is unlikely.

A key finding from this analysis is that hospitals that were in the bottom 20% in Year 1 were more likely to be missing from the Demonstration in Year 4, as compared with hospitals in the top 20% in Year 1. This finding has implications for the evaluation of the Demonstration since those experiencing low performance are more likely to drop out of participation and not place themselves at risk for penalties associated with poor performance. The finding clearly indicates that self-selection is occurring; hospitals that were high performers early in the Demonstration are highly likely to remain in the Demonstration throughout (and thus push the quality scores higher over time), whereas hospitals that were low performers are likely to drop out (of the 266 participating hospitals in Year 1, 220 hospitals (82.7%) participated in all Years 1 through 4).

Decile thresholds increased during the Demonstration years for all clinical focus areas, indicating that quality is improving. Furthermore, the decile thresholds became compressed over time, such that in Year 4 there is often little difference in the threshold that would place a hospital in one decile or another. We will continue to examine this phenomenon in Phase II of

the evaluation, to determine if the trend continues (i.e., differences between deciles get smaller and smaller).

The expansion of payment opportunities to include the attainment award resulted in substantially more hospitals receiving rewards in Year 4 compared with the previous years. Hospitals in Year 4 must have reached the median set in Year 2 to receive the attainment award, and our results reveal that the median in Year 2 was often close to the ninth or tenth decile in Year 4 (because quality scores increased dramatically over time). Therefore, for all clinical focus areas, a large proportion of hospitals (often more than 90%) received the Attainment Award bonuses. Phase II of the Demonstration evaluation will examine whether this trend continues. Because our trend analyses show that quality scores are improving at a much slower rate in Year 4 compared with previous years, we may find that not as many hospitals are rewarded for median attainment in Year 6 if the rates of improvement continue to slow between Year 4 and Year 6.

In Phase II of the analyses, we will continue to examine trends in quality measures for the full 6 years of the Demonstration and to examine whether the rates of improvement are indeed slowing, as they appear to be based on analysis of Years 1 through 4. Key analyses to be performed in Phase II will include comparison hospitals that did not participate in the Demonstration to examine whether Demonstration participants improved their quality measures during the period of the 6 Demonstration years significantly more so than the comparison hospitals. Other key analyses to be performed include incorporating Medicare claims data to examine whether improvements in quality of care, as measured by the process, outcome, and CQS measures in the Demonstration, affect Medicare beneficiary acute care inpatient hospital length of stay and mortality for the conditions included in the Demonstration.

SECTION 1
BACKGROUND OF THE PREMIER HOSPITAL QUALITY INCENTIVE DEMONSTRATION (PHQID)

In July 2003, the Centers for Medicare & Medicaid Services (CMS) contracted with Premier, Inc., to launch a hospital pay-for-performance Demonstration whereby hospitals would receive a financial bonus for high-quality care and be at risk for paying penalties for poor-quality care that has not improved. Called the Premier Hospital Quality Incentive Demonstration (PHQID or Demonstration), this national pay-for-performance project was designed to determine whether economic incentives are effective at improving the quality of inpatient hospital care. The initial Demonstration ran for 3 fiscal years, beginning on October 1, 2003, and ending on September 30, 2006; the Demonstration was then extended for an additional 3 years beginning October 1, 2006, and ending September 30, 2009.

The Demonstration, managed by Premier, Inc., was open to hospitals that were reporting quality measures through Premier's Perspective™ quality measurement program. Premier, Inc. is a health care purchasing and service company that is owned by a collection of hospitals and systems; in 2004, at the start of the Demonstration it was owned by 203 entities and had 1,418 members. Perspective™ is a reporting system marketed by Premier that includes data on the quality measures included in the PHQID. Prior to the start of the Demonstration, the 444 hospitals that were reporting under the Perspective system were offered participation in the Demonstration; more than 250 hospitals chose to participate with participants located in 37 states. During the first 3 years of the Demonstration, approximately 17% of the participants were located in rural regions with 83% in urban locations, and approximately 14% were Council of Teaching Hospital members. Hospitals participating in the first 3 years of the Demonstration were offered the opportunity to continue for another 3 years; 227 hospitals elected to participate starting in Year 4 of the Demonstration. This 6-year project tracked hospital-specific performance on a set of standardized and widely accepted clinical quality indicators as well as additional quality measures that were tested over the course of the Demonstration. During Demonstration Years 1–4, quality indicators and test measures were tracked for five clinical conditions; an additional clinical condition was added for Years 5 and 6. Details on the quality measures are presented in Section 1.1 below.

PHQID was designed to test the effects of pay-for-performance incentives on the quality of hospital care. To motivate improvements in quality, the Demonstration included an evolving array of annual payments beginning with payments to top performers in each of five clinical areas. The addition of penalties for poor performers in each of five clinical areas began in the third year of the Demonstration. For the second 3 years of the Demonstration, the pay-for-performance incentives changed and a sixth clinical condition was added.

In this Phase I Report, we examine trends for Years 1–4 of the Demonstration focusing on trends in quality within the Premier Demonstration hospitals that remained in the Demonstration for all of Year 1–Year 4. The report is organized as follows: this section provides an introduction as well as a summary of the evaluation of the first 3 years of the Demonstration. Section 2 describes the methodology used to perform the various analyses. Section 3 describes the trends in the quality measures for which we had data on all 16 quarters of Year 1 through Year 4 (including those hospitals that reported measures in all 16 quarters). Section 4 describes

the results of the Pay for Performance in Year 4, including analyses of the hospital decile thresholds and the payments that were made to hospitals in Year 4. Section 5 describes the test measures and the other measures for which we had data for fewer than 16 quarters. Finally, Section 6 presents an overview of the findings and conclusions as well as directions for next steps in the evaluation.

1.1 Pay-for-Performance

PHQID was designed to test the effects of pay-for-performance incentives on the quality of hospital care. To motivate improvements in quality, the Demonstration included an evolving array of annual payments beginning with payments to top performers in each of five clinical areas; the addition of penalties for poor performers in each of five clinical areas (in the third year of the Demonstration); the addition of a sixth clinical area (as a test measure in the fourth year of the Demonstration and as part of the calculation for annual payments in the fifth year of the Demonstration); and additional incentive opportunities starting the fourth year of the Demonstration. Incentive payments made by CMS totaled $24.6 million during the first 3 years, while penalties in Year 3 totaled less than $104,000.

The six clinical areas examined for payment purposes are listed below, with the first five included during the first 3 years and the last tested during Demonstration Year 4 and incorporated into payment models beginning in Year 5. The clinical areas include:

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

Inpatients were categorized into clinical areas using primary and secondary diagnostic and procedure codes. Although patients could be categorized in more than one clinical area, during the first 3 years such categorization was unusual. Each clinical area included between 4 and 9 area-specific quality measures; performance in each clinical area was assessed based on composites of the scores for the relevant area-specific performance measures. During the first 3 years of the Demonstration, CMS made incentives to top performers in each clinical area in each year; in the third year, CMS began imposing financial penalties on poor performers. Beginning in Year 4, in addition to giving incentives to top performers CMS also gave Attainment Awards, by clinical area, to hospitals that attained or exceeded median-level performance and improvement awards to hospitals that achieved attainment and were among the top 20% of hospitals with the largest improvement; CMS also continued to penalize poor performers. In addition, in Year 4 CMS began to publish the scores for all participating hospitals on the CMS Web site.

With the exception of HK,[1] admissions eligible for inclusion included all adult inpatients, regardless of payer, with the clinical-area specific diagnostic or procedure codes. Not all Demonstration hospitals had patients in each of the clinical areas. The numbers and percentages of Demonstration hospitals with admissions in a given clinical area during the first 3 years ranged from 50% (CABG) to nearly 100% (PN) (Kennedy et al., 2008).

During the first 3 years, the Demonstration included a total of 34 quality measures for the five clinical areas (see **Table 1**). The number of clinical area-specific measures ranged from a low of four for HF to a high of nine for AMI. Clinical area specific measures included both process and outcome measures, and were taken from measures developed by CMS and its Quality Improvement Organizations (QIO), the Joint Commission core measures, the Hospital Quality Alliance, and the Agency for Healthcare Research and Quality (AHRQ) patient safety indicators, among others.

Measures were revised and/or suppressed during the course of the Demonstration for consistency and alignment with National Hospital Quality Inpatient Measures (NHQIM), CMS, and JCAHO guidelines. For example, "Appropriate prophylactic antibiotic selection" measure for both the CABG and HK clinical areas was suppressed effective October 1, 2004, Year 2 of the Demonstration. The concerns with the measure involved the increasing prevalence of MRSA, national shortages of antibiotics for colorectal and hysterectomy procedures, and clinical evidence recommendations for prevention of endocarditis. The measure was suppressed to align with CMS and JCAHO, and was unsuppressed effective July 1, 2006, Year 3 of the Demonstration. Similarly, the "influenza vaccination" for PN was suppressed during Year 2 of the Demonstration due to the influenza vaccine shortage during the 2004–2005 influenza season; CMS and JCAHO suppressed this measure, and to maintain alignment with the National Hospital Quality Measures, the Demonstration project incorporated this measure suppression in Year 2. This measure was unsuppressed and resumed use starting in Year 3 of the Demonstration. Also, "Isolated CABG using Internal Mammary Artery (IMA)" for the CABG clinical area was suppressed during the first 4 years of the Demonstration, because several exclusionary diagnosis codes were omitted from the original measure specification. For all suppressed measures, trends over Demonstration Year 1–Year 4 are not examined in this report.

For the CABG clinical area only, "prophylactic antibiotics discontinued after surgery end" was revised from 24 hours after surgery end to 48 hours after surgery end, effective with discharges after January 1, 2006 (during Year 3 of the Demonstration). In addition, for the AMI and HF clinical areas, "ACEI for LVSD" was revised to be named "ACEI or ARB for LVSD" effective with January 1, 2005 discharges (Year 2, Quarter 2 of the Demonstration). Detailed information regarding the measure revisions, suppression, and alignment can be found at http://www.premierinc.com/quality-safety/tools-services/p4p/hqi/specifications-by-focus-area-year1-3.jsp.

[1] HK included post hospitalization outcome measures such as readmission to an acute care hospital within 30 days so that only Medicare patients were included in those measures for this condition during Years 1–3 of the Demonstration. However, beginning in Year 4, this measure was inclusive of all payors and was not restricted to Medicare patients.

Table 1
Demonstration measures

N	Condition	Measure	Measure type
1	AMI	Aspirin at arrival	Process
2	AMI	Aspirin prescribed at discharge	Process
3	AMI	ACE or ARB for LVSD [a]	Process
4	AMI	Adult smoking cessation advice/counseling	Process
5	AMI	Beta blocker prescribed at discharge	Process
6	AMI	Beta blocker at arrival	Process
7	AMI	Thrombolytic agent within 30 minutes of hospital arrival	Process
8	AMI	PCI within 120 minutes of hospital arrival	Process
9	AMI	Inpatient mortality	Outcome
10	AMI	AMI composite	Composite
11	AMI	Total appropriateness of care [b]	Composite
12	CABG	Aspirin prescribed at discharge	Process
13	CABG	CABG using internal mammary artery (IMA) [c]	Process
14	CABG	Prophylactic antibiotics received within 1 hour prior to surgical incision	Process
15	CABG	Prophylactic antibiotics selection for surgical patients	Process
16	CABG	Prophylactic antibiotics discontinued within 24/48 hours after surgery end time [d]	Process
17	CABG	CABG mortality (APR-DRG mortality risk adjustment)	Outcome
18	CABG	Postoperative hemorrhage or hematoma	Outcome
19	CABG	Postoperative physiologic and metabolic derangement	Outcome
20	CABG	CABG composite	Composite
21	CABG	Total appropriateness of care [b]	Composite
22	HF	LVF assessment	Process
23	HF	Detailed discharge instructions	Process
24	HF	ACE or ARB or LVSD [a]	Process
25	HF	Adult smoking cessation advice/counseling	Process
26	HF	HF composite	Composite
27	HF	Total appropriateness of care [b]	Composite
28	PN	Oxygenation counseling within 24 hours	Process
29	PN	Initial antibiotic selection consistent with current recommendations—intensive care unit (ICU)	Process
30	PN	Initial antibiotic selection consistent with current recommendations—non-ICU	Process
31	PN	Blood cultures collected priori to first antibiotic administration	Process
32	PN	Influenza screening/vaccination	Process

(continued)

Table 1
Demonstration measures (continued)

N	Condition	Measure	Measure type
33	PN	Pneumococcal screening/vaccination	Process
34	PN	Initial antibiotic received within 4 hours of hospital arrival	Process
35	PN	Adult smoking cessation advice/counseling	Process
36	PN	PN composite	Composite
37	PN	Total appropriateness of care [b]	Composite
38	HK	Prophylactic antibiotics received within 1 hour prior to surgical incision	Process
39	HK	Prophylactic antibiotics selection for surgical patients	Process
40	HK	Prophylactic antibiotics discontinued within 24 hours after surgery end time	Process
41	HK	Postoperative hemorrhage or hematoma	Outcome
42	HK	Postoperative physiologic and metabolic derangement	Outcome
43	HK	Readmissions 30 days postdischarge (APR-DRG severity adjustment)	Outcome
44	HK	HK composite	Composite
45	HK	Total appropriateness of care [b]	Composite
46	SCIP [e]	Prophylactic antibiotics received within 1 hour prior to surgical incision	Process
47	SCIP [e]	Prophylactic antibiotics selection for surgical patients	Process
48	SCIP [e]	Prophylactic antibiotics discontinued within 24 hours after surgery end (48 hours for CABG or other cardiac surgery)	Process
49	SCIP [e]	Surgery patients with recommended venous thromboembolism prophylaxis ordered	Process
50	SCIP [e]	Surgery patients who received appropriate VTE prophylaxis within 24 hours Prior to surgery up to 24 hours after surgery end time	Process
51	SCIP [e]	SCIP composite	Composite
52	SCIP [e]	SCIP Total Appropriate Care Score (ACS) [b]	Composite

ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; APR-DRG = all patient refined diagnosis related group; ARB = angiotensin receptor blocker; CABG = coronary artery bypass graft; HF = heart failure; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention; PN = pneumonia; HK = hip/knee replacement; SCIP = surgical care improvement project.

[a] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

[b] Appropriateness of Care Scores (ACS) are test measures, not considered for payment incentives. Other test measures examined in Years 5 and 6 of the Demonstration will be addressed in future reports.

[c] This measure was suppressed during the first 4 years of the Demonstration.

[d] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

[e] SCIP measure began being used for payment incentives in Year 5 of Demonstration.

Scores for process measures reflected the proportion of patients whose treatment conformed to certain recommended practices.[2] Only three of the clinical areas also had outcome measures (AMI, CABG, and HK). The outcome measures included the incidence of adverse events either during the inpatient stay (for AMI and CABG) or within a short time following discharge (HK).[3] In addition to process and outcome measures, clinical-condition specific composite scores were constructed using the individual process and outcome scores for each of the clinical areas. In Year 4 of the Demonstration SCIP, a sixth clinical area was tested and became part of the payment incentive calculations in the fifth year.

Beginning in Year 4 of the Demonstration, CMS requested that the Demonstration collect and test additional measures. Although monitored by the Demonstration and included in the evaluation, test measures are not part of the financial incentives and penalties calculations. In this Phase I evaluation, we provide a descriptive analysis of the test measures as they come available, and as more than 1 year of data become available for a measure (usually during the first year of Phase II), we will trend the measures.

Incentive payments and penalties were determined separately for each clinical area. Performance was assessed for each area based on a hospital's annual area-specific composite quality score.

1.1.1 Financial Incentives for Demonstration Years 1–3

Incentives for the first 3 years of the Demonstration were paid to hospitals whose composite quality score (CQS) for a given clinical area was in the first or second decile of scores for participating hospitals. An incentive payment was equal to 2% of the hospital's basic Medicare reimbursements for patients in that clinical area for hospitals scoring in the first decile. Hospitals whose CQSs were in the second decile (the top 20% to 11%) for that clinical area received an incentive payment equal to 1% of the hospital's basic Medicare reimbursements for patients in that clinical area. Beginning in Year 3, penalties were applied to hospitals whose composite scores fell below thresholds established in the first year based on the scores for the hospitals scoring in the bottom two deciles.

1.1.2 Years 4–6

Whereas hospitals received incentive payments when their composite scores were in the top 20% for a given fiscal year during the first 3 years of the Demonstration, beginning in Year 4, the incentive system was changed substantially. Incentives were awarded based on threshold attainment, top performance, and significant improvement:

[2] These recommendations were subject to various exceptions. The process measures were a calculation of the measure of relevant cases whose treatment conformed to the recommended guidelines over the number of eligible patients, after exclusions.

[3] These included inpatient mortality rates (AMI and CABG), surgical complications (CABG and HK), and 30-day readmissions (HK). As with process measures, some patients may be excluded from the denominator.

- **Median-Level Attainment Award:** Hospitals that attain or exceed the median level as measured 2 years prior CQS in a given clinical area, will receive an incentive payment.

- **Top Performer Award:** Hospitals that have a CQS for a given clinical area that falls in the top 20% of scores will receive an additional payment. These hospitals will receive the median-level attainment award as well.

- **Top Improvement Award:** Hospitals that attain median-level performance and are among the top 20% of hospitals with the largest percentage quality improvement in each clinical area will receive an additional incentive payment. Improvement will be calculated based on the performance year compared to 2 years prior. This group will receive the median-level attainment award as well. Top performers are not eligible for the improvement award.

Furthermore, in Years 4–6, the penalty thresholds were allowed to adjust every year, so that penalties in one Demonstration measurement year were applied to hospitals whose composite scores fell below thresholds established 2 years prior. However, a maximum penalty threshold was set, above which a penalty would not be assessed no matter where the 9th and 10th reference deciles fell. An absolute maximum required attainment to avoid penalty was set at a score of 85%, which represented the highest reference threshold in the first year of the Demonstration (AMI 9th decile threshold). Therefore, although the 9th decile reference threshold may be higher than 85% in Year 2, hospitals with attainment scores above 85% in measurement Year 2 were not penalized.

1.2 Summary of Results from Years 1–3 Evaluation

In addition to findings from the Abt report (Kennedy et al., 2008), results were also published on the Premier Web site as well as in the peer-reviewed literature (Lindenauer et al., 2007; Ryan, 2009). In this section, we summarize the findings from the Abt evaluation of the first 3 years of the Demonstration.

The evaluation described the changes in the quality measure scores over the first 3 years of the Demonstration. It also reported on the portion of the changes in the scores that could be attributed to the Demonstration as well as the effects of the Demonstration on Medicare reimbursement and outlays and Medicare beneficiary average length of stay. Results of the evaluation suggested that quality increased considerably among hospitals that participated in the Demonstration. However, quality also increased substantially among hospitals that posted their scores on the Hospital Compare Web site and who were not part of the Demonstration. The evaluation could only examine the Demonstration effect for three conditions for which comparison data were available (AMI, HF, and PN). Among these measures, participation in the Demonstration contributed to a small portion of the increase in quality (between 10% and 18% of the total change in quality). The evaluation results also suggested that the Demonstration hospitals did not experience savings associated with reductions in length of stay that could be attributed to the Demonstration. In general, during the first 3 years of the Demonstration, length of stay decreased over time for both Demonstration and comparison hospitals. Finally, the Abt report concluded that since the PHQID did not appear to reduce Medicare reimbursement, it is

unlikely that the Demonstration was budget-neutral—in the aggregate, payments were made to high-performing hospitals, net of penalty payment and in the absence of savings that could be attributed to the Demonstration (Kennedy et al., 2008). Complete budget neutrality for the Medicare program was not explored since differences in post-acute care were not examined.

The current evaluation extends the evaluation to all 6 years of the Demonstration in two Phases. The Phase I Report provides results of the analysis of Demonstration data from Years 1–4; Phase II will extend the analysis to all 6 years of the Demonstration.

1.3 Research Questions

The research questions, including subquestions, posed by CMS are provided in **Table 2**. The questions examine whether the Demonstration

1. had a positive impact on the quality of care provided,

2. resulted in quality improvements over time that exceeded what may have occurred in an overall national focus on improved quality of care, and

3. had a positive impact on Medicare beneficiaries.

The Phase I analysis addresses the first of these three focal areas—whether the Demonstration had a positive impact on the quality of care provided. In addition, we also examine trends in payments for performance, and given the change in payment scheme, the distribution of payments by type of award.

1.3.1 Quality of Care/Outcomes

The Demonstration utilized established measures of process and outcomes of care reflecting quality in five clinical areas during the first 3 years and six clinical areas during the latter 3 years. The two key summary quality of care scores used in this Demonstration were the Composite Quality Score (CQS) and the Appropriate Care Score (ACS).The individual clinical area measures are the basis for developing a Composite Quality Score (CQS) using a standardized methodology for each clinical area. Hospital quality scores are computed based on the entire patient population who met the Demonstration criteria. Although a completely separate (and test) measure, the Appropriate Care Score (ACS) is computed as follows: patients receiving all the recommended processes for a given clinical area have a score of 1; the score otherwise the score is 0. Premier calculated the ACS for each of the conditions for the hospitals participating in the Demonstration.

1.3.2 Medicare Outcomes

The PHQID provided incentive payments to hospitals that demonstrated high-quality care and imposed payment penalties for hospitals that did not demonstrate improvement from poor baseline performance (lower two deciles). In Phase I, we examine the distribution of incentive payments; in subsequent reports (for Phase II), we will examine the differential impact of the Demonstration on the quality of care measures as well as on Medicare beneficiary length of stay and mortality.

Table 2
Research questions

Quality of care/outcomes

1. For participating hospitals, how did measuring, advertising, and paying for quality care improve the quality of care for the conditions included in the Demonstration? [a]
   a. Did one clinical area show a clear improvement over another?
   b. Was improvement in one clinical area correlated with improvements in the other clinical areas or was each clinical area independent of the other areas?
   c. Do hospitals that show improvements in process measures also demonstrate improvements in outcome measures?
   d. How much of the improvement in quality can be attributed to the Premier Demonstration? [b]
2. How did quality scores change over time? [a]
   a. How did the distribution of composite quality scores for the participating Premier hospitals change in each clinical area during the Demonstration?
   b. What are the characteristics of hospitals that demonstrate improvements over time and what are the characteristics of hospitals that do not?
   c. Do these differ for participating and for comparison hospitals? [b]
3. How did the different incentive payment systems used during the initial and extension phases of the Demonstration impact distribution of incentive payments? [c]
   a. Does rewarding a threshold (median-level attainment) raise the median?
   b. Does rewarding hospitals for improvement above their own prior performance result in larger percentage improvement?
   c. How does the distribution of deciles change over time?
4. Which did hospitals withdraw from the Demonstration over the 6-year period? What was the relationship of withdrawals to quality performance and incentive payments? [d]
   a. Are lack of improvement and lack of incentives associated with withdrawal?
   b. Do including the incentives for median level attainment and improvement provide an added incentive to continue participation?

Medicare outcomes (2 research questions)

1. Did changes in the quality of care affect Medicare beneficiary acute-care inpatient hospital length of stay (LOS) and mortality for the six clinical conditions included in the Demonstration? [e]
2. What were the effects of the demonstration on Medicare beneficiary utilization and Medicare reimbursements for all 6 years of the demonstration? ***(to be addressed in Optional Task)***

[a] Will use data from Demonstration hospital only (no comparative analysis will be performed using control hospitals).

[b] Related to the secular trend of a national focus on quality, and will use control hospitals for analysis (from Hospital Compare).

[c] During the first 3 years of the program, those in the top decile and second decile were rewarded. Starting in the fourth year, incentives were awarded based on threshold attainment, top performance, and significant improvement. Threshold penalties for hospitals that did not score above the ninth decile remain.

[d] Hospitals that withdrew from the Demonstration will not be contacted and asked about their decision to withdraw. Rather, RTI will examine the association among hospital characteristics, quality performance and incentive payments, and the timing of hospitals' withdrawal from the Demonstration.

[e] We propose to examine both inpatient mortality and 30-day mortality to capture those that may have been discharged and then died either at a post-acute care (PAC) setting, at home, or during a related inpatient readmission.

SECTION 2
METHODOLOGY

This section summarizes the analytic approach used for the Phase I analyses, which focus on trends in quality within the Demonstration hospitals. No comparison group is used for this phase of the analysis, and no Medicare claims data are used for analyses that go beyond the inpatient experience (e.g., readmissions or 30-day mortality), with the exception of clinical areas that use that data in construction of measures.

2.1 Analytic Framework

The general analytic framework used is a quasi-experimental design. Demonstration impacts are quantified using trend and difference-in-difference approaches, depending on the outcome variable. Trend analysis is one way to assess ceiling effects for particular measures, sustainability of performance, how rapidly improvement is obtained, and whether pattern of performance is comparable among clinical areas and among measures. Thus, trend methods are conducted for individual measures as well as composite measures thus enabling comparisons among measures and composites. This method is useful when control group data are unavailable and when control groups are available. However, the flaw in this approach is that other site-specific factors may influence the size or nature of the impacts over time. We propose to use trend analyses to illustrate changes over time in quality measures for the hospitals that participated in the Demonstration and to assess the distribution changes in hospital deciles, which are used to determine Demonstration bonuses and penalties.

2.2 Generalizability

There are several threats to generalizability from this Demonstration. First, hospitals opted to participate in the initial 3 years of the Demonstration and then opted to remain for the continuation of the Demonstration beyond the first 3 years; the self-selection of enthusiastic hospitals may result in hospitals that have been more successful in improving quality or are more focused on improving quality of care than other similar hospitals. Another threat to generalizability is the services provided by Premier. In addition to managing the Demonstration, Premier offered services to participating hospitals such as shared learning among participants, communicating successful interventions, providing on-site consultation and other services. This analysis cannot disaggregate the impact of financial incentives versus the impact of the services provided by Premier.[4] Therefore, it is not clear to what extent the findings can be generalized to a pay-for-performance approach that does not include the extensive support provided by Premier.

[4] RTI and CMS learned of a group of hospitals that "shadowed" the Demonstration hospitals. The shadowing hospitals participated in the interventions provided by Premier; however, they did not participate in the financial incentives provided by CMS. We explored the possibility of using the shadowing hospitals to disaggregate the incentive effect from the support effect; however, it is not likely to be feasible. These hospitals change during the course of the Demonstration, and Premier would not be able to provide us with data on a consistent cohort of shadow hospitals.

2.3 Data Source

For this Phase I analysis, RTI used data from Premier on the Demonstration participants. The specifics and years of data are discussed below.

2.3.1 Premier Demonstration Data

To generate the quality scores for participating Demonstration hospitals, inpatients at the hospital are categorized into a clinical area using certain primary and secondary diagnostic and procedure codes. For most measures, all adult inpatients with diagnostic or procedure codes for a given clinical area were included in the measure. However, for the HK measures, during Demonstration Years 1–3, outcomes analyses included only Medicare beneficiaries; non-Medicare patients were excluded from outcomes analyses during this time. This changed in Demonstration Years 4–6, when all patients were included for HK measures with the exception of the HK readmissions measure; this measure included only Medicare beneficiaries. This change aligned the other measure populations with the SCIP population criteria.

Demonstration hospitals do not necessarily have patients in each of the five clinical areas. Only about half of the Demonstration hospitals performed coronary artery bypass grafts (CABG) in each of the first 3 years. Most performed HK or had acute myocardial infarction (AMI) admissions, and almost all had heart failure (HF) and pneumonia (PN), in all 3 years.

In Demonstration Years 1–3, there were a total of 34 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 1. The measures were taken from measures developed by CMS and its Quality Improvement Organizations (QIOs), from the Joint Commission core measures, from the Hospital Quality Alliance, and from the Agency for Healthcare Research and Quality (AHRQ) patient safety indicators (PSIs), among others.

The Premier data file contained key hospital quality measures, the process and outcome measures that are part of the PHQID, as well as other organizational (hospital) indicators such as hospital ownership and teaching status. Premier hospital data are provided for the evaluation analysis at the hospital level and therefore do not indicate whether a given patient received the condition-specific intervention included in the process measures nor information on the characteristics of the individuals that make up the quality scores. The quality data are validated by CMS. The Demonstration quality measures include both process and outcome measures, as well as composite measures, described below. Composite quality scores are calculated annually for each Demonstration hospital by rolling up individual measures into an overall quality score, by hospital, for each clinical condition.

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 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 (postoperative hemorrhages or hematomas, and postoperative 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 the hospital's control.

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 3 years.

Composite Scores

Composite scores, or composite quality scores (CQSs), are constructed from the individual process and outcome scores for each of the clinical areas and consist of both the Composite Process Score (CPS) and the Composite Outcome Score (COS), for clinical areas that have outcome measures. A patient represents many opportunities for evidence-based interventions that can be measured by the performance indicators in this Demonstration. The Composite Process Scores (CPS) are developed for each disease category by dividing the number of achieved interventions by the total number of opportunities for the same targeted interventions. The Composite Outcome Score (COS) is calculated based on the hospital's actual mortality or adverse event rate and the expected mortality or adverse event rate derived from adjusting the actual rate for the presence of various risk factors. The observed and risk-adjusted mortality rates are transposed to create a survival index. The observed and risk-adjusted adverse event rates and the observed and risk-adjusted readmission rates are transposed to create an avoidance index. If a hospital does not have any patients eligible for an outcome measure, that hospital's weights are modified and adjusted down by each missing outcome measure.

After the weights are applied to both the CPS and COS components, a composite score for each of the five clinical conditions is calculated by adding together the CPS and the COS.

The resulting CQS is used to identify hospitals participating in the Demonstration for incentive payments. Proportional weighting is used to account for the relative contribution of each CQS component. AMI has nine total indicators, eight of which are process indicators. Therefore, a weighting factor of (8/9) is applied to the process indicators, and 1/9 is applied to the one outcome indicator. Each Demonstration participant's CQS is provided to RTI with the other measures by Premier, Inc.

Appropriateness of Care

Appropriateness of care scores are constructed from the individual process scores for each clinical area. The score represents the proportion of patients that received all of the measured interventions for which they were eligible in a given clinical area. A hospital receives an appropriateness-of-care score for each of the clinical areas for which they report data.

Trend Analysis

Most of the analyses performed for the Phase I evaluation involved trending and examining changes over time between Q1 (first quarter of the Demonstration) and Q16 (last quarter of Year 4 of the Demonstration). We examined within-clinical area differences in measures over time and performed t-tests for differences in means between Q1 and Q16. We also examined differences in rate of change between years by looking at the difference between Q1 and Q4 (i.e., the rate of change during Year 1) and comparing that statistically to the difference between Q13 and Q16 (i.e., the rate of change during Year 4). In addition to the mean scores, we also examined measure standard deviations over time to assess whether variation hospital performance is increasing or decreasing over Year 1 to Year 4, and to examine whether the range between high and low performers is shrinking. When performing t-tests of significance, we accounted for multiple testing using the Bonferonni correction.

SECTION 3
PARTICIPATING HOSPITALS TRENDS IN QUALITY FOR DEMONSTRATION YEARS 1 TO 4

3.1 Introduction

This section of the report describes the trend analyses performed using hospitals that had measures in all 16 quarters of the Demonstration Year 1 to Year 4, including differences in measures over time and t-tests for differences in means and for differences in rates of change between Year 1 and Year 4. This section first presents the process and outcome measures for each clinical focus area: acute myocardial infarction (AMI); coronary artery bypass graft (CABG); heart failure (HF); hip/knee replacement (HK); and pneumonia (PN), followed by the Composite Quality Scores (CQS) for each clinical focus area.

For each clinical focus area, the difference in the number of hospitals (N's) by measure is largely due to the fact that not all hospitals provided care for each clinical area. For example, more hospitals provide care for PN than provide CABG surgeries. Therefore, the number of hospitals for the PN measures is greater than the number of hospitals for the CABG measures.

Note that this chapter presents analyses that focus on changes in quality that may not be due strictly to the Demonstration alone. Because we are not presenting a comparative analysis using hospitals that did and did not participate in the Demonstration, we cannot attribute the changes in quality scores described in this chapter directly to the Demonstration. This chapter provides the trends in quality scores over time for hospitals that participated in the Demonstration. Analyses in future years will incorporate comparison hospitals to parse out the demonstration effect on quality.

3.2 Acute Myocardial Infarction (AMI)

This section presents results for the analyses of the process and outcome measures for AMI that include hospitals reporting 16 quarters of data. **Table 3** shows the measures and the number of hospitals included in the analyses (the number of hospitals with 16 quarters of data, ranged from 112 to 193 depending on the measure). A total of 112 hospitals reported 16 quarters of data on PCI received within 120 (90) minutes of arrival, whereas 193 hospitals reported 16 quarters of data on Beta blocker at arrival.

The AMI measure, Thrombolytic agent within 30 minutes of hospital arrival, is not included in these analyses because only four hospitals reported this measure for all 16 quarters of Years 1–4. It is also important to note that the PCI measure for the AMI population changed definition during the third year of the Demonstration; from administration within 120 minutes after arrival to change within 90 minutes after arrival.

Table 3
Demonstration: AMI measures for hospitals reporting results, Q1–Q16

AMI measure	N hospitals with measures in all 16 quarters
Adult smoking cessation advice/counseling	138
Aspirin at arrival	192
Aspirin prescribed at discharge	183
ACE or ARB for LVSD[a]	135
Beta blocker prescribed at discharge	188
Beta blocker at arrival	193
PCI received within 120 (90) minutes of arrival	112
Survival index	192

ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention

[a] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

3.2.1 AMI Measure Means

Table 4 shows the means for each AMI measure across the first 16 quarters of the Demonstration (Years 1–4). **Figure 1** details in graphical form the trends in AMI measures across the 16 quarters. As shown in Table 4 and Figure 1, the mean for each AMI measure increased during the first 4 years of the Demonstration. For all measures except the survival index, the increases were more pronounced during the first couple of years and the leveled off during Years 3 and 4. The increase for the survival index measure was very high to begin with and remained high throughout the 4 years. As Figure 1 shows, the measures for the use of aspirin at admission and discharge were quite high at the start of the Demonstration, leaving little room to detect significant improvement over time; however, performance gains were observed and sustained during Year 4.

Table 4
Demonstration: AMI measures for Years 1–4, [a, b] means

AMI measure	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Adult smoking cessation advice/counseling (n=138)	0.804	0.837	0.854	0.895	0.912	0.939	0.960	0.969	0.978	0.983	0.982	0.990	0.992	0.993	0.996	0.996
Aspirin at arrival (n=192)	0.940	0.948	0.948	0.951	0.955	0.956	0.960	0.966	0.970	0.976	0.978	0.978	0.978	0.980	0.983	0.983
Aspirin prescribed at discharge (n=183)	0.947	0.944	0.953	0.956	0.956	0.958	0.964	0.967	0.971	0.974	0.980	0.979	0.981	0.982	0.981	0.982
ACE or ARB for LVSD (n=135) [c]	0.775	0.809	0.812	0.837	0.858	0.833	0.854	0.877	0.885	0.884	0.897	0.907	0.908	0.917	0.964	0.947
Beta blocker prescribed at discharge (n=188)	0.901	0.922	0.929	0.938	0.945	0.954	0.953	0.967	0.967	0.969	0.976	0.977	0.977	0.980	0.983	0.984
Beta blocker at arrival (n=193)	0.888	0.893	0.917	0.916	0.931	0.928	0.936	0.944	0.949	0.947	0.959	0.956	0.958	0.955	0.958	0.965
PCI received within 120 (90) minutes of arrival (n=112)	0.551	0.580	0.592	0.665	0.684	0.664	0.733	0.725	0.721	0.764	0.802	0.622	0.659	0.688	0.746	0.764
Survival index (n=192)	0.989	0.987	0.992	0.997	0.996	0.998	1.006	1.007	1.006	1.007	1.011	1.029	0.996	1.003	1.003	0.973

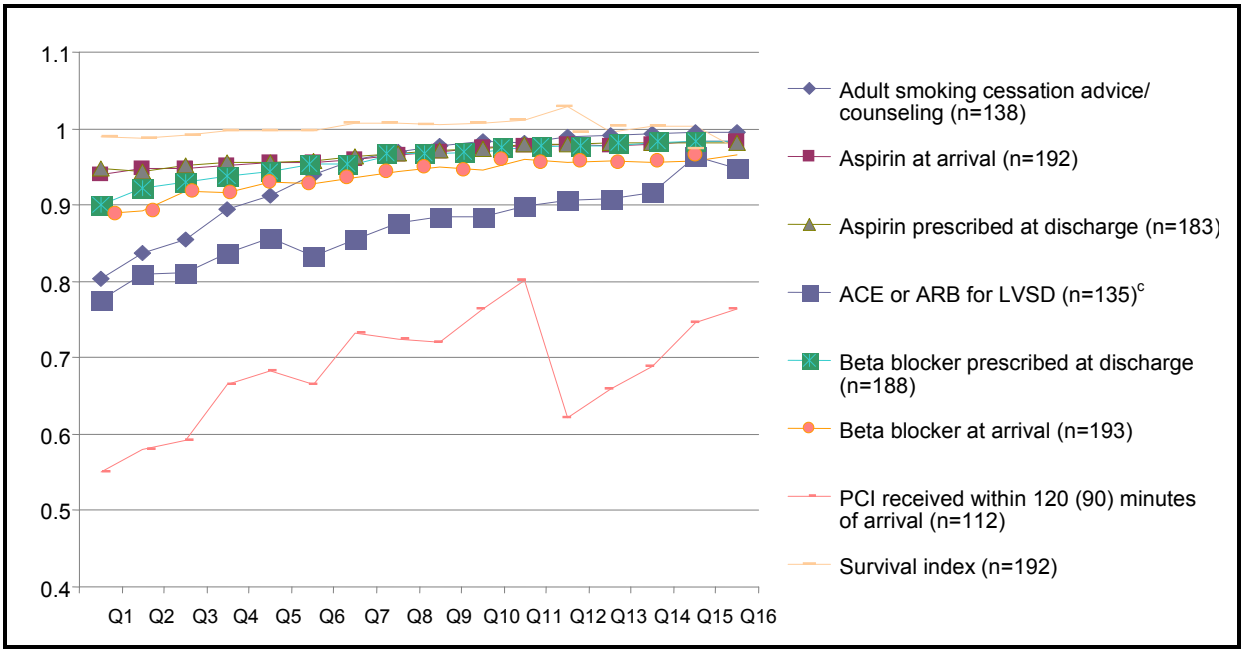
ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Figure 1
Demonstration: Trends in AMI measures, Q1–Q16 [a, b]



[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

To test whether the AMI measure differences between Q1 and Q16 were significantly different from zero, we performed paired t-tests using the Bonferroni correction for multiple testing. Results are shown in **Table 5**. This table shows, for each AMI measure, the mean score at Q1 and Q16, the percentage change in the mean between Q1 and Q16, computed as ((Q16–Q1)/Q16), the absolute difference between Q1 and Q16, and the *p*-value for the test for whether the absolute difference was significantly different from zero. This table shows how the AMI measures change over time across the 16 quarters. A positive percentage change indicates that the means have increased between Q1 and Q16.

As shown in the table, all AMI measures except Survival index were significantly different between Q1 and Q16. The largest percentage increase was noted for PCI received within 120 (90) minutes of arrival, which increased by 38.7% (from 0.551 in Q1 to 0.764 in Q16). Adult smoking cessation advice/counseling increased by 24% between Q1 and Q16, and was virtually 100% by the end of Year 4. ACE or ARB for LVSD also increased by a large percentage between Q1 and Q16 (22.2%). Aspirin prescribed at discharge showed the smallest, but statistically significant increase: 3.7% between Q1 and Q16 (yet it was 98% by the end of Year 4). The Survival index measure showed a nonsignificant decrease between Q1 and Q16, from 0.989 to 0.973.

Table 5
Demonstration: Change in AMI measures, Q1–Q16 [a, b]

AMI measure	Q1	Q16	% change Q1–Q16	Absolute change Q1–Q16	*p*-value
Adult smoking cessation advice/ counseling (n=138)	0.804	0.996	23.96%	0.192	<.001
Aspirin at arrival (n=192)	0.940	0.983	4.61%	0.043	<.001
Aspirin prescribed at discharge (n=183)	0.947	0.982	3.69%	0.035	<.001
ACE or ARB for LVSD (n=135)[c]	0.775	0.947	22.16%	0.172	<.001
Beta blocker prescribed at discharge (n=188)	0.901	0.984	9.22%	0.083	<.001
Beta blocker at arrival (n=193)	0.888	0.965	8.68%	0.077	<.001
PCI received within 120 (90) minutes of arrival (n=112)	0.551	0.764	38.71%	0.213	<.001
Survival index (n=192)	0.989	0.973	−1.63%	−0.016	ns

ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; NS = not significant; PCI = percutaneous coronary intervention.

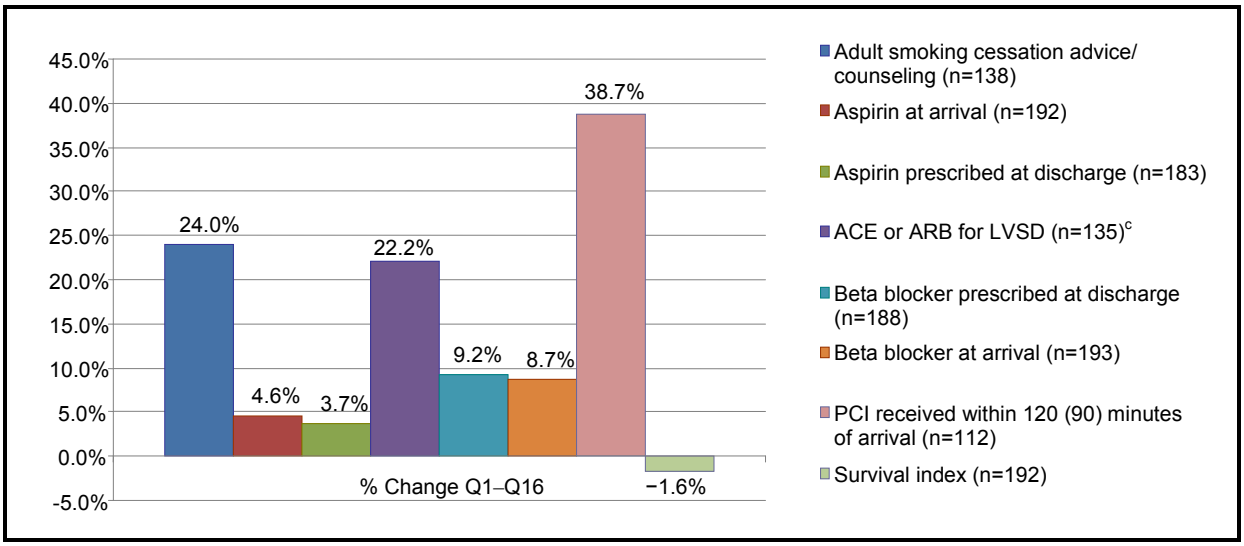
[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Statistically significant improvements are evident across time during the first 4 years of the PHQID. However, in addition to examining where a statistically significant difference existed between Q1 and Q16 for each AMI measure, we performed tests for trends in the measures to examine whether differential rates of improvement occurred across the different quarters or years of the Demonstration. These analyses provide information on whether, for example, the rate of improvement in AMI measures between Q1 and Q4 (during Year 1) was the same as the rate of improvement between Q13 and Q14 (during Year 4), or whether the rates were higher or lower during the earlier years of the Demonstration.

Figure 2
Demonstration: Percentage change in AMI measures, Q1–Q16 [a, b]



[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Results from these trend tests are shown in **Table 6**. In this table, the first eight columns show the results from paired t-tests between the first and last quarters of Years 1–4; they also show the rate at which CQSs improved each year and whether the improvement was significantly different from zero. The last two columns in the table show the result of the paired t-test examining whether the rate of change in CQS between Q1 and Q4 (i.e., during Year 1), was significantly different from the rate of change in CQS between Q13 and Q16 (i.e., during Year 4).

During Year 1, five AMI measures increased significantly between Q1 and Q4: (Adult smoking cessation advice/counseling, ACE or ARB for LVSD, Beta blocker prescribed at discharge, Beta blocker at arrival, and PCI received within 120 (90) minutes of arrival). Of these, the PCI measure increased at the fastest rate during Q1–Q4 (0.114 points). By Year 2, fewer measures showed significant increases between Q5 and Q8 (on three AMI measures), and by Year 3, *none* of the AMI measures increased significantly between Q9 and Q12.

During Year 4 (Q13 to Q16), two AMI measures showed small but significant increases: ACE or ARB for LVSD (increased 0.039 points) and PCI received within 120 (90) minutes of arrival (increased 0.105 points). Moreover, the Survival index measure showed a small but significant decrease between Q13 and Q16 (−0.022 points).

Table 6
Demonstration: Trends in AMI measures, [a, b] means Years 1–4

AMI measure	Year 1 Q4–Q1	*p*-value	Year 2 Q8–Q5	*p*-value	Year 3 Q12–Q9	*p*-value	Year 4 Q16–Q13	*p*-value	Q4–Q1 vs. Q16–Q13	*p*-value
Adult smoking cessation advice/ counseling (n=138)	0.0917	<.001	0.0569	<.001	0.0119	NS	0.0039	NS	0.0878	<.001
Aspirin at arrival (n=192)	0.0112	NS	0.0117	<.01	0.0085	NS	0.0057	NS	0.0055	NS
Aspirin prescribed at discharge (n=183)	0.0092	NS	0.0115	NS	0.0078	NS	0.0004	NS	0.0088	NS
ACE or ARB for LVSD (n=135)[c]	0.0612	<.001	0.0192	NS	0.0214	NS	0.0389	<.01	0.0223	NS
Beta blocker prescribed at discharge (n=188)	0.0371	<.001	0.0221	<.001	0.0105	NS	0.0066	NS	0.0305	<.01
Beta blocker at arrival (n=193)	0.0277	<.001	0.0131	NS	0.0062	NS	0.0075	NS	0.0202	<.01
PCI received within 120 (90) minutes of arrival (n=112)	0.1139	<.001	0.0415	NS	-0.0986	NS	0.1052	<.01	0.0087	NS
Survival index (n=192)	0.0073	NS	0.0108	NS	0.0228	NS	-0.0223	<.05	0.0296	NS

ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

The final two columns in Table 6 denote whether the measures' rates of change between Q1 and Q4 differ from the rates of change between Q13 and Q16 (i.e., a "difference in difference" between Year 1 and Year 4). Significant differences were noted for only three AMI measures: Adult smoking cessation advice/counseling, Beta blocker prescribed at discharge, and Beta blocker at arrival. For these measures, each difference was positive and significant, which indicates that the rates of change during Year 1 were significantly faster than the rates of change during Year 4.

3.2.2 AMI Measures Standard Deviation and Range

The standard deviation in the means for the AMI measures gives an indication for the amount of variation present. Larger standard deviations indicate more variation across participating hospitals, and smaller standard deviations indicate less variation. A negative percentage change between Year 1 and Year 4 indicates that the standard deviation is decreasing over time thereby suggesting that variation in performance among hospitals is diminishing. **Table 7** and **Figure 3** display the standard deviation in the AMI measures during Years 1 to 4 of the Demonstration.

Table 7
Demonstration: Trends in AMI measures, [a, b] standard deviations Years 1–4

AMI measure	Year 1	Year 2	Year 3	Year 4	% change Year 1– Year 4
Adult smoking cessation advice/counseling (n=138)	3.186	1.630	0.791	0.239	−92.5%
Aspirin at arrival (n=192)	0.698	0.548	0.413	0.307	−56.0%
Aspirin prescribed at discharge (n=183)	1.040	0.855	0.584	0.449	−56.8%
ACE or ARB for LVSD (n=135) [c]	2.019	1.723	1.613	1.180	−41.6%
Beta blocker prescribed at discharge (n=188)	1.232	0.942	0.615	0.402	−67.4%
Beta blocker at arrival (n=193)	1.325	0.962	0.700	0.612	−53.8%
PCI received within 120 (90) minutes of arrival (n=112)	3.454	3.397	3.302	3.596	4.1%
Survival index (n=192)	0.729	0.564	0.564	0.452	−38.0%

ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

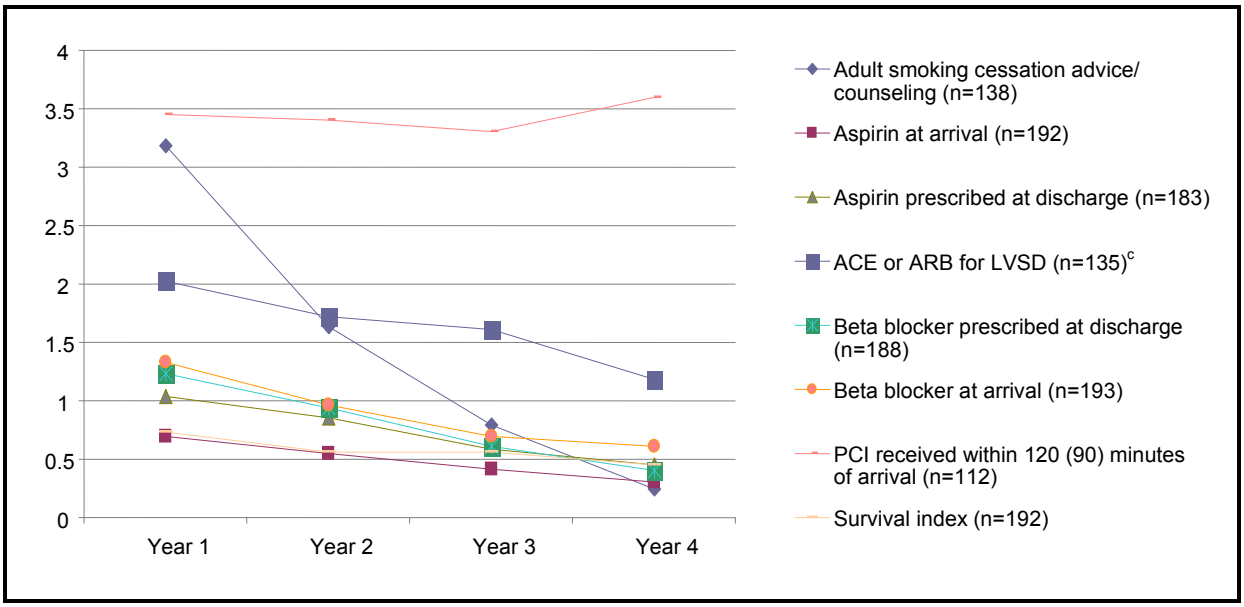
[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

As Table 7 shows, the AMI measures became less variable over the first 4 Demonstration years. The largest percentage change in standard deviations was for Adult smoking cessation advice/counseling, which decreased from 3.186 in Year 1 to 0.239 in Year 4 (a 92.5% reduction), suggesting that performance did not vary much among hospitals. During Year 1, Aspirin at arrival had the smallest standard deviation (0.698), and this measure decreased by 56% by Year 4, to 0.307. Only one measure, PCI received within 120 (90) minutes of arrival, did not show a decrease in standard deviation over PHQID Years 1 to 4. The standard deviation for this measure actually increased slightly by Year 4, and remained at over 3.3 for all 4 years.

Decreases in the standard deviation of the AMI measures indicates that they are becoming less variable over time and that hospital, by Year 4, are beginning to "cluster" around the mean without the wide range between high and low performers.

Figure 3
Demonstration: Trends in AMI measures, [a, b] standard deviations Years 1–4



ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

The "range" in AMI measures is the difference between the maximum and the minimum for all hospitals. Similar to the standard deviation, a smaller range indicates less spread (variation) across the hospitals in terms of their AMI measure. **Table 8** and **Figure 4** show the ranges for Years 1 to 4 for every AMI measure, as well as the percentage change in the range between Year 1 and Year 4. Every AMI measure decreased the range between Year 1 and Year 4, most notably Adult smoking cessation advice/counseling (88.85% decrease) and Beta blocker prescribed at discharge (55.11% decrease). Only one measure showed an increased range between Year 1 and Year 4—PCI received within 120 (90) minutes of arrival increased by 15.15%.

Table 8
Demonstration: Trends in AMI measures, [a, b] range in scores Years 1–4

AMI measure	Range Year 1	Range Year 2	Range Year 3	Range Year 4	% change Year 1– Year 4
Adult smoking cessation advice/counseling (n=138)	0.852	0.645	0.529	0.095	−88.85%
Aspirin at arrival (n=192)	0.250	0.221	0.175	0.113	−54.80%
Aspirin prescribed at discharge (n=183)	0.511	0.399	0.287	0.319	−37.57%
ACE or ARB for LVSD (n=135)[c]	0.573	0.499	0.428	0.366	−36.13%
Beta blocker prescribed at discharge (n=188)	0.450	0.373	0.241	0.202	−55.11%
Beta blocker at arrival (n=193)	0.465	0.250	0.224	0.269	−42.15%
PCI received within 120 (90) minutes of arrival (n=112)	0.845	0.722	0.849	0.973	15.15%
Survival index (n=192)	0.380	0.252	0.527	0.342	−10.00%

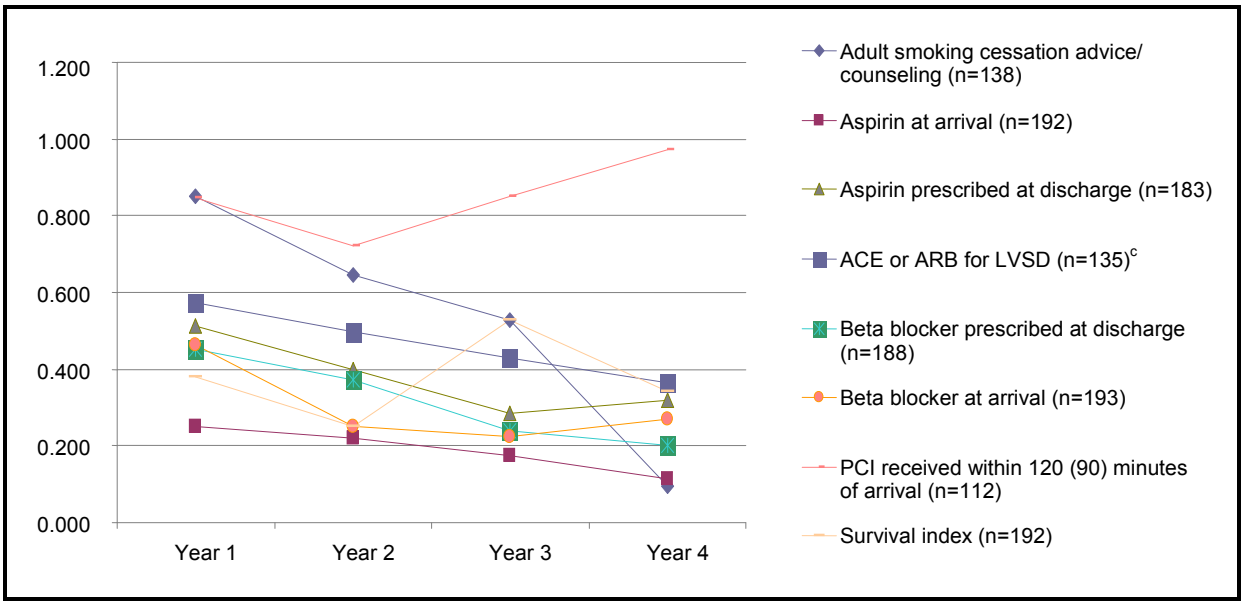
ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Figure 4
Demonstration: Trends in AMI measures, [a, b] range in scores Years 1–4



ACE = angiotensin converting enzyme; AMI = acute myocardial infarction; ARB = angiotensin receptor blocker; LVSD = left ventricular systolic dysfunction; PCI = percutaneous coronary intervention.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

3.3 Coronary Artery Bypass Graft Surgery (CABG)

This section presents results for the analyses of the process and outcome measures for CABG; included in the analysis are hospitals reporting 16 quarters of data. **Table 9** shows the measures and the number of hospitals included in the analyses (the number of hospitals with 16 quarters of data, between 94 and 109). A total of 94 hospitals reported 16 quarters of data on Post-op physical/metabolic derangement avoidance index, and 109 hospitals reported 16 quarters of data on Aspirin prescribed at discharge.

Table 9
Hospitals with CABG measures in all 16 quarters of PHQID

CABG measure	N hospitals with measures in all 16 quarters
Aspirin prescribed at discharged	109
Prophylactic antibiotics within 1 hour prior to surgical incision	108
Prophylactic antibiotics discontinued within 24/48 hours after surgery end time [a]	105
Post-op physical/metabolic derangement avoidance index	94
Post-op hemorrhage/hematoma avoidance index	109
Survival index	109

CABG = coronary artery bypass graft; PHQID = Premier Hospital Quality Incentive Demonstration.

[a] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

3.3.1 CABG Measure Means

Table 10 shows the means for each CABG measure across 16 quarters of the Demonstration (Years 1–4). These measures are shown for a panel of hospitals that reported the CABG measures in each of the 16 quarters. **Figure 5** details in graphical form the trends in CABG measures across the 16 quarters. As shown in Table 10 and Figure 5, the mean for each CABG measure increased during the first 4 years of the Demonstration. For all measures, the increases were steady until the 10th quarter and then leveled off and were sustained the remainder of the quarters. Some measures reached a near ceiling; some measures started high and remained high (e.g., post operative hemorrhage avoidance) while others had very impressive gains during the first couple of years of the Demonstration (e.g., stopping of prophylactic antibiotics within 24 hours after surgery end time).

To test whether the CABG measure differences between Q1 and Q16 were significantly different from zero, we performed paired t-tests using the Bonferroni correction for multiple testing. Results are shown in **Table 11**. This table shows, for each CABG measure, the mean score at Q1 and Q16, the percentage change in the mean between Q1 and Q16, computed as ((Q16–Q1)/Q16), the absolute difference between Q1 and Q16, and the *p*-value for the test for whether the absolute difference was significantly different from zero.

Table 10
CABG measures for Demonstration Years 1–4, means [a, b]

Measure code	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Aspirin prescribed at discharge (n=109)	0.947	0.950	0.961	0.962	0.966	0.974	0.974	0.979	0.982	0.981	0.985	0.987	0.988	0.991	0.988	0.992
Prophylactic antibiotics within 1 hour prior to surgical incision (n=108)	0.697	0.762	0.809	0.844	0.868	0.896	0.917	0.922	0.934	0.942	0.954	0.948	0.955	0.961	0.944	0.962
Prophylactic antibiotics discontinued within 24/48 hours after surgery end time (n=105) [c]	0.475	0.524	0.578	0.611	0.698	0.768	0.802	0.820	0.823	0.949	0.945	0.939	0.945	0.955	0.961	0.968
Post-op phy/metabolic derangement avoidance index (n=94)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.996	0.996	0.995	0.995	0.995	0.993	0.995
Post-op hemorrhage/ hematoma avoidance index (n=109)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	0.999	1.000	1.000	0.999	1.000
Survival index (n=109)	1.004	1.002	1.005	1.007	1.002	1.003	1.004	1.007	1.002	1.004	1.006	1.006	1.005	1.004	1.006	1.006

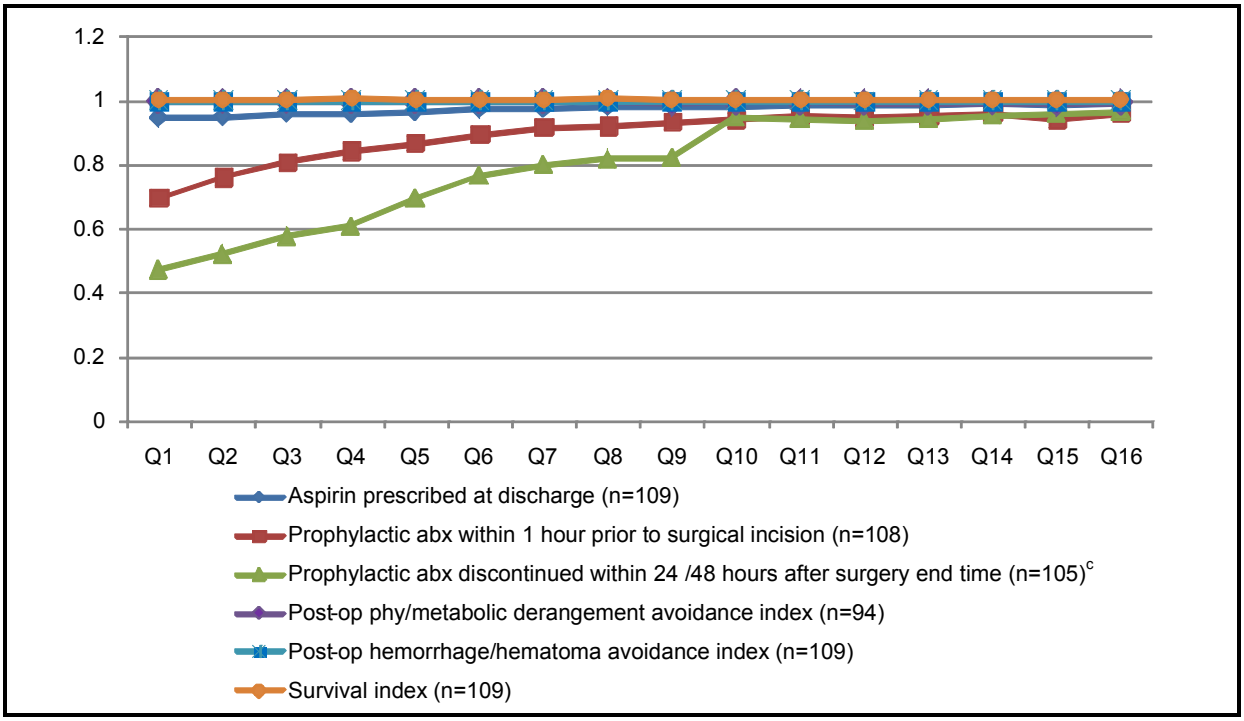
CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

Figure 5
Demonstration: Trends in CABG measures, [a, b] Q1–Q16



CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

As shown in the table, all AMI measures except Post-op physical/metabolic derangement avoidance index, Post-op hemorrhage/hematoma avoidance index, and Survival index were significantly different between Q1 and Q16; however, they were very high at the outset. The largest percentage increase was noted for Prophylactic antibiotics discontinued within 24 hours after surgery end time, which increased by over 100% (from 0.475 in Q1 to 0.968 in Q16). Prophylactic antibiotics within 1 hour prior to surgical incision increased by 38.1% between Q1 and Q16, and was at 97% by the end of Demonstration Year 4. The CABG measure that showed the smallest, but statistically significant increase was Aspirin prescribed at discharge, which increased by 4.8% between Q1 and Q16 (yet was 99% by the end of Year 4).

In addition to examining where a statistically significant difference existed between Q1 and Q16 for each CABG measure, we performed tests for trends in the measures to examine whether differential rates of improvement occurred across the different quarters or years of the Demonstration. These analyses provide information on whether, for example, the rate of improvement in AMI measures between Q1 and Q4 (during Year 1) was the same as the rate of improvement between Q13 and Q16 (during Year 4), or whether the rates were higher or lower during the earlier years of the Demonstration.

Table 11
Demonstration: Change in CABG measures, [a, b] Q1–Q16

Measure code	Q1	Q16	% change Q1–Q16	Absolute difference Q1 to Q16	*p*-value
Aspirin prescribed at discharge (n=109)	0.947	0.992	0.048	0.045	<.01
Prophylactic antibiotics within 1 hour prior to surgical incision (n=108)	0.697	0.962	0.381	0.265	<.001
Prophylactic antibiotics discontinued within 24/48 hours after surgery end time (n=105)[c]	0.475	0.968	1.037	0.493	<.001
Post-op physical/metabolic derangement avoidance index (n=94)	1.000	0.995	−0.004	−0.004	NS
Post-op hemorrhage/hematoma avoidance index (n=109)	1.000	1.000	0.000	0.000	NS
Survival index (n=109)	1.004	1.006	0.002	0.002	NS

CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

Results from these trend tests are shown in **Table 12**. During Year 1, only two of the CABG measures increased significantly between Q1 and Q4 (both of the prophylactic antibiotic measures). Of these, Prophylactic antibiotic within 1 hour prior to surgical incision increased at the fastest rate during Q1–Q4 (0.147 points). These same two measures increased at a significant rate during Year 2 (but at a slower rate). By Year 4, none of the CABG measures were increasing at a significant rate; they reached a plateau and were sustaining the gains made during the early years of the Demonstration.

The final two columns in Table 12 denote whether the measures' rates of change between Q1 and Q4 are different from the rates of change between Q13 and Q16 (i.e., a "difference in difference" between Year 1 and Year 4). Significant differences were noted for only the two CABG measures that also showed significant increases in Year 1 (both prophylactic antibiotic measures). For these measures, each difference was positive and significant, which indicates that the rates of change during Year 1 were significantly faster than the rates of change during Year 4.

Table 12
Trends in rates of change in CABG measures Q1–Q16, means [a, b]

Measure code	Year 1 Q4–Q1	*p*-value	Year 2 Q8–Q5	*p*-value	Year 3 Q12–Q9	*p*-value	Year 4 Q16–Q13	*p*-value	Year 1 Q4–Q1 vs. Year 4 Q16–Q13	*p*-value
Aspirin prescribed at discharge (n=109)	0.015	NS	0.013	NS	0.005	NS	0.005	NS	0.010	NS
Prophylactic antibiotics within 1 hour prior to surgical incision (n=108)	0.147	<.001	0.054	<.004	0.014	NS	0.007	NS	0.140	<.001
Prophylactic antibiotics discontinued within 24/48 hours after surgery end time (n=105) [c]	0.136	<.001	0.122	<.001	0.115	<.001	0.023	NS	0.113	<.003
Post-op physical/metabolic derangement avoidance index (n=94)	0.000	NS	0.000	NS	0.001	NS	0.000	NS	0.000	NS
Post-op hemorrhage/ hematoma avoidance index (n=109)	0.000	NS	0.000	NS	0.000	NS	0.000	NS	0.000	NS
Survival index (n=109)	0.003	NS	0.005	NS	0.004	NS	0.001	NS	0.002	NS

CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

3.3.2 CABG Measures Standard Deviation and Range

The standard deviation in the means for the CABG measures gives an indication of the amount of variation present. **Table 13** and **Figure 6** display the standard deviation in the CABG measures during Years 1–4 of the Demonstration.

Table 13
Demonstration: Trends in CABG measures, [a, b] standard deviations Years 1–4

Measure code	Year 1	Year 2	Year 3	Year 4	% change Year 1 to Year 4
Aspirin prescribed at discharge (n=109)	0.843	0.634	0.429	0.224	−73.41%
Prophylactic antibiotics within 1 hour prior to surgical incision (n=108)	3.048	1.651	0.977	0.884	−71.02%
Prophylactic antibiotics discontinued within 24/48 hours after surgery end time (n=105) [c]	5.485	4.178	1.974	0.824	−84.98%
Post-op phy/metabolic derangement avoidance index (n=94)	0.007	0.009	0.097	0.110	1,404.04%
Post-op hemorrhage/hematoma avoidance index (n=109)	0.003	0.003	0.015	0.012	254.92%
Survival Index (n=109)	0.214	0.201	0.230	0.230	7.68%

CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

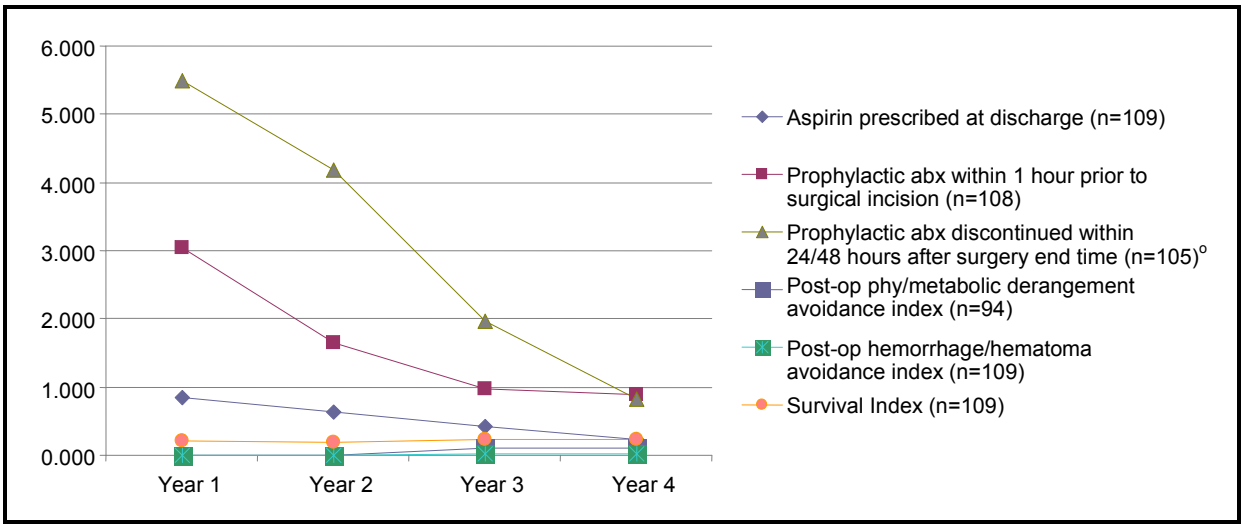
[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

As the table shows, the CABG process measures became less variable over the 4 Demonstration years. The largest percentage decrease in standard deviations was for Prophylactic antibiotic discontinued within 24 hours after surgery end time, which decreased by 95%; this measure had the highest standard deviation in Year 1 suggesting large variation in practice among hospitals at the outset. Decreases in the standard deviation of the CABG measures indicates that hospital performance is becoming less variable over time as performance "clusters" around the mean.

The standard deviations for the three outcome measures actually increased during the first 4 years of the Demonstration; however, as they were so low in Year 1 and remained low in Year 4, these increases are not meaningful.

Figure 6
Demonstration: Trends in CABG measures, [a, b] standard deviations Years 1–4



CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

As with the standard deviation, a decrease in range among hospital CABG measures indicates that there was less spread (variation) between hospitals performing at the high vs low end of the spectrum. **Table 14** and **Figure 7** show the ranges for Years 1–4 for each of the CABG measure, as well as the percentage change in the range between Year 1 and Year 4. Results for the range trend the same as results from the standard deviation: three process measures showed decreases in their range (over 50% for all measures), while the three outcome measures increased their range (yet their ranges remain very small in Year 4).

Table 14
Demonstration: Trends in CABG measures, [a, b] ranges Years 1–4

CABG measure	Range Year 1	Range Year 2	Range Year 3	Range Year 4	% change Year 1 to Year 4
Aspirin prescribed at discharge (n=109)	0.25	0.25	0.37	0.08	−66.77%
Prophylactic antibiotics within 1 hour prior to surgical incision (n=108)	0.92	0.78	0.57	0.33	−63.94%
Prophylactic antibiotics discontinued within 24/48 hours after surgery end time (n=105) [c]	0.99	0.97	0.79	0.47	−52.98%
Post-op phy/metabolic derangement avoidance index (n=94)	0.00	0.00	0.04	0.04	1,696.09%
Post-op hemorrhage/hematoma avoidance index (n=109)	0.00	0.00	0.01	0.01	869.17%
Survival index (n=109)	0.12	0.08	0.16	0.13	7.11%

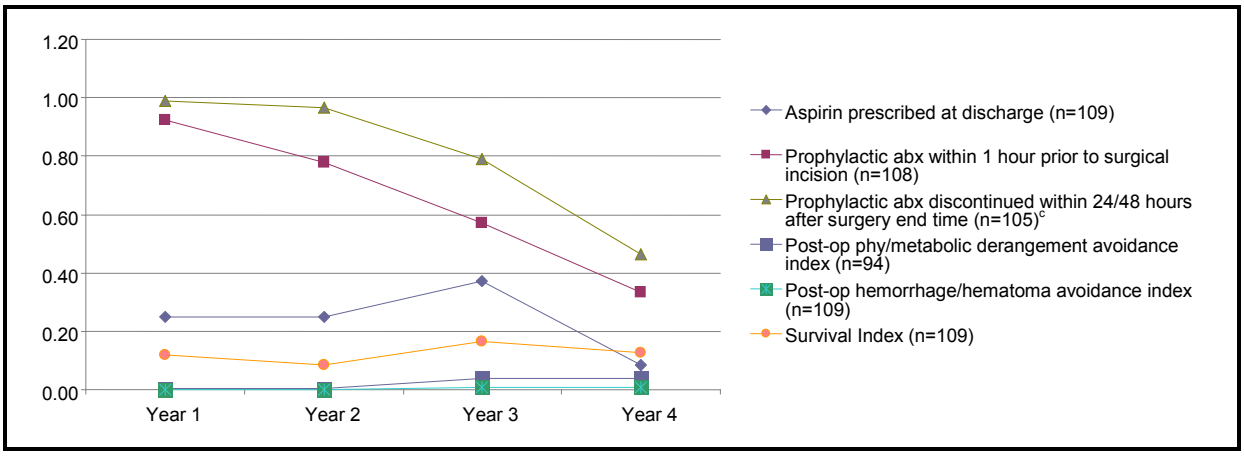
CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

Figure 7
Demonstration: Trends in CABG measures, [a, b] ranges Years 1–4



CABG = coronary artery bypass graft.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from 24 to 48 hours in Year 3 of the Demonstration (effective for discharges after January 1, 2006).

3.4 Heart Failure (HF)

This section presents results for the analyses of the process and outcome measures for HF among hospitals reporting 16 quarters of data. **Table 15** shows the measures and the number of hospitals included in the analyses (the number of hospitals with 16 quarters of data, between 191 and 215). A total of 191 hospitals reported 16 quarters of data on Smoking cessation advice/counseling, and 215 hospitals reported 16 quarters of data on both Discharge instructions and LVF assessment.

Table 15
Hospitals with HF measures in all 16 quarters of PHQID

HF measure	N hospitals with measures in all 16 quarters
Discharge instructions	215
LVF assessment	215
Smoking cessation advice/counseling	191
ACEI or ARB for LVSD [a]	207

ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction; PHQID = Premier Hospital Quality Incentive Demonstration.

[a] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

3.4.1 HF Measure Means

Table 16 shows the means for each HF measure across the first 16 quarters of the Demonstration (Years 1–4). These measures are shown for a panel of hospitals that reported the HF measures in each of the 16 quarters. **Figure 8** details in graphical form the trends in HF measures across the 16 quarters. As shown in Table 16 and Figure 8, the mean for each HF measure increased during the first 4 years of the Demonstration. For all measures, the increases were steady over time. The increase in mean performance for Discharge instructions is far greater than the increase for all other measures, because the mean performance was so low at the start of the Demonstration (0.42).

Table 16
Demonstration: Trends in HF measures, [a, b] Q1–Q16

HF measure	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Discharge instructions (n=215)	0.424	0.476	0.528	0.541	0.595	0.626	0.644	0.648	0.678	0.734	0.755	0.775	0.783	0.811	0.827	0.846
LVF assessment (n=215)	0.858	0.873	0.892	0.902	0.913	0.923	0.927	0.937	0.945	0.952	0.960	0.961	0.964	0.969	0.974	0.976
Smoking cessation advice/ counseling (n=191)	0.617	0.672	0.743	0.791	0.835	0.882	0.908	0.918	0.930	0.947	0.962	0.975	0.975	0.981	0.980	0.984
ACEI or ARB for LVSD (n=207)[c]	0.768	0.763	0.776	0.790	0.799	0.837	0.852	0.852	0.877	0.882	0.884	0.892	0.898	0.914	0.934	0.940

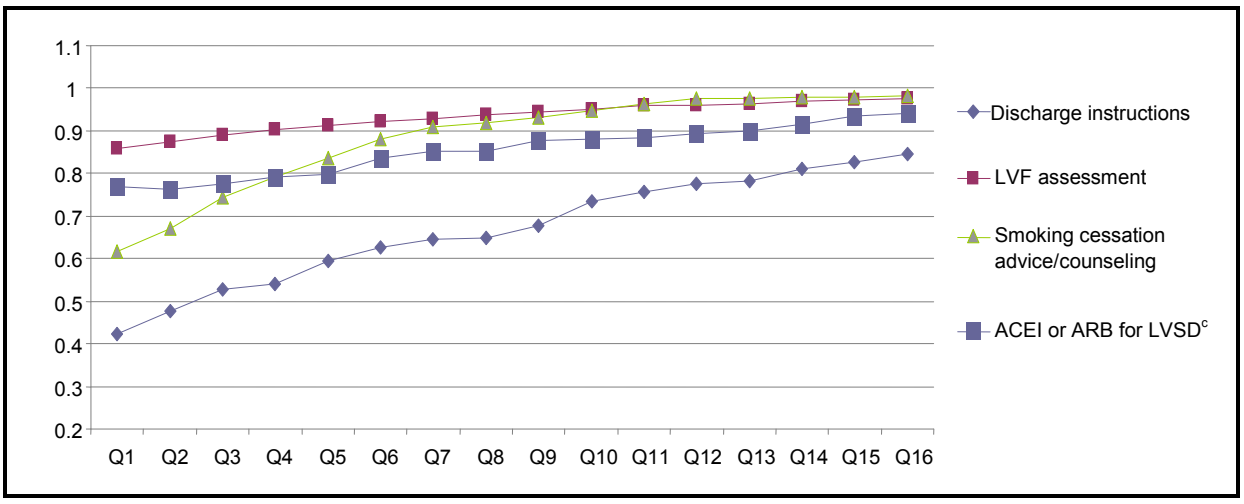
ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Figure 8
Demonstration: Trends in HF measures, [a, b] Q1–Q16



ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Results for the paired t-tests for differences in mean HF scores between Q1 and Q16 are shown in **Table 17**. This table shows, for each HF measure, the mean score at Q1 and Q16, the percentage change in the mean between Q1 and Q16, computed as ((Q16–Q1)/Q16), the absolute difference between Q1 and Q16, and the *p*-value for the test for whether the absolute difference was significantly different from zero.

As shown in the table, all HF measures were significantly different between Q1 and Q16. The largest percentage increase was noted for Discharge instructions, which increased by essentially 100% (from 0.424 in Q1 to 0.846 in Q16). Smoking cessation advice/counseling increased by 59.5% between Q1 and Q16, and was over 98% by the end of Demonstration Year 4 (the highest of all HF measures). The HF measure LVS Assessment showed the smallest, but statistically significant increase: 13.8% between Q1 and Q16. Despite the 100% increase between Q1 and Q16, Discharge instructions remained the lowest HF measure in terms of performance at the end of the fourth Demonstration year (84.6%).

While HF measures improved significantly during the first 4 years of the Demonstration, much of the improvement was observed during the first 3 years. In Year 4, the means for some measures leveled off and approached the ceiling of the range. We also performed tests for trends in the measures to examine whether there were differential rates of improvement across the different quarters or years of the Demonstration. These analyses provide information on whether, for example, the rate of improvement in HF measures between Q1 and Q4 (during Year 1) was

the same as the rate of improvement between Q13 and Q14 (during Year 4), or whether the rates were higher or lower during the earlier years of the Demonstration.

Table 17
Demonstration: Change in HF measures, [a, b] Q1–Q16

HF measure	Q1	Q16	% change Q1–Q16	Absolute difference Q1 to Q16	*p*-value
Discharge instructions (n=215)	0.424	0.846	99.8%	0.423	<.001
LVF assessment (n=215)	0.858	0.976	13.8%	0.118	<.001
Smoking cessation advice/counseling (n=191)	0.617	0.984	59.5%	0.367	<.001
ACEI or ARB for LVSD (n=207)[c]	0.768	0.940	22.5%	0.172	<.001

ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Results from these trend tests are shown in **Table 18**. During Year 1, all HF measures except ACEI or ARB for LVSD increased significantly between Q1 and Q4. Smoking cessation advice/counseling increased at the fastest rate during Q1–Q4 (0.174 points). By Year 2, all four HF measures showed significant increases between Q5 and Q8, although the rates of increase were approximately half those seen in Year 1.By Year 3, *none* of the HF measures showed significant increases between Q9 and Q12.

During Year 4 (Q13 to Q16), all HF measures but Smoking cessation advice/counseling showed small but significant increases: Discharge instructions (increased 0.063 points), LVF assessment (increased 0.012 points), and ACEI or ARB for LVSD (increased 0.042 points).

The final two columns in Table 18 denote whether the measures' rates of change between Q1 and Q4 are different from the rates of change between Q13 and Q16 (i.e., a "difference in difference" between Year 1 and Year 4). Significant differences were noted for only three HF measures: Discharge instructions, LVF assessment, and Smoking cessation advice/counseling. For these measures, each difference was positive and significant, which indicates that the rates of change during Year 1 were significantly faster than the rates of change during Year 4.

Table 18
Demonstration: Trends in HF measure means, [a, b] Years 1–4

HF measure	Year 1 Q4–Q1	*p*-value	Year 2 Q8–Q5	*p*-value	Year 3 Q12–Q9	*p*-value	Year 4 Q16–Q13	*p*-value	Q4–Q1 vs. Q16–Q13	*p*-value
Discharge instructions (n=215)	0.117	<.001	0.053	<.005	0.097	<.001	0.063	<.003	0.054	<.001
LVF assessment (n=215)	0.045	<.001	0.024	<.002	0.016	<.001	0.012	<.003	0.033	<.001
Smoking cessation advice/counseling (n=191)	0.174	<.001	0.082	<.001	0.045	<.001	0.009	NS	0.165	<.001
ACEI or ARB for LVSD (n=207)[c]	0.022	NS	0.053	<.001	0.015	NS	0.042	<.001	-0.019	NS

ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

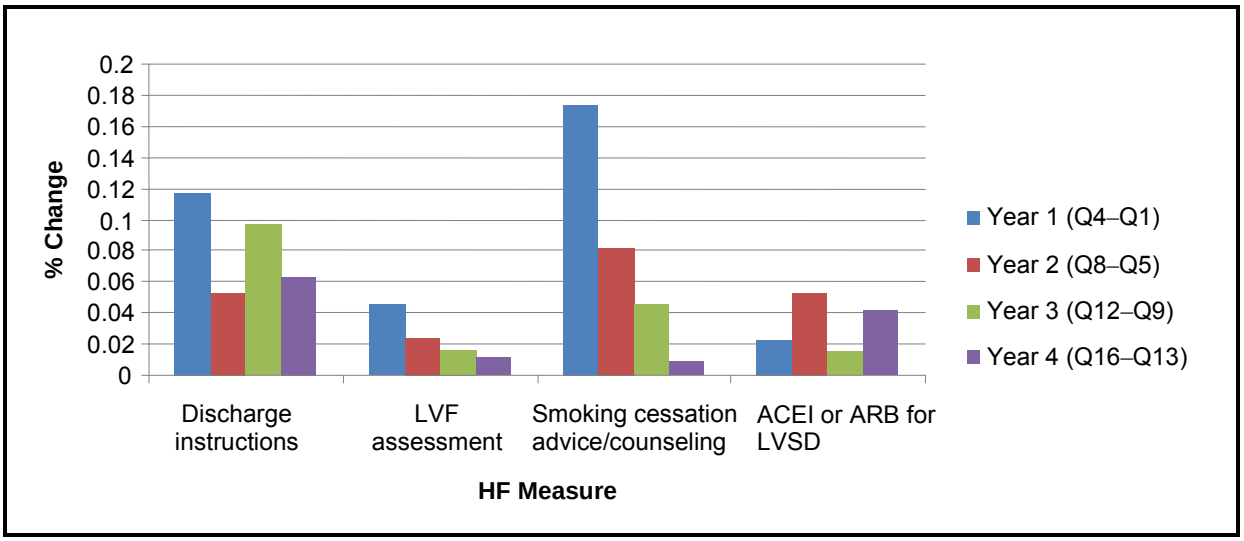
Figure 9 depicts the percentage change in each HF measures by Demonstration year. This figure illustrates that for most measures, the percentage change between the first and last quarter of Year 1 was much larger than the percentage change between the first and last quarter of the other years, which indicates that the rate of increase in the HF measures is slowing.

3.4.2 HF Measures Standard Deviation and Range

The standard deviation in the means for the HF measures indicates the amount of variation present. **Table 19** and **Figure 10** display the standard deviation in the HF measures during Years 1–4 of the Demonstration.

As the table shows, hospital performance among the HF measures became less variable over time. All HF measures experienced a reduction in the standard deviation of the mean scores by 44.5% to 81.2%. As with the findings for the AMI measures, the greatest percentage change in standard deviations was observed for Smoking cessation advice/counseling, which decreased from 4.63 in Year 1 to 0.873 in Year 4 (a reduction of 81.2%). During Year 1 of the Demonstration, LVF assessment had the smallest standard deviation (1.78); by Year 4 this measure had a 65.9% reduction in the standard deviation (to 0.607). The observed increase in mean performance along with a decrease in the standard deviation of the mean HF measures indicate that the over time, there has been improvement among participating hospitals and that the variation in performance among these hospitals has decreased.

Figure 9
Demonstration: Percentage change in HF measure means, [a, b] Years 1–4



ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Table 19
Demonstration: Trends in HF measures, [a, b] standard deviations Years 1–4

HF measure	Standard deviation Year 1	Standard deviation Year 2	Standard deviation Year 3	Standard deviation Year 4	% change Year 1 to Year 4
Discharge instructions (n=215)	5.777	4.979	3.713	2.596	−55.06%
LVF assessment (n=215)	1.780	1.343	0.949	0.607	−65.90%
Smoking cessation advice/counseling (n=191)	4.633	2.800	1.669	0.873	−81.16%
ACEI or ARB for LVSD (n=207)[c]	2.433	2.026	1.661	1.349	−44.55%

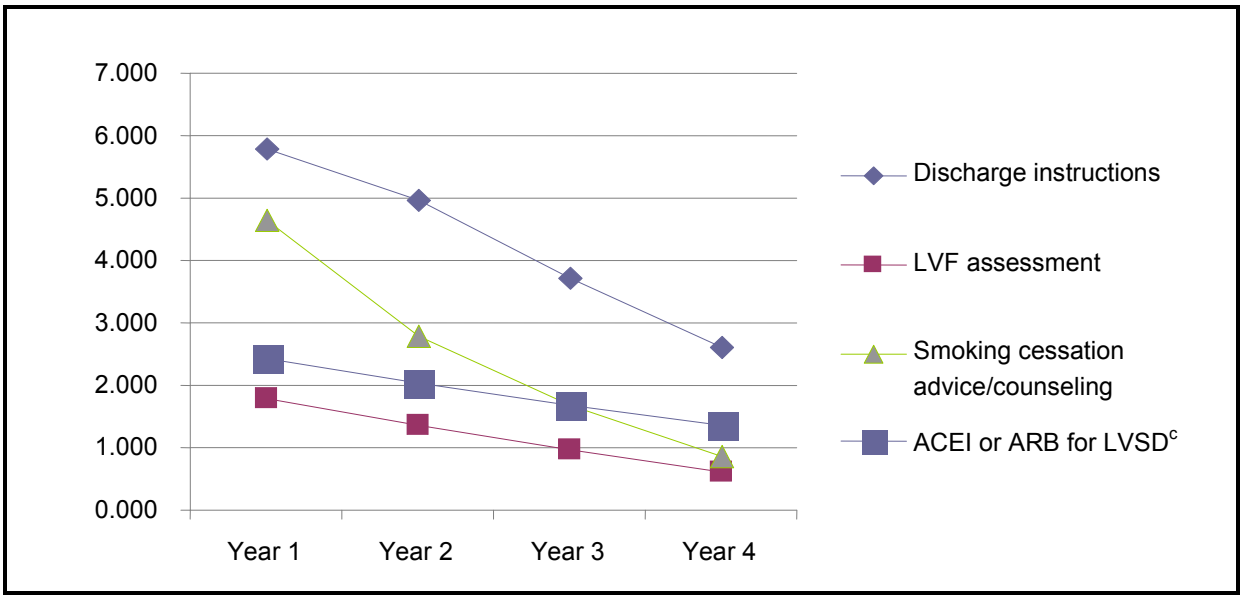
ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Figure 10
Demonstration: Trends in HF measures, [a, b] standard deviations Years 1–4



ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

The trends in the range in HF measures are depicted in **Table 20** and **Figure 11.** Every HF measure decreased the range between Year 1 and Year 4, most notably Smoking cessation advice/counseling (74.4% decrease) and LVF assessment (54.6% decrease). The measure that showed the smallest decrease in range between Year 1 and Year 4 was Discharge instructions, which had the highest range in both Year 1 and Year 4 (and only decreased by 16.1% over the 4 years).

Table 20
Demonstration: Trends in HF measures, [a, b] range of scores Years 1–4

HF measure	Range Year 1	Range Year 2	Range Year 3	Range Year 4	% change Year 1 to Year 4
Discharge instructions (n=215)	0.990	0.967	0.794	0.831	−16.06%
LVF assessment (n=215)	0.471	0.426	0.358	0.214	−54.56%
Smoking cessation advice/counseling (n=191)	0.948	0.779	0.479	0.262	−72.36%
ACEI or ARB for LVSD (n=207)[c]	0.826	0.519	0.428	0.482	−41.65%

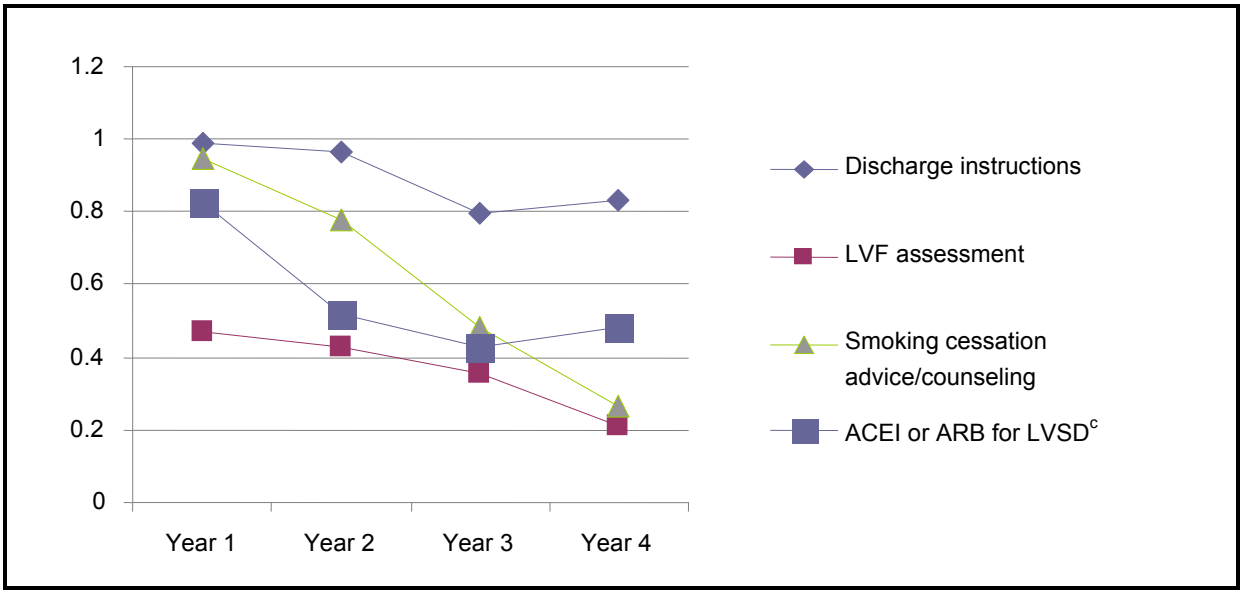
ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

Figure 11
Demonstration: Trends in HF measures, [a, b] range of scores Years 1–4



ACEI = angiotensin I-converting enzyme inhibitor; HF = heart failure; LVF = left ventricular failure; LVSD = left ventricular systolic dysfunction.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

[c] This measure definition changed from "ACE for LVSD" to "ACE or ARB for LVSD" in Quarter 2, Year 2 of the Demonstration (effective for discharges after January 1, 2005).

3.5 Hip/Knee Replacement Surgery (HK)

This section presents results for the analyses of the process and outcome measures for HK that included hospitals reporting all 16 quarters of data. **Table 21** shows the measures and the number of hospitals included in the analyses (the number of hospitals with 16 quarters of data, between 149 and 166). A total of 149 hospitals reported 16 quarters of data on Readmission (30 day) avoidance index, and 166 hospitals reported 16 quarters of data on Post-op hemorrhage/ hematoma avoidance index.

Table 21
Hospitals with HK measures in all 16 quarters of PHQID

HK measure	N hospitals with measures in all 16 quarters
Prophylactic antibiotics within 1 hour prior to surgical incision	165
Prophylactic antibiotics selection for surgical patients	162
Post-op physical/metabolic derangement avoidance index	151
Post-op hemorrhage/hematoma avoidance index	166
Readmission (30 day) avoidance index	149

HK = hip/knee replacement; PHQID = Premier Hospital Quality Incentive Demonstration.

3.5.1 Hip/Knee Replacement Surgery (HK) Measure Means [a, b]

Table 22 shows the means for each HK measure across the 16 quarters of the Demonstration (Years 1–4). These measures are shown for a panel of hospitals that reported the HK measures in each of the 16 quarters. **Figure 12** details in graphical form the trends in HK measures across the 16 quarters. As shown in Table 22 and Figure 12, the mean for the two process measures prophylactic antibiotics within 1 hour prior to surgical incision and Prophylactic antibiotics selection for surgical patients increased steadily over the first 4 Demonstration years, while the three outcome measures remained consistently high over the 4 Demonstration years, suggesting that adverse outcomes are rare events among all hospitals.

Results for the paired t-tests for differences in mean HK scores between Q1 and Q16 are shown in **Table 23**. This table shows, for each HK measure, the mean score at Q1 and Q16, the percentage change in the mean between Q1 and Q16, computed as ((Q16–Q1)/Q16), the absolute difference between Q1 and Q16, and the *p*-value for the test for whether the absolute difference was significantly different from zero.

As shown in the table, both HK process measures (prophylactic antibiotics) were significantly different between Q1 and Q16. The largest percentage increase was noted for Prophylactic antibiotics selection for surgical patients, which increased by 84.9% (from 0.512 in Q1 to 0.948 in Q16). Only one outcome measure, Readmission (30 day) avoidance index was significantly different between Q1 and Q16. However, this difference, although statistically significant, was not clinically substantial (a decrease of 2.4%).

Table 22
HK measures for Demonstration Years 1–4, [a, b] means

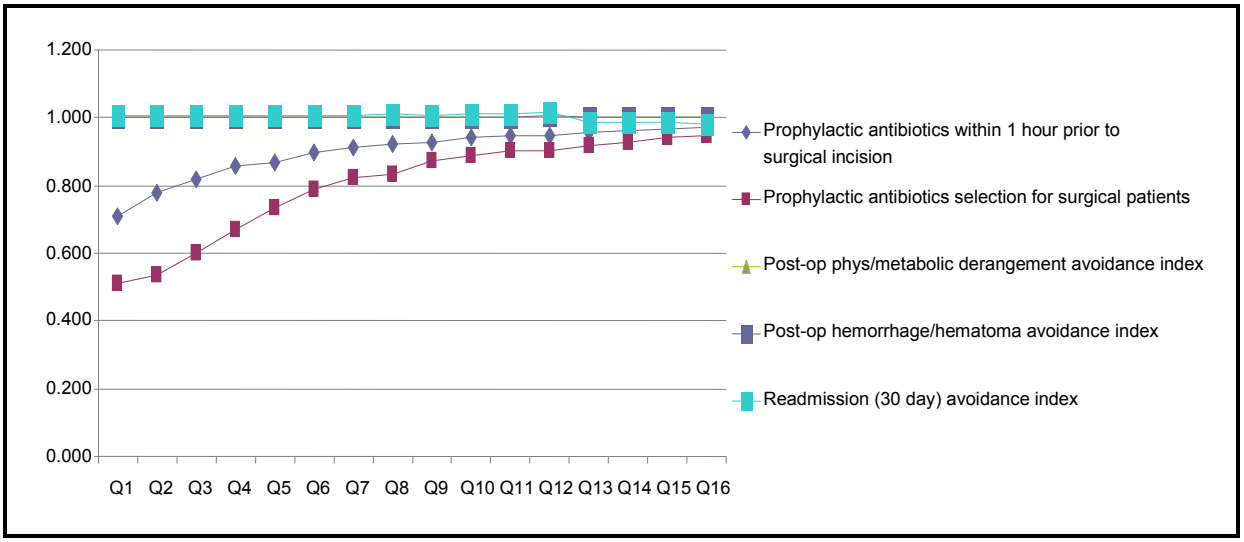
HK measure	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Prophylactic antibiotics within 1 hour prior to surgical incision (n=165)	0.71	0.78	0.82	0.86	0.87	0.90	0.91	0.92	0.93	0.94	0.95	0.95	0.96	0.96	0.97	0.97
Prophylactic antibiotics selection for surgical patients (n=162)	0.51	0.54	0.60	0.67	0.74	0.79	0.82	0.84	0.87	0.89	0.90	0.90	0.92	0.93	0.94	0.95
Post-op phys/metabolic derangement avoidance index (n=151)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Post-op hemorrhage/hematoma avoidance index (n=166)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Readmission (30 day) avoidance index (n=149)	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.01	1.01	1.01	1.01	1.02	0.99	0.99	0.99	0.98

HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 12
Demonstration: Trends in HK measures, [a, b] Q1–Q16



HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Table 23
Demonstration: Change in HK measures, [a, b] Q1–Q16

HK measure	Q1	Q16	% change Q1–Q16	Absolute difference Q1 to Q16	*p*-value
Prophylactic antibiotics within 1 hour prior to surgical incision (n=165)	0.708	0.970	36.9%	0.261	<.001
Prophylactic antibiotics selection for surgical patients (n=162)	0.512	0.948	84.9%	0.435	<.001
Post-op phys/metabolic derangement avoidance index (n=151)	1.000	1.000	0.0%	0.000	NS
Post-op hemorrhage/hematoma avoidance index (n=166)	1.002	1.001	−0.1%	−0.001	NS
Readmission (30 day) avoidance index (n=149)	1.007	0.983	−2.4%	−0.024	<.001

HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

The HK process measures improved significantly during the first 4 years of the Demonstration; however, most of the improvement took place during the first 3 years of the Demonstration, with a slight increase or leveling off during Year 4. We also tested trends in the measures to examine whether there were differential rates of improvement across the different quarters or years of the Demonstration. Results from these trend tests are shown in **Table 24**.

The process measure Prophylactic antibiotics within 1 hour prior to surgical incision increased significantly during the first 3 years of the Demonstration (at a higher rate in Year 1 than in Year 2 or Year 3). By Year 4, the rate of change in this measure was not significant (i.e., the growth in the measure was flattening out). The difference-in-difference test examining whether the rate of change in Year 1 was significantly different than the rate of change in Year 4 was statistically significant and positive, which indicates that this process measure increased at a faster rate during Year 1.

The other HK process measure Prophylactic antibiotics selection for surgical patients showed significant increases in each of the first 4 years of the Demonstration, but at decreasing rates each year. For example, the rate of increase for Year 1 was 0.158 points, which dropped to 0.100 points by Year 2, then to 0.032 points by Year 3, and 0.030 points by Year 4. Similar to the findings for the other process measure, the difference-in-difference test examining whether the rate of change in Year 1 was different than the rate of change in Year 4 was significant and positive, which indicates that this process measure increased at a faster rate during Year 1.

Table 24
Demonstration: Trends in HK measure means, [a, b] Years 1–4

HK measure	Year 1 Q4–Q1	*p*-value	Year 2 Q8–Q5	*p*-value	Year 3 Q12–Q9	*p*-value	Year 4 Q16–Q13	*p*-value	Year 1 Q4–Q1 vs. Year 4 Q16–Q13	*p*-value
Prophylactic antibiotics within 1 hour prior to surgical incision (n=165)	0.150	<.001	0.054	<.001	0.024	<.001	0.012	NS	0.138	<.001
Prophylactic antibiotics selection for surgical patients (n=162)	0.158	<.001	0.100	<.001	0.032	<.001	0.030	<.001	0.128	<.001
Post-op phys/metabolic derangement avoidance index (n=151)	0.000	NS	0.000	NS	0.000	NS	0.000	NS	0.000	NS
Post-op hemorrhage/ hematoma avoidance index (n=166)	0.000	NS	0.000	NS	0.002	NS	0.001	NS	−0.001	NS
Readmission (30 day) avoidance index (n=149)	0.000	NS	0.005	NS	0.009	NS	−0.002	NS	0.002	NS

HK = hip/knee replacement; NS = not significant.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

The rates of change for the three HK outcome measures were not statistically significant across any of the 4 years examined.

3.5.2 HK Measures Standard Deviation and Range

The standard deviation in the means for the HK measures gives an indication for the amount of variation present. **Table 25** and **Figure 13** display the standard deviation in the HK measures during Years 1–4 of the Demonstration.

As the table shows, the two HF process measures became less variable over the first 4 Demonstration years. Standard deviations for these prophylactic antibiotics decreased by 64.2%–70.5% between Year 1 and Year 2, which indicates that the variation in the process measures decreased over time.

The standard deviation for the three outcome measures did not change in a consistent pattern; however, performance among these measures was high at the start of the Demonstration with little variation among hospitals.

Table 25
Demonstration: Trends in HK measures, [a, b] standard deviations Years 1–4

HK measure	Standard deviation Year 1	Standard deviation Year 2	Standard deviation Year 3	Standard deviation Year 4	% change Year 1 to Year 4
Prophylactic antibiotics within 1 hour prior to surgical incision (n=165)	2.306	1.488	0.879	0.681	−70.47%
Prophylactic antibiotics selection for surgical patients (n=162)	3.837	2.974	1.881	1.375	−64.16%
Post-op phys/metabolic derangement avoidance index (n=151)	0.001	0.001	0.024	0.038	3,700.00%
Post-op hemorrhage/hematoma avoidance index (n=166)	0.000	0.000	0.153	0.078	N/A
Readmission (30 day) avoidance index (n=149)	0.364	0.357	0.376	0.584	60.44%

HK = hip/knee replacement; NA = not applicable.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

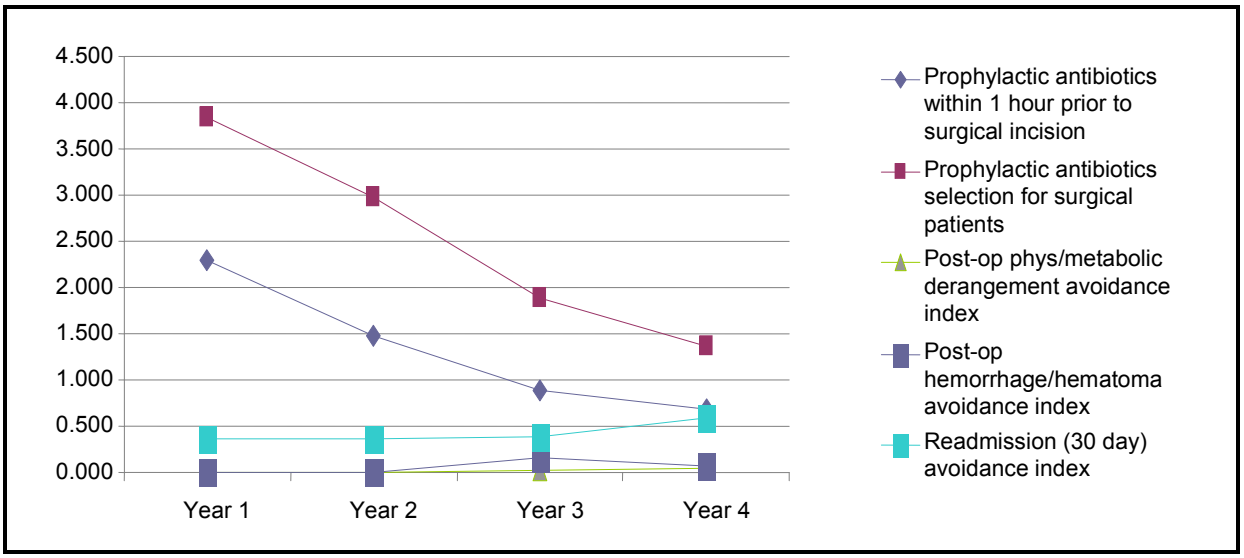
The trends in the range in HK measures are depicted in **Table 26** and **Figure 14**. These findings are similar to the findings for the standard deviation trends during the first 4 years: the range in the two process measures decreased between 60% and 70% over time, while the range in the outcome measures changed very little (because the range was small at the start of the Demonstration).

3.6 Community-Acquired Pneumonia (PN)

This section presents results for the analyses of the process and outcome measures for PN among Demonstration hospitals reporting 16 quarters of data. **Table 27** shows the measures and the number of hospitals included in the analyses (the number of hospitals with 16 quarters of data, between 214 and 220). A total of 214 hospitals reported 16 quarters of data on Smoking cessation advice/counseling, and 220 hospitals reported 16 quarters of data on both Oxygenation assessment and Pneumococcal vaccination.

RTI, International received hospital data for the fall and winter quarters of Year 1 (Q1 and Q2 of the Demonstration), and for the fall and winter quarters of Year 4 (Q13 and Q14 of the Demonstration). The results for the influenza vaccination measure are presented in Section 4.

Figure 13
Demonstration: Trends in HK measures, [a, b] standard deviations Years 1–4



HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Table 26
Demonstration: Trends in HK measures, [a, b] range Years 1–4

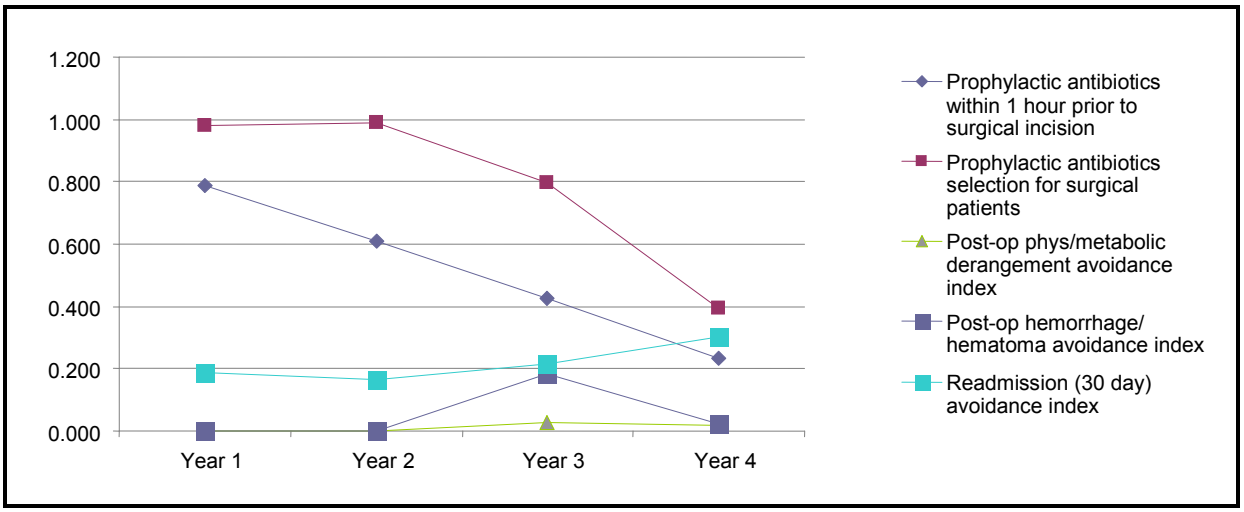
HK measure	Range Year 1	Range Year 2	Range Year 3	Range Year 4	% change Year 1 to Year 4
Prophylactic antibiotics within 1 hour prior to surgical incision (n=165)	0.789	0.608	0.426	0.234	−70.34%
Prophylactic antibiotics selection for surgical patients (n=162)	0.982	0.991	0.799	0.394	−59.88%
Post-op phys/metabolic derangement avoidance index (n=151)	0.001	0.001	0.025	0.016	1,500.00%
Post-op hemorrhage/hematoma avoidance index (n=166)	0.000	0.000	0.185	0.025	N/A
Readmission (30 day) avoidance index (n=149)	0.189	0.164	0.214	0.300	58.73%

HK = hip/knee replacement; NA = not applicable.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 14
Demonstration: Trends in HK measures, [a, b] range Years 1–4



HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Table 27
Hospitals with PN measures in all 16 quarters of PHQID.

Pneumonia measure	# hospitals with measures in all 16 quarters
Oxygenation assessment	220
Pneumococcal vaccination	220
Blood culture before first antibiotic	215
Smoking cessation advice/counseling	214
Initial antibiotic within 4 hours of arrival	216

PHQID = Premier Hospital Quality Incentive Demonstration; PN = pneumonia.

3.6.1 Pneumonia Measures Means

Table 28 shows the means for each Pneumonia measure across the first 16 quarters of the Demonstration. These measures are shown for a panel of hospitals that reported the Pneumonia measures in each of the 16 quarters. As shown in **Table 28** and **Figure 15**, the mean for each Pneumonia measure increased during the first 4 years of the Demonstration. Oxygen assessment was highest in the first quarter (0.978) and remained high throughout the quarters.

Table 28
PN measures for Demonstration Years 1–4, means [a, b]

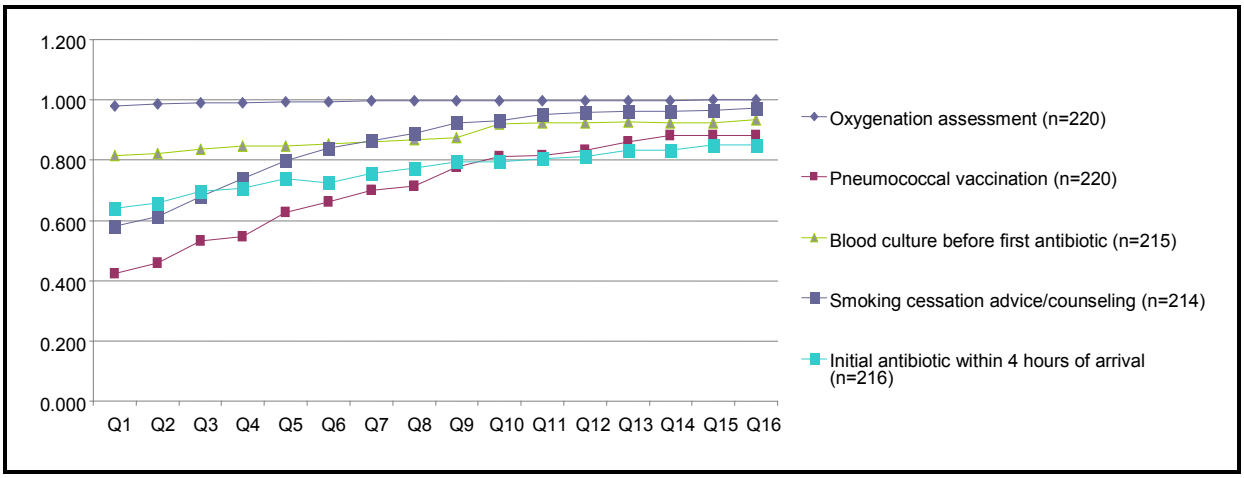
Measure code	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Oxygenation assessment (n=220)	0.978	0.986	0.989	0.991	0.993	0.995	0.997	0.996	0.996	0.997	0.997	0.999	0.998	0.999	0.999	0.999
Pneumococcal vaccination (n=220)	0.422	0.459	0.532	0.547	0.625	0.660	0.699	0.713	0.775	0.813	0.816	0.832	0.860	0.882	0.880	0.883
Blood culture before first antibiotic (n=215)	0.816	0.821	0.836	0.846	0.846	0.854	0.861	0.867	0.874	0.921	0.925	0.923	0.926	0.925	0.925	0.934
Smoking cessation advice/counseling (n=214)	0.582	0.612	0.679	0.740	0.797	0.838	0.864	0.889	0.924	0.931	0.951	0.957	0.960	0.963	0.966	0.972
Initial antibiotic within 4 hours of arrival (n=216)	0.642	0.658	0.697	0.707	0.738	0.723	0.757	0.775	0.796	0.793	0.805	0.812	0.833	0.832	0.849	0.848

PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 15
Demonstration: Trends in PN measures, [a, b] Q1–Q16



PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

To test whether the Pneumonia measure differences between Q1 and Q16 were significantly different from zero, we performed paired t-tests using the Bonferroni correction for multiple testing. Results are shown in **Table 29** and **Figure 16**. All Pneumonia measures increased at a significant rate between Q1 and Q16. The largest percentage increase was noted for Pneumococcal vaccination, which increased by 109% (from 0.442 in Q1 to 0.883 in Q16). The measure Oxygenation assessment showed the smallest, but statistically significant increase, 2.0%, between Q1 and Q16 (yet was essentially 100% by the end of Year 4).

Results from these trend tests (rates of increase) are shown in **Table 30** and **Figure 17**. During Year 1, all of the Pneumonia measures increased significantly between Q1 and Q4. Of these, the Smoking cessation measure increased at the fastest rate during Q1–Q4 (0.158 points). By Years 2 and 3, all measures except Oxygenation assessment continued to increase at a significant rate (but at a smaller rate than during Year 1). By Year 4, only Pneumococcal vaccination continued to increase at a significant rate between Q13 and Q16; the other measures had stopped increasing significantly.

Table 29
Demonstration: Change in PN measures, [a, b] Q1 and Q16

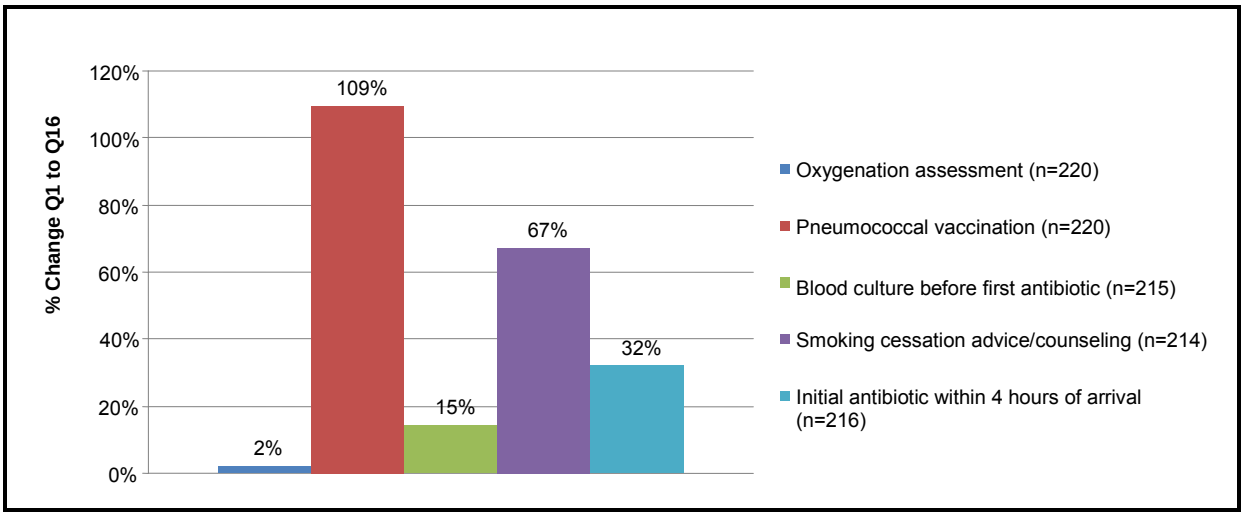
Measure code	Q1	Q16	% change Q1–Q16	Absolute difference Q1 to Q16	*p*-value
Oxygenation assessment (n=220)	0.978	0.999	2%	0.021	<.001
Pneumococcal vaccination (n=220)	0.422	0.883	109%	0.461	<.001
Blood culture before first antibiotic (n=215)	0.816	0.934	15%	0.119	<.001
Smoking cessation advice/ counseling (n=214)	0.582	0.972	67%	0.391	<.001
Initial antibiotic within 4 hours of arrival (n=216)	0.642	0.848	32%	0.207	<.001

PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 16
Demonstration: Percentage change in PN measures, [a, b] Q1 and Q16



PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Table 30
Demonstration: Trends in rates of change in PN measures, [a, b] Q1 and Q16

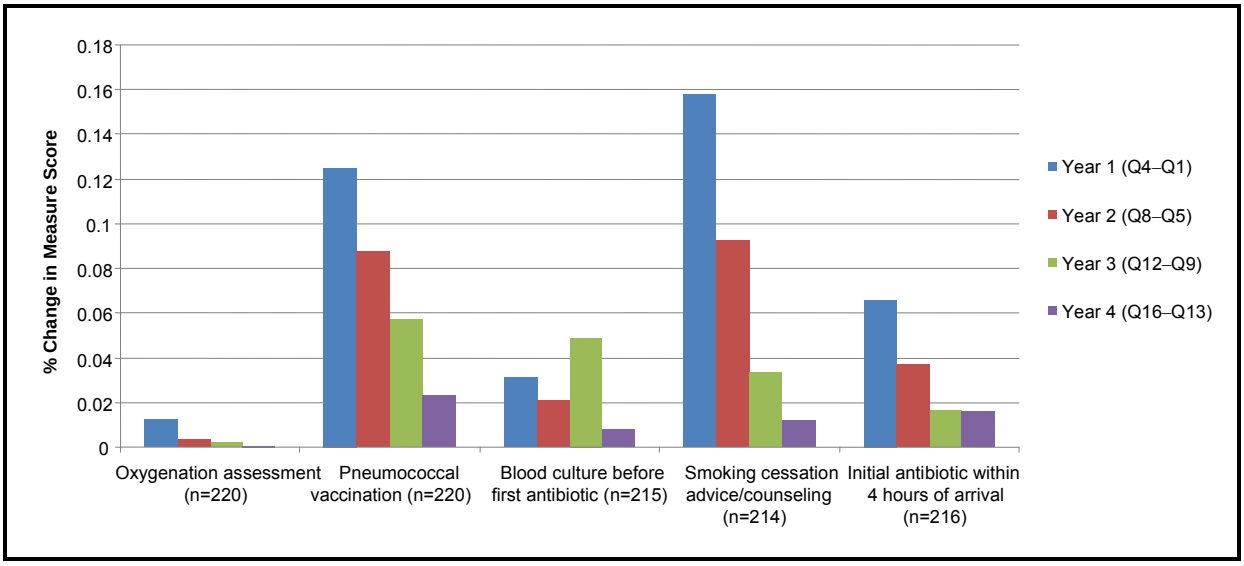
Measure code	Q4–Q1	*p*-value	Q8–Q5	*p*-value	Q12–Q9	*p*-value	Q16–Q13	*p*-value	Q4–Q1 vs Q16–Q13	*p*-value
Oxygenation assessment (n=220)	0.013	<.001	0.003	Ns	0.002	NS	0.001	NS	0.012	<.001
Pneumococcal vaccination (n=220)	0.125	<.001	0.087	<.001	0.057	<.001	0.023	<.001	0.102	<.001
Blood culture before first antibiotic (n=215)	0.031	<.001	0.021	<.02	0.049	<.001	0.008	NS	0.023	<.01
Smoking cessation advice/counseling (n=214)	0.158	<.001	0.093	<.001	0.033	<.02	0.012	NS	0.146	<.001
Initial antibiotic within 4 hours of arrival (n=216)	0.066	<.001	0.037	<.001	0.016	<.001	0.016	NS	0.050	<.001

PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 17
Demonstration: Trends in rates of change in PN measures, [a, b] Q1 and Q16 (Years 1–4)



PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

We examined whether the measures' rates of change between Q1 and Q4 were different from the rates of change between Q13 and Q16 (i.e., a "difference in difference" between Year 1 and Year 4). Significant differences were noted for all Pneumonia measures. For these measures, each difference was positive and significant, which indicates that the rates of change during Year 1 were significantly faster than the rates of change during Year 4. The largest rate of increase difference was noted for Smoking cessation advice/counseling.

3.6.2 Pneumonia Measures Standard Deviation and Range

Table 31 and **Figure 18** display the standard deviation in the Pneumonia measures during Years 1–4 of the Demonstration. As the table shows, the Pneumonia measures became less variable over the first 4 Demonstration years. The largest percentage change in standard deviations is for Oxygenation assessment measure, followed by the measure for Smoking cessation advice/counseling. Decreases in the standard deviation of the measures indicate that performance among hospitals is becoming less variable over time.

Table 31
Demonstration: Changes in PN measures, [a, b] Years 1–4, standard deviation

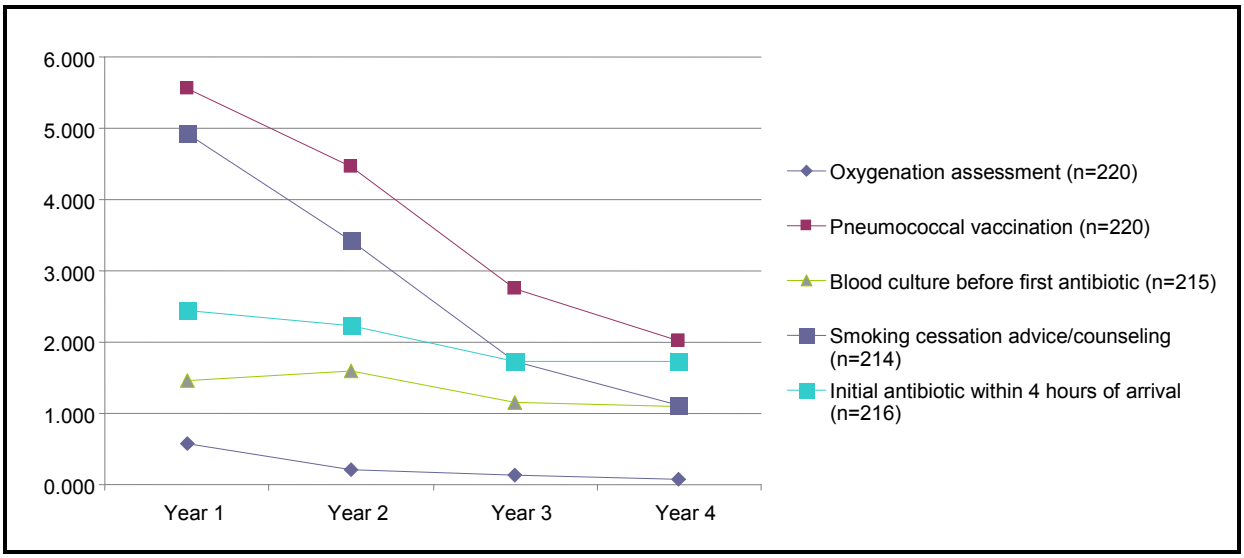
Measure code	Standard deviation Year 1	Standard deviation Year 2	Standard deviation Year 3	Standard deviation Year 4	% change Year 1 to Year 4
Oxygenation assessment (n=220)	0.574	0.217	0.134	0.071	−87.71%
Pneumococcal vaccination (n=220)	5.563	4.453	2.749	2.023	−63.63%
Blood culture before first antibiotic (n=215)	1.469	1.598	1.144	1.104	−24.85%
Smoking cessation advice/ counseling (n=214)	4.930	3.414	1.740	1.121	−77.26%
Initial antibiotic within 4 hours of arrival (n=216)	2.440	2.235	1.723	1.733	−28.96%

PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 18
Demonstration: Changes in PN measures [a, b] Years 1–4, standard deviation



PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

The range in Pneumonia measures is the difference between the maximum and the minimum for all hospitals. Similar to the standard deviation, a smaller range indicates less spread (variation) across the hospitals in terms of their Pneumonia measure. **Table 32** and **Figure 19** show the ranges for Years 1–4 for every Pneumonia measure, as well as the percentage change in the range between Year 1 and Year 4. Every Pneumonia measure decreased during the range between Year 1 and Year 4, most notably Oxygenation assessment (87.6% decrease) and Smoking cessation advice/counseling (67.6% decrease).

3.7 Composite Quality Scores

This section presents results for the analyses of the CQS among hospitals that report 16 quarters of data. Individual hospital process and outcome measures for each clinical focus area are aggregated into a composite score that is used to establish relative performance among hospitals participating in the Demonstration. Details on the methodology by which the CQS is computed can be found at https://www.cms.gov/HospitalQualityInits/35_HospitalPremier.asp.

Table 32
Demonstration: Changes in PN measures, [a, b] Years 1–4, range

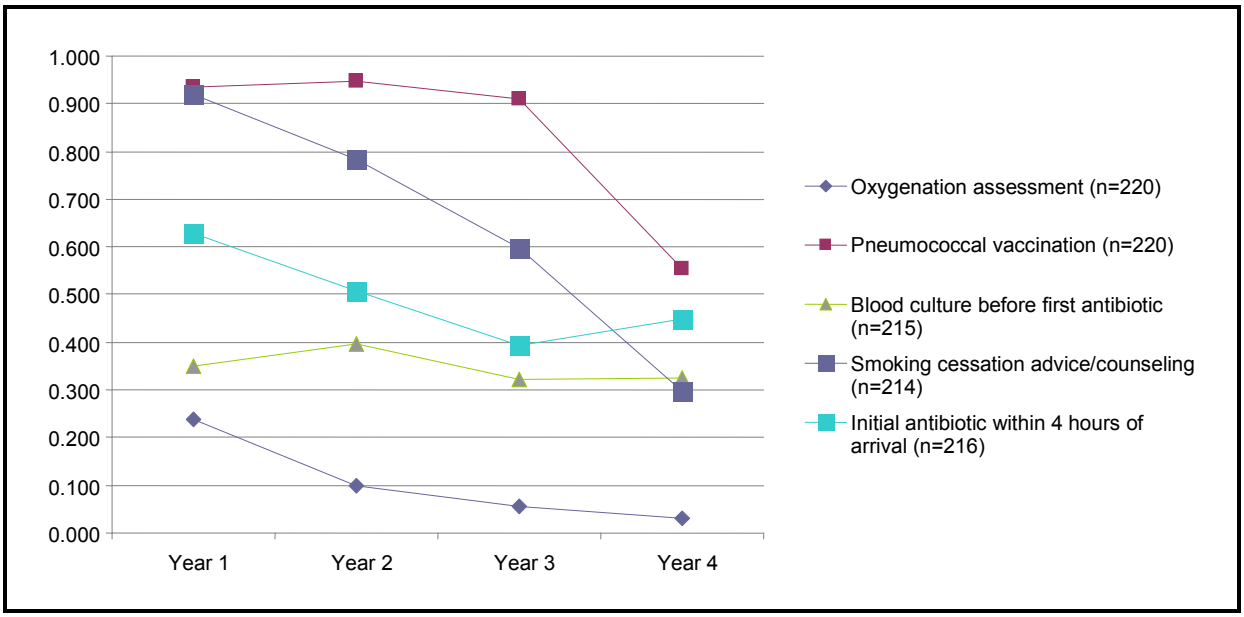
Measure code	Range Year 1	Range Year 2	Range Year 3	Range Year 4	% change Year 1 to Year 4
Oxygenation assessment (n=220)	0.237	0.099	0.057	0.029	−87.61%
Pneumococcal vaccination (n=220)	0.935	0.948	0.909	0.556	−40.59%
Blood culture before first antibiotic (n=215)	0.351	0.397	0.321	0.326	−7.11%
Smoking cessation advice/ counseling (n=214)	0.919	0.783	0.597	0.297	−67.70%
Initial antibiotic within 4 hours of arrival (n=216)	0.630	0.507	0.394	0.447	−28.97%

PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Figure 19
Demonstration: Changes in PN measures [a, b] Years 1–4, range



PN = pneumonia.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

[b] Hospital scores were weighted by the denominator when computing the measures.

Table 33 shows the CQS for each clinical focus area and the number of hospitals included in the analyses (the number of hospitals with 16 quarters of data, between 112 and 193). A total of 112 hospitals reported 16 quarters of data on PCI received within 120 (90) minutes of arrival, and 193 hospitals reported 16 quarters of data on Beta blocker at arrival. Recall that the difference in the N's by measure is due to the fact that not all hospitals provided care for each clinical area.

Table 33
Hospitals with CQS measures in all 16 quarters of PHQID

Clinical focus area	N hospitals with CQS In Year 4	N (%) hospitals with CQS In all 16 quarters
AMI	207	194 (93.7%)
CABG	116	109 (93.7%)
HF	224	215 (96.0%)
HK	215	164 (76.3%)
PN	222	220 (99.1%)

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement; PHQID = Premier Hospital Quality Incentive Demonstration.

3.7.1 CQS Means

Table 34 displays the means for the CQS across 16 quarters of Demonstration Years 1–4. The CQS is displayed for five clinical focus areas for a panel of hospitals that reported CQS measures in each of the 16 quarters. **Figure 20** details in graphical form the trends in CQS measures across the 16 quarters.

As shown in Table 34 and Figure 20, each clinical focus area increased its average CQS score during Demonstration Years 1 to 4. HF was the focus area showing the greatest improvement. In Q1, hospitals had an average CQS for HF of 67.2%. By the end of Year 4 (Q16), hospitals had an average CQS for HF of 92.7%. Nonetheless, by the end of Year 4, the mean for the HF CQS remained the lowest along with the PN CQS while the highest CQS was for CABG at 98.8%.

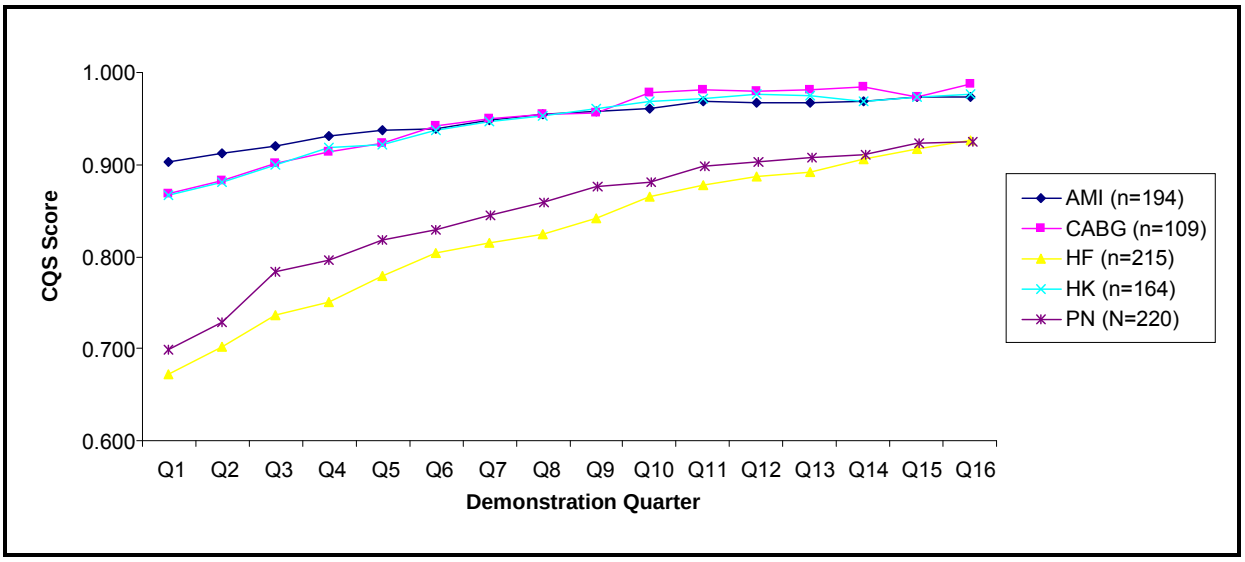
Table 34
CQS Measures for Demonstration Years 1–4, [a] means

Clinical focus	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
AMI (n=194)	0.902	0.912	0.921	0.930	0.937	0.938	0.948	0.954	0.958	0.962	0.969	0.967	0.966	0.969	0.974	0.973
CABG (n=109)	0.869	0.883	0.901	0.914	0.924	0.941	0.950	0.955	0.956	0.978	0.980	0.980	0.982	0.985	0.973	0.988
HF (n=215)	0.672	0.702	0.736	0.751	0.779	0.804	0.816	0.824	0.841	0.865	0.878	0.887	0.892	0.906	0.917	0.927
HK (n=164)	0.867	0.882	0.900	0.919	0.921	0.937	0.947	0.953	0.961	0.969	0.973	0.977	0.975	0.968	0.973	0.976
PN (N=220)	0.699	0.728	0.783	0.796	0.819	0.829	0.845	0.859	0.877	0.881	0.898	0.903	0.908	0.911	0.924	0.924

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

Figure 20
CQS measures for Demonstration Years 1–4, [a] means



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

To test whether differences in CQS between Q1 and Q16 were significantly different from zero, we performed paired t-tests using the Bonferroni correction for multiple testing. Results are shown in **Table 35**. For each clinical focus area, the mean CQS at Q1 and Q16, the percentage change in the mean between Q1 and Q16, the absolute difference between Q1 and Q16, and the *p*-value for the test for whether the absolute difference was significantly different from zero.

Table 35
T-tests for differences in CQS measures during Q1 and Q16 of Demonstration [a]

Clinical focus area	Q1	Q16	% change Q1–Q16	Absolute difference Q1 to Q16	*p*-value
AMI (n=194)	0.902	0.973	7.9%	0.07	*p*<.0001
CABG (n=109)	0.869	0.988	13.7%	0.12	*p*<.0001
HF (n=215)	0.672	0.927	37.9%	0.25	*p*<.0001
HK (n=164)	0.867	0.976	12.6%	0.11	*p*<.0001
PN (N=220)	0.699	0.924	32.2%	0.23	*p*<.0001

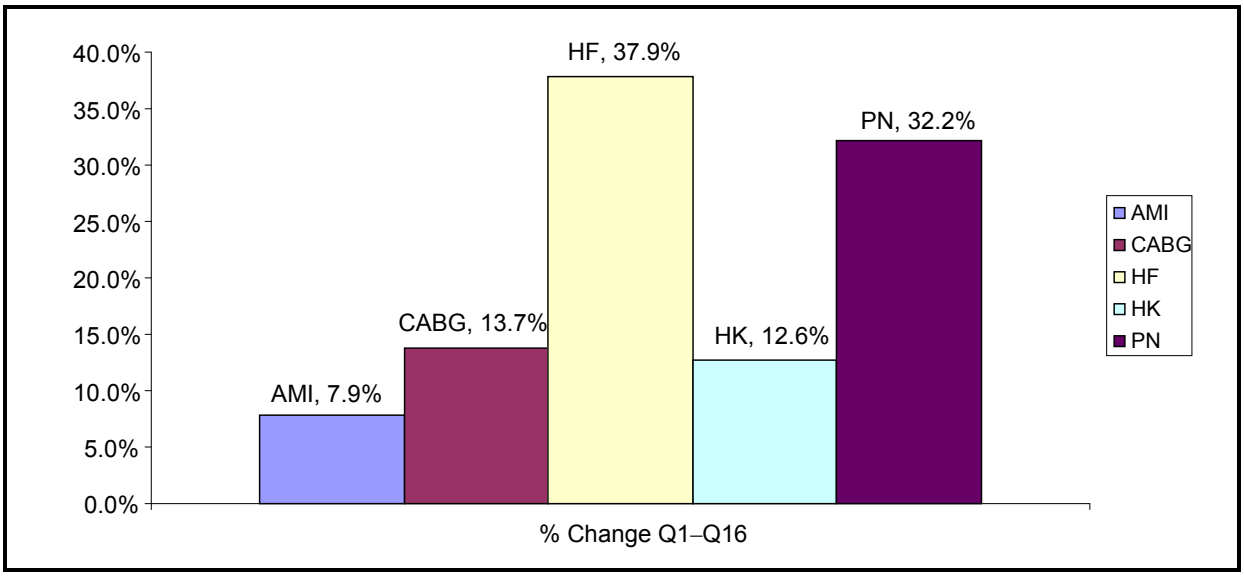
AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration.

As also shown in Table 35, all CQS measures experienced statistically significant increases during the 4 years of the Demonstration. AMI increased from 0.902 to 0.973, an absolute difference of 0.07 points and a percentage increase of 7.9%. CABG increased from 0.869 to 0.988, an absolute increase of 0.12 points, and a percentage increase of 13.7%. HF increased from 0.672 to 0.927, an absolute increase of 0.25 points and a percentage increase of 37.9% (the largest percentage increase among all clinical focus areas). HK increased from 0.867 to 0.976, an absolute increase of 0.11 points and a percentage increase of 12.6%. PN increased from 0.699 to 0.924, an absolute increase of 0.23 points and a percentage increase of 32.2%. **Figure 21** displays graphically the percentage change in CQS measures between Q1 and Q16. Note that the percentage change in CQS between Q1 and Q16 was most dramatic for HF and PN CQS.

As the results for the CQS measures show, clinical quality scores improved significantly during the first 4 years of the PHQID. In addition to examining if a statistically significant difference existed between Q1 and Q16 for each CQS measure, we also tested for trends in the measures to examine whether there were differential rates of improvement across the different quarters or years of the Demonstration.

Figure 21
Demonstration Years 1–4: CQS percentage change Q1–Q16 [a]



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes hospitals with measures in all 16 quarters of the Demonstration

Results from these trend tests are shown in **Table 36**. In this table the first eight columns show the results from paired t-tests between the first and last quarters of Years 1–4, to examine the rate at which CQS improvement was happening each year and whether the improvement was significantly different from zero. These results are also displayed graphically in **Figure 22**. The last two columns in the table show the result of the paired t-test (difference-in-difference test) examining whether the rate of change in CQS between Q1 and Q4 (i.e., during Year 1), was significantly different from the rate of change in CQS between Q13 and Q16 (i.e., during Year 4).

During Year 1 (Q4–Q1), every CQS measure increased significantly. The CQS rate for PN increased at a faster rate than the other CQS measures (0.097 points between Q1 and Q4), and AMI increased at a slower rate than the other CQS measures (0.028 points between Q1 and Q4). Nonetheless, all changes during Year 1 were significant (p<.001).

Similar results were noted for changes during Year 2 (Q8–Q5), although the increases were at a slower rate than during Year 1. For example, the CQS rate for PN increased by 0.097 points during Year 1, and 0.040 points during Year 2. Similarly, the CQS rate for CABG was 0.045 in Year 1 and 0.032 in Year 2. Nonetheless, all of the changes during Year 2 were significant (p<.001).Changes in CQS measures during Year 3 (Q12–Q9) were all significant as well, but at a slower rate than was noted for changes during Years 1 and 2. For example, the CQS rate for AMI increased by 0.028 points during Year 1, 0.017 points during Year 2, and 0.009 points during Year 3. Similar trends were noted for the other clinical focus areas: significant, but smaller increases, in Year 3 compared with Years 1 and 2.

Table 36
Demonstration: Trends in CQS, means Years 1–4

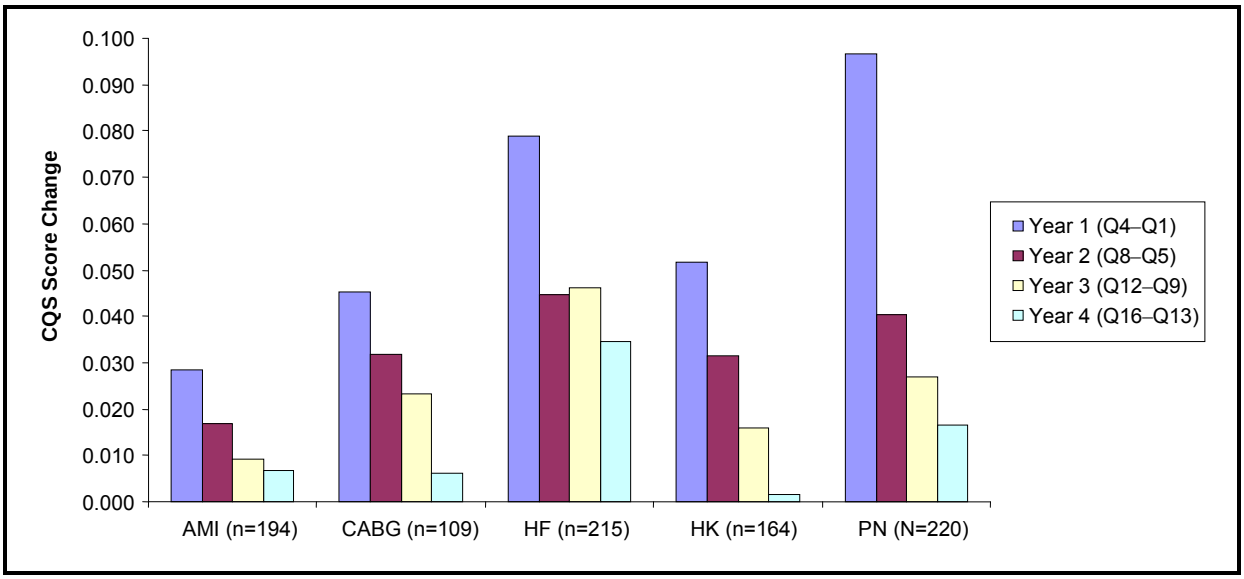
Clinical Focus Area (n hospitals)	Year 1 Q4–Q1	*p*-value	Year 2 Q8–Q5	*p*-value	Year 3 Q12–Q9	*p*-value	Year 4 Q16–Q13	*p*-value	Year 1 Q4–Q1 vs. Year 4 Q16–Q13	*p*-value
AMI (n=194)	0.028	p<.001	0.017	p<.001	0.009	p<.05	0.007	NS	0.022	*p*<.0001
CABG (n=109)	0.045	p<.001	0.032	p<.001	0.023	p<.001	0.006	NS	0.039	p<.0001
HF (n=215)	0.079	p<.001	0.045	p<.001	0.046	p<.001	0.035	p<.001	0.044	p<.0001
HK (n=164)	0.052	p<.001	0.032	p<.001	0.016	p<.001	0.001	NS	0.050	p<.0001
PN (N=220)	0.097	p<.001	0.040	p<.001	0.027	p<.001	0.017	p<.001	0.080	p<.0001

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

By Year 4, however, the CQS rate changed between the quarters during the year (Q16–Q13) and was nonsignificant for three of the five clinical focus areas: AMI, CABG, and HK. The CQS measures essentially did not increase during Year 4 for these three clinical focus groups. In other words, for AMI, CABG, and HK, the rate of improvement in CQS flattened out after Year 3 and was no longer improving. However, CQS rates for the other two clinical focus areas, HF and PN, were still increasing during Year 4 at a statistically significant rate. The CQS for HF increased by 0.035 points between Q13 and Q16, and the CQS for PN increased by 0.017 points between Q13 and Q16. These rates were slower than the rates noted for the other Demonstration years, however, which indicates that the rate of improvement in CQS scores declined over time.

The final two columns in Table 36 show the results from the difference-in-difference analysis between Years 1–4 of the Demonstration. All measures are positive and significant, which indicates that for every clinical focus area, the rate of increase in CQS measure was higher during Year 1 than during Year 4 (i.e., the rate of change has slowed across time during the PHQID).

Figure 22
Demonstration: Yearly trends in CQS measures



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

3.7.2 CQS Standard Deviation and Range

The standard deviation for each clinical area in the mean CQS gives an indication of the amount of variation in performance among Demonstration hospitals. Larger standard deviations indicate more CQS variation across participating hospitals, and smaller standard deviations indicate less variation. A negative percent change between Year 1 and Year 4 indicates that the standard deviation is decreasing over time, indicating that hospital performance is closer to the mean. **Table 37** and **Figure 23** below display the standard deviation in the CQS measures for each clinical focus area during Years 1–4 of the Demonstration.

As the table shows, the CQS measures have become less variable over the first 4 Demonstration years. All have become less variable by nearly 50% or more (as shown by the % change column). The largest percent change in standard deviations is for the CABG CQS, which decreased from 1.079 in Year 1 to 0.375 in Year 4 (a 65% reduction in the mean CQS between Year 1 and Year 4). During Year 1 of the Demonstration, the HK CQS had the smallest standard deviation (0.823), and by Year 4 the standard deviation for this measure decreased by 48% to 0.425. Decrease in the standard deviation of the mean CQS indicates that the measures are becoming less variable over time and that hospital CQS measures are beginning to "cluster" around the mean by Year 4 more so than in Year 1.

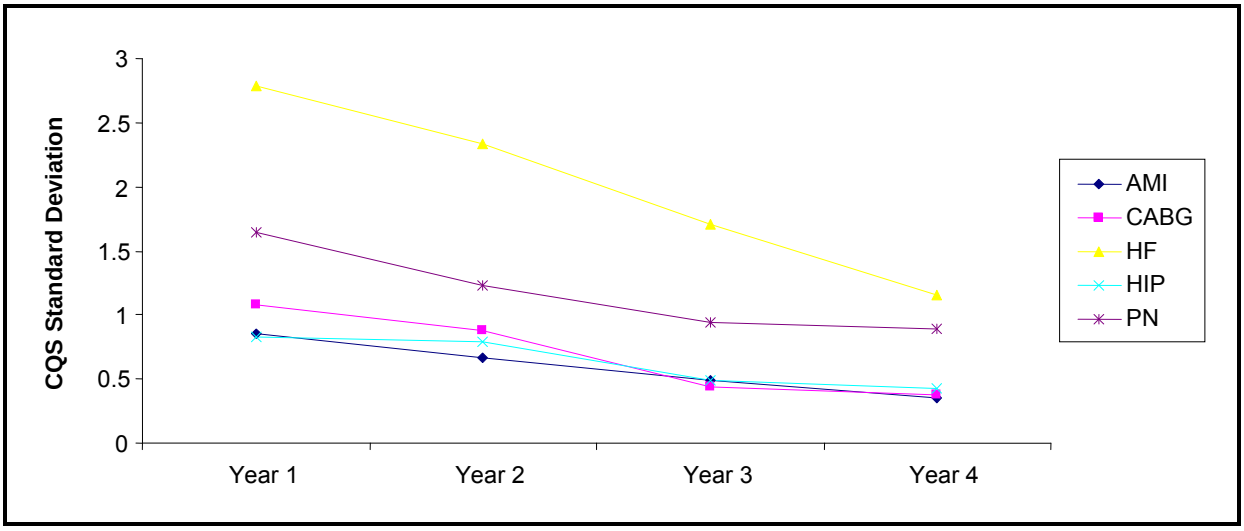
Table 37
Demonstration: Standard deviations in CQS measures Years 1–4 [a]

	Year 1	Year 2	Year 3	Year 4	% change Year 1 to Year 4
AMI (n=194)	0.859	0.660	0.484	0.356	−58.55%
CABG (n=109)	1.079	0.879	0.438	0.375	−65.22%
HF (n=215)	2.788	2.337	1.702	1.156	−58.53%
HK (n=164)	0.823	0.796	0.489	0.425	−48.39%
PN (n=220)	1.644	1.234	0.941	0.896	−45.53%

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes only hospitals with CQS measures in all 16 quarters of Demonstration.

Figure 23
Demonstration: CQS Standard Deviations Years 1–4 [a]



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes only hospitals with CQS measures in all 16 quarters of Demonstration.

The "range" in CQS measures is the difference between the maximum CQS measure score and the minimum CQS among participating hospitals. Similar to the standard deviation, a smaller range indicates that there is less spread (variation) across the hospitals in the CQS measure. **Table 38** and **Figure 24** below show the CQS ranges for Years 1–4 by clinical focus area, as well as the percent change in the range between Year 1 and Year 4. The range for each of the clinical focus area decreased between Year 1 and Year 4, most notably AMI (55.4% decrease) and HK (52.7% decrease).

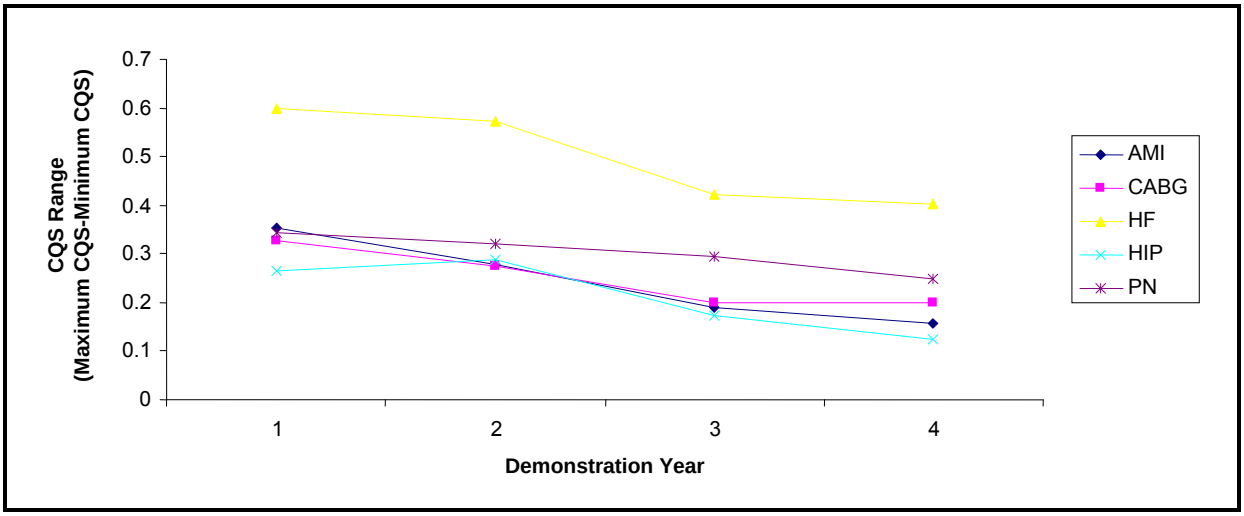
Table 38
Demonstration: CQS range (maximum CQS—minimum CQS) Years 1–4 [a]

	Range Year 1	Range Year 2	Range Year 3	Range Year 4	% change Year 1 to Year 4
AMI (n=194)	0.354	0.279	0.190	0.158	−55.4%
CABG (n=109)	0.326	0.274	0.201	0.198	−39.3%
HF (n=215)	0.600	0.572	0.422	0.403	−32.8%
HK (n=164)	0.264	0.289	0.175	0.125	−52.7%
PN (N=220)	0.345	0.319	0.295	0.248	−28.1%

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

[a] Includes only hospitals with CQS measures in all 16 quarters of Demonstration.

Figure 24
Demonstration: Range in clinical quality score values Years 1–4



AMI = acute myocardial infarction; CABG = coronary artery bypass graft; CQS = composite quality score; HF = heart failure; PN = pneumonia; HK = hip/knee replacement.

SECTION 4
MEDICARE INCENTIVE PAYMENTS

During the first 3 years of the Demonstration, hospitals were eligible for an incentive payment for each clinical area based on performance. In a given fiscal year, top performers receiving an incentive were those whose clinical area specific composite score was in the top 20% (or 1st and 2nd decile). Beginning in the third year of the Demonstration, hospitals were penalized for poor performance, defined as being in the bottom two deciles in a given clinical area relative to the base year (2 years prior). Beginning in Year 4, the incentive system was changed substantially. Incentives for top performers continued; however, there were also awards based on threshold attainment and significant improvement. Furthermore, beginning in Year 4 penalties were not assessed for hospitals with a CQS score of at least 85%. Performance awards are defined as:

- **Median-Level Attainment Award:** Hospitals that attain or exceed the median-level CQS in a given clinical area, as measured 2 years prior for each clinical area receive an incentive payment. For example, a hospital will achieve median-level attainment in Year 4 if that hospital's CQS in Year 4 exceeds the median CQS for all participating hospitals in Year 2.

- **Top Performer Award:** Hospitals that have a CQS for a given clinical area that falls in the top 20% of scores receive an additional payment. These hospitals receive the median level attainment award as well.

- **Top Improvement Award:** Hospitals that attain median-level performance and are among the top 20% of hospitals with the largest percentage quality improvement in a clinical area receive an additional incentive payment. Improvements are calculated based on the performance year compared to 2 years prior. This group receives the median-level attainment award as well. Top performers are not eligible for the improvement award.

In this Phase I report we focus on the first 4 years of the Demonstration, which includes the first year of this change in pay for performance awards. The evaluation of the first 3 years is summarized in an Abt report (Kennedy et al., 2008) so it is not repeated here. This report focuses on the distribution of awards in this new award method as well as change in the decile distribution. In Phase II of the evaluation, we will be able to look at all 6 years of the Demonstration and detect change in performance associated with the change in award incentives.

It is notable that the number of hospitals receiving awards increased dramatically during the fourth year of the Demonstration. The number of hospitals receiving awards in Years 1–4 were 123 (Year 1), 115 (Year 2), 112 (Year 3), and 225 (Year 4). It is not surprising that more hospitals received awards in Year 4, given the substantial change to the incentive system.

4.1 Decile Analyses: Movement of Hospitals in Deciles between Year 1 and Year 4

For this analysis, the sample included all hospitals that reported a CQS measure in Year 1. We examined decile distribution in Year 1, particularly among the top two and bottom

two deciles, and compared these hospitals' decile placement in Year 4. We also looked at continuation in the Demonstration in Year 4 by decile distribution in Year 1.

Table 39 below shows the decile rankings in Year 4 for hospitals that were in the top two deciles in Year 1. Between 28% (HK) and 39% (PN) of hospitals in the top two deciles in Year 1 remained in the top 20% in Year 4. Moreover, hospitals in the top 20% in Year 1 were highly *unlikely* to drop to the bottom 20% by Year 4. Only 2% to 8% of hospitals dropped from the top two deciles in Year 1 to the bottom two deciles by Year 4.

As **Table 39** also shows, few hospitals in the top 20% in Year 1 were missing a CQS score in Year 4—i.e., these hospitals are highly likely to remain in the Demonstration between Year 1 and Year 4; however, this varied by clinical focus group. Only 4% of hospitals in the top two deciles for PN CQS in Year 1 were missing a Pneumonia CQS measure in Year 4 and a high of 15% of hospitals in the top two deciles for CABG CQS in Year 1 no longer had a CABG CQS by Year 4. Although the number of hospitals missing a Year 4 CABG CQS score is similar to the number missing CQS scores by Year 4 for other clinical areas (see **Table ES-2**), e.g., 4 hospitals missing Year 4 data for CABG and for AMI, the percentage is higher as there are fewer hospitals with a CABG clinical focus area (134 at Year 1 for CABG vs. 243 for AMI).

Table 39
Hospitals in deciles 1 and 2 in Year 1 (top 20%): Comparison of decile ranking in Year 1 and Year 4

Clinical focus area	N hospitals with decile rankings Year 1	N hospitals in deciles 1 or 2 Year 1	Deciles 1 or 2 (top 20%) Year 4	Deciles 3 to 8 (middle 60%) Year 4	Deciles 9 or 10 (bottom 20%) Year 4	Missing Year 4[a]
AMI	243	49	18 (36.7%)	24 (49.0%)	3 (6.1%)	4 (8.2%)
CABG	134	27	10 (37.0%)	12 (44.4%)	1 (3.7%)	4 (14.8%)
HF	259	52	15 (28.9%)	28 (53.9%)	4 (7.7%)	5 (9.6%)
HK	214	43	12 (27.9%)	26 (60.5%)	1 (2.3%)	4 (9.3%)
PN	261	52	20 (38.5%)	28 (53.9%)	2 (3.9%)	2 (3.9%)

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Missing data may be due to a variety of issues including: hospital withdrawal from the Demonstration, insufficient volume for the timeframe (<30 cases for the year), and/or hospitals failing validation.

As **Table 40** shows, in addition to examining the hospitals in the top two deciles in Year 1, we also examined hospitals in the bottom two deciles in Year 1 to see whether they were more or less likely to improve deciles by Year 4. Of the hospitals that were in the bottom 20% in terms of CQS score in Year 1, between 23% (HK) and 40% (CABG) remained in the bottom 20% of hospitals in Year 4. Relatively few hospitals improved enough to be in the top two deciles in Year 4—but 11% of both HK and CABG hospitals improved enough to move from the bottom 20% in Year 1 to the top 20% in Year 4. For the other clinical areas, the percentage moving from the bottom to the top deciles between Years 1 and 4 was much smaller.

As Table 40 also shows, in general, with the exception of HK, hospitals that were in the bottom 20% in Year 1 were much more likely to be missing from the Demonstration in Year 4, as compared with hospitals in the top 20% in Year 1. Not including HK, between 25% (PN) and 35% (AMI) of hospitals that were in the bottom two deciles in Year 1 were missing CQS scores for those clinical focus areas by Year 4. This indicates that hospitals that were the bottom performers early in the Demonstration were more likely not to continue participating in the Demonstration by Year 4, as compared with hospitals that were top performers early in the Demonstration.

Table 40
Hospitals in deciles 9 and 10 in Year 1 (bottom 20%): Comparison of decile ranking in Year 1 and Year 4

Clinical focus area	N hospitals with decile rankings Year 1	N hospitals in deciles 9 or 10 Year 1	Deciles 1 or 2 (top 20%) Year 4	Deciles 3 to 8 (middle 60%) Year 4	Deciles 9 or 10 (bottom 20%) Year 4	Not in Demonstration Year 4[a]
AMI	243	49	2 (4.1%)	15 (30.6%)	15 (30.6%)	17 (34.7%)
CABG	134	27	3 (11.1%)	4 (14.8%)	11 (40.7%)	9 (33.3%)
HF	259	52	4 (7.7%)	20 (38.5%)	14 (26.9%)	14 (26.9%)
HK	214	43	5 (11.6%)	15 (34.9%)	10 (23.3%)	1 (2.3%)
PN	261	52	3 (5.8%)	23 (44.2%)	13 (25.0%)	13 (25.0%)

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Missing data may be due to a variety of issues including: hospital withdrawal from the Demonstration, insufficient volume for the timeframe (<30 cases for the year), and/or hospitals failing validation.

4.2 Decile Analyses: Changing of CQS across Deciles: Years 1–4

In addition to examining the movement of hospitals through deciles between Years 1–4, we also performed two additional decile analyses:

a. Change in the range of values used to determine each decile changed between Year 2 and Year 4, and

b. Change in the mean CQS for each decile during the 4 years of the Demonstration.

The results of these analyses are described below.

4.2.1 Decile Ranges Year 2 and Year 4

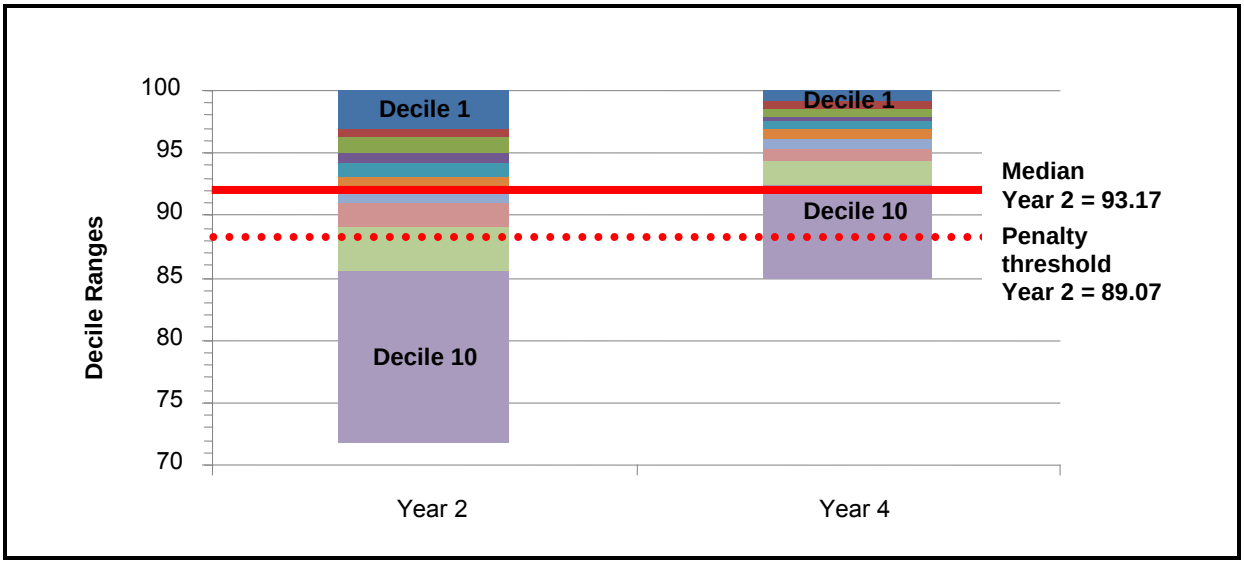
In this analysis, we compare the decile threshold ranges for Years 2 and 4. This analysis was performed using the decile thresholds that were provided for Years 1–4 on the Premier Web site: (http://www.premierinc.com/p4p/hqi/year4/decile-threshold-year-4.pdf). The decile threshold is the CQS that defines the upper limit of that particular decile. This analysis was designed to examine how the upper and lower values in each threshold changed over time. Results for each clinical focus area are described below using figures. In the graphs we show

stacked bars representing the deciles for Year 2 and Year 4. We also show a solid red line that represents the median for Year 2, because this value is used to determine what hospitals received the attainment award in Year 4. Finally, we also show a dotted red line that represents the value below which hospitals in Year 4 were penalized. Penalties in Year 4 were assessed based on the value of the ninth and tenth deciles in Year 2, and penalties were not assessed for hospitals scoring at least 85%. **Figures 25–29** show these results. These figures depict the reference deciles set based on the Year 2 CQS scores.

4.2.1.1 Decile Ranges: AMI

As **Figure 25** shows, the ranges in AMI decile thresholds have been compressed significantly between Year 2 and Year 4. In Year 2, the upper limit for decile 10, indicating low performance, was 85.51, and by Year 4 it was 92.44. The lower limit for decile 2, which was the lower limit for receiving a high performer award, was 96.27 in Year 2 and 98.55 in Year 4.

Figure 25
Demonstration: Decile ranges Year 2 and Year 4: AMI [a]



[a] The figure depicts the reference deciles set based on the Year 2 CQS scores. Beginning in Year 4, penalties were not assessed for hospitals scoring at least 85%.

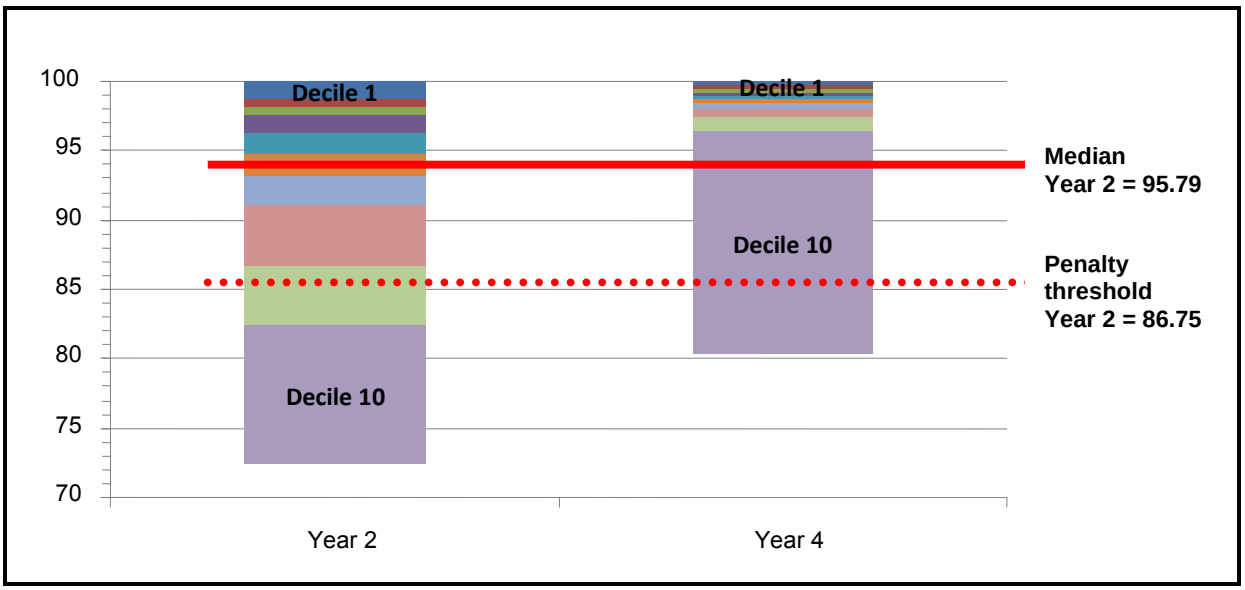
The median value in Year 2 was 93.16 so that any hospital in Year 4 with an AMI CQS score of at least 93.16 received the attainment bonus. As this solid red line shows, this value in Year 4 represents virtually all hospitals except those in the tenth decile. The dotted red line indicates that hospitals in Year 4 that had an AMI CQS score of 89.07 or below paid a penalty. However, as previously noted, CMS set a cap of 85% starting in Year 4, so that penalties were not assessed for hospitals that had an HK CQS score of at least 85%.

4.2.1.2 Decile Ranges: CABG

As **Figure 26** shows, the ranges in CABG decile thresholds have also been compressed significantly between Year 2 and Year 4. In Year 2, the upper limit for decile 10 was 82.47, and

by Year 4 had increased to 96.44. The lower limit for decile 2, which was the lower limit for receiving a high performer award, was 98.09 in Year 2 and 99.49 in Year 4. Interestingly, the difference in threshold between decile 2 and decile 3 (i.e., the difference between receiving a high performer award and not receiving one) decreased between Year 2 (0.6915) and Year 4 (0.2277). This indicates that there is little difference between CABG CQS scores for hospitals in decile 2 and decile 3 in Year 4.

Figure 26
Demonstration: Decile ranges Year 2 and Year 4: CABG [a]



[a] The figure depicts the reference deciles set based on the Year 2 CQS scores. Beginning in Year 4, penalties were not assessed for hospitals scoring at least 85%.

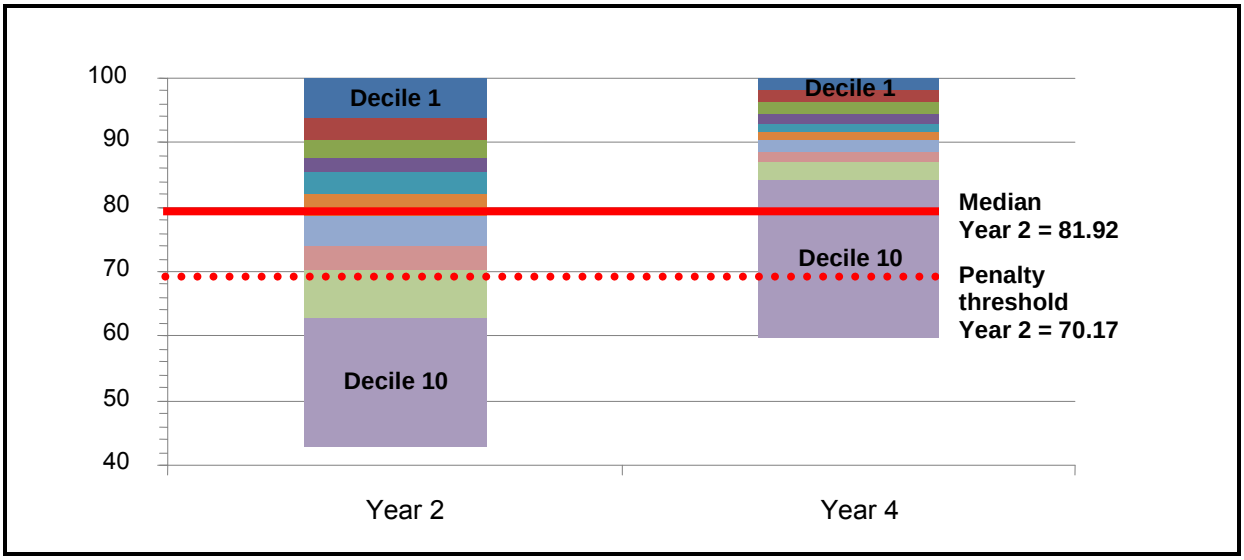
The median value in Year 2 was 95.79, so that any hospital in Year 4 with a CABG CQS score of at least 95.79 received the Attainment bonus. As this solid red line shows, this value in Year 4 represents virtually all hospitals except those in the tenth decile. The dotted red line indicates that hospitals in Year 4 that had a CABG CQS score of 86.75 or below paid a penalty. However, as previously noted, penalties were not assessed for hospitals that had an CABG CQS score of at least 85% in Year 4.

4.2.1.3 Decile Ranges: HF

As **Figure 27** shows, the ranges in HF decile thresholds have also been compressed significantly between Year 2 and Year 4. In Year 2, the upper limit for decile 10 was 62.73 (which was lowest among all clinical focus areas), and by Year 4 had increased to 84.17. The lower limit for decile 2, which was the lower limit for receiving a high performer award, was 90.36 in Year 2 and 96.43 in Year 4. Like the decile thresholds for HF, the difference in threshold between decile 2 and decile 3 for HF (the difference between receiving a high performer award and not receiving one) decreased between Year 2 (3.42) and Year 4 (1.75), indicating that the difference between HF CQS scores for hospitals in decile 2 and decile 3 is decreasing over time.

The median value in Year 2 was 81.92. As this solid red line shows, this value in Year 4 indicates that virtually all hospitals received the attainment award except most of those in the tenth decile. The dotted red line indicates that hospitals in Year 4 that had a HF CQS score of 70.17 or below paid a penalty (approximately the lower 1/3 of decile 10 in Year 4).

Figure 27
Demonstration: Decile ranges Year 2 and Year 4: HF [a]

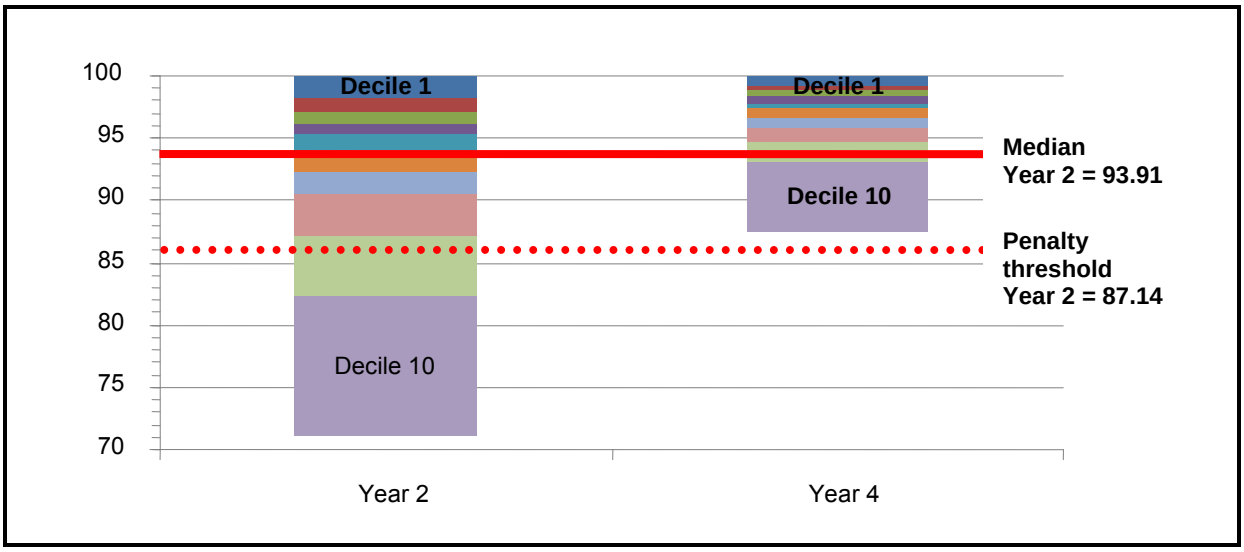


[a] The figure depicts the reference deciles set based on the Year 2 CQS scores. Beginning in Year 4, penalties were not assessed for hospitals scoring at least 85%.

4.2.1.4 Decile Ranges: HK

As **Figure 28** shows, the ranges in HK decile thresholds have been compressed significantly between Year 2 and Year 4 in a similar manner to the other clinical focus areas. In Year 2, the upper limit for decile 10 was 82.37, and by Year 4 had increased to 93.04. The lower limit for receiving a high performer award (decile 2) was 97.13 in Year 2 and increased only to 98.81 in Year 4. Like the decile thresholds for HF and CABG, the difference in threshold between decile 2 and decile 3 for HK (the difference between receiving a high performer award and not receiving one) decreased dramatically between Year 2 (1.10) and Year 4 (0.37). By Year 4 there was very little difference between deciles 2 and 3.

Figure 28
Demonstration: Decile ranges Year 2 and Year 4: HK [a]



[a] The figure depicts the reference deciles set based on the Year 2 CQS scores. Beginning in Year 4, penalties were not assessed for hospitals scoring at least 85%.

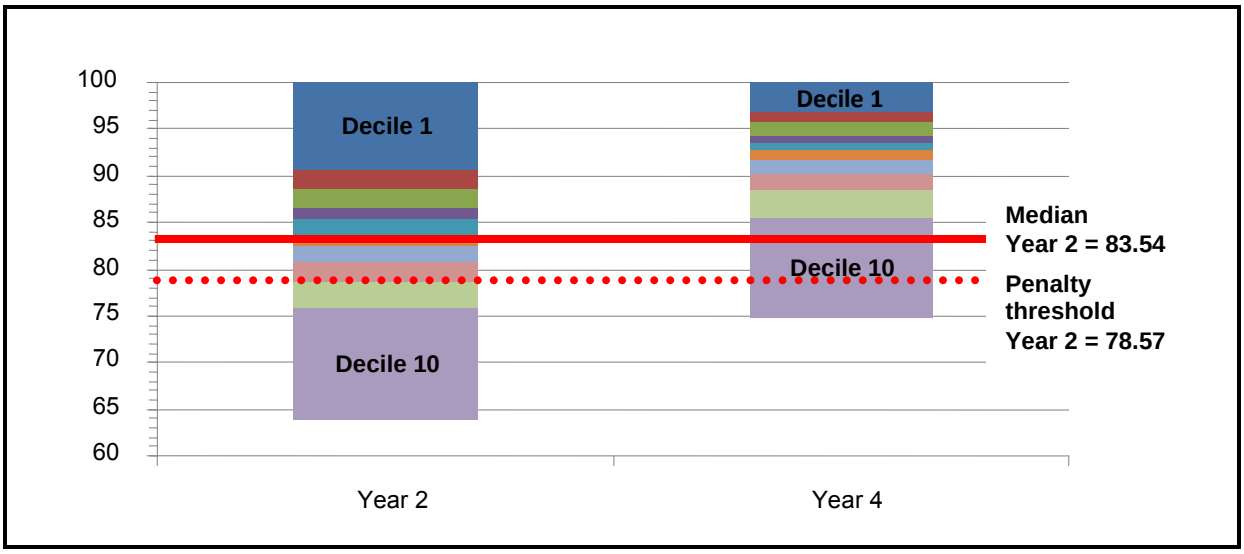
The median value in Year 2 was 93.91. As this solid red line shows, this value in Year 4 indicates that virtually all hospitals received the attainment award except those in the tenth decile. The dotted red line indicates that hospitals in Year 4 that had a HK CQS score of 87.13 or below paid a penalty (no hospitals for HK in Year 4). As previously noted, CMS set a cap of 85% starting in Year 4, so that penalties were not assessed for hospitals that had an HK CQS score of at least 85%.

4.2.1.5 Decile Ranges: PN

Figure 29 shows the ranges in PN decile thresholds over time. Like the other clinical areas, the thresholds have been compressed significantly between Year 2 and Year 4. In Year 2, the upper limit for decile 10 was 75.89, and by Year 4 had increased to 85.49. The lower limit for receiving a high performer award (decile 2) was 88.68 in Year 2 and increased only to 95.64 in Year 4. The difference in threshold between decile 2 and decile 3 for HK (the difference between receiving a high performer award and not receiving one) did not decrease substantially between Year 2 (1.86) and Year 4 (1.13).

The median value in Year 2 was 83.54. As the solid red line shows, this value in Year 4 indicates that virtually all hospitals received the attainment award except those in the bottom half of the tenth decile. The dotted red line indicates that hospitals in Year 4 that had a PN CQS score of 78.57 or below paid a penalty (which was virtually no hospitals for PN in except those in bottom of decile 10 in Year 4).

Figure 29
Demonstration: Decile ranges Year 2 and Year 4: PN [a]



[a] The figure depicts the reference deciles set based on the Year 2 CQS scores. Beginning in Year 4, penalties were not assessed for hospitals scoring at least 85%.

4.2.2 Average CQS across Deciles

This section shows the results from analyses that examined the mean value of each CQS in each decile during Years 1–4. These analyses were performed using the CQS scores that were provided to RTI by CMS, and were performed for all hospitals that were in each year of the Demonstration (not for the panel of hospitals that reported measures for all 16 quarters). As these analyses show, in general, the means within each decile have increased over the course of the Demonstration (similar to the findings described for the decile thresholds over time), and the difference in CQS measures between deciles has compressed over time. These results are described for each clinical focus group in the sections below.

4.2.2.1 Average CQS across Deciles: AMI

As **Table 41** shows, the mean AMI CQS for decile 5 in Year 3 was the same mean AMI CQS for decile 3 in Year 2. By Year 4, the mean AMI CQS for decile 9 (one of the bottom two deciles) was the same as the mean CQS for decile 3 in Year 1. Similar findings were noted for all of the clinical areas—by Year 4, the mean CQS for the bottom two deciles were similar to the mean CQS for the higher deciles.

It is also noteworthy that the difference in CQS between deciles has compressed significantly over time. For example, the difference between decile 2 and decile 3 for AMI in Year 1 was 0.014 points (0.950–0.936). By Year 4, the difference between decile 2 and decile 3 for AMI was half that (0.990–0.983, or 0.07 points). Similarly, the difference between the top and bottom deciles greatly decreased over time. In Year 1, it was 0.970–0.756, or 0.214 points. By Year 4, the difference between the top and bottom deciles was merely half that, 0.102 points.

Table 41
Mean CQS scores by decile: Demonstration Years 1–4, clinical focus area AMI

Decile	Mean CQS Year 1	Mean CQS Year 2	Mean CQS Year 3	Mean CQS Year 4
1	0.970	0.985	1.001	0.998
2	0.950	0.972	0.988	0.990
3	0.936	0.964	0.981	0.983
4	0.923	0.954	0.973	0.978
5	0.912	0.945	0.964	0.973
6	0.897	0.932	0.958	0.966
7	0.885	0.923	0.949	0.958
8	0.865	0.907	0.935	0.949
9	0.836	0.877	0.916	0.934
10	0.756	0.822	0.880	0.896

CQS = composite quality score; AMI = acute myocardial infarction.

4.2.2.2 Average CQS across Deciles: Coronary Artery Bypass Graft

As **Table 42** shows, the trends are the same for CABG CQS as were noted for AMI CQS. The mean CABG CQS for decile 10 in Year 4 was the same mean CABG CQS for decile 1 in Year 1. The difference in CQS between deciles compressed significantly over time for CABG. For example, the difference between decile 2 and decile 3 for CABG in Year 1 was 0.021 points. By Year 4, the difference between decile 2 and decile 3 for CABG was merely 0.004 points. Similarly, the difference between the top and bottom deciles greatly decreased over time. In Year 1, it was 0.977–0.736, or 0.241 points. By Year 4, the difference between the top and bottom deciles was only 0.070.

4.2.2.3 Average CQS across Deciles: Heart Failure

As **Table 43** shows, the trends are the same for HF CQS as were noted for both AMI and CABG CQS. The mean CQS for decile 10 in Year 4 did not reach the same mean CQS for the higher deciles in the earlier PHQID years, because the mean CQS for HF in decile 10, Year 1 was extremely low. However, the mean CQS for decile 9 in Year 4 was higher than that for decile 2 in Year 1.

The difference in CQS between deciles compressed significantly over time for HF. The difference between the top and bottom deciles also greatly decreased over time. In Year 1, it was 0.910–0.441, or 0.469 points. By Year 4, the difference between the top and bottom deciles was half that, only 0.214.

Table 42
Mean CQS Scores by decile: Demonstration Years 1–4, clinical focus area CABG

Decile	Mean CQS Year 1	Mean CQS Year 2	Mean CQS Year 3	Mean CQS Year 4
1	0.977	0.992	0.999	1.000
2	0.954	0.985	0.994	0.996
3	0.933	0.979	0.990	0.993
4	0.904	0.971	0.986	0.991
5	0.885	0.958	0.980	0.989
6	0.869	0.940	0.973	0.986
7	0.848	0.925	0.963	0.982
8	0.828	0.887	0.947	0.979
9	0.799	0.846	0.928	0.970
10	0.736	0.785	0.869	0.930

CQS = composite quality score; CABG = coronary artery bypass graft.

Table 43
Mean CQS Scores by decile: Demonstration Years 1–4, clinical focus area HF

Decile	Mean CQS Year 1	Mean CQS Year 2	Mean CQS Year 3	Mean CQS Year 4
1	0.910	0.969	0.986	0.991
2	0.840	0.920	0.959	0.975
3	0.805	0.890	0.932	0.954
4	0.768	0.864	0.914	0.936
5	0.722	0.837	0.895	0.922
6	0.675	0.806	0.874	0.910
7	0.642	0.768	0.848	0.895
8	0.603	0.726	0.825	0.878
9	0.554	0.665	0.779	0.857
10	0.441	0.557	0.695	0.777

CQS = composite quality score; HF = heart failure.

4.2.2.4 Average CQS across Deciles: Hip/Knee Replacement

As **Table 44** shows, the mean CQS across deciles has improved greatly for HK over time, just as for the other clinical focus areas. The mean CQS for decile 9 in Year 4 was almost that of the mean CQS for decile 2 in Year 1. Also, the difference in CQS between deciles compressed significantly over time for HK. The difference between decile 2 and decile 3 in Year 1 was 0.013 points, and by Year 4 it was merely 0.003 points. Also, the difference between the top and bottom deciles greatly decreased over time. In Year 1, it was 0.213 points. By Year 4, the difference between the top and bottom deciles was half that, only 0.108.

Table 44
Mean CQS scores by decile: Demonstration Years 1–4, clinical focus area HK

Decile	Mean CQS Year 1	Mean CQS Year 2	Mean CQS Year 3	Mean CQS Year 4
1	0.969	0.991	0.998	0.997
2	0.942	0.977	0.992	0.990
3	0.929	0.967	0.986	0.987
4	0.910	0.959	0.980	0.981
5	0.892	0.946	0.972	0.976
6	0.872	0.931	0.964	0.970
7	0.847	0.914	0.953	0.964
8	0.829	0.893	0.934	0.955
9	0.804	0.849	0.897	0.938
10	0.756	0.778	0.819	0.889

CQS = composite quality score; HK = hip/knee replacement.

4.2.2.5 Average CQS across Deciles: Pneumonia

As **Table 45** shows, the mean CQS across deciles improved greatly for PN over time, and the means for this clinical focus area were among the lowest for all clinical focus areas in Year 1 The mean CQS for decile 1 in Year 1 was 0.863, and by Year 4 it was 0.980. Also, the mean CQS for decile 9 in Year 4 was more than that of the mean CQS for decile 1 in Year 1.

The difference in CQS between deciles compressed significantly over time for PN. The difference between decile 2 and decile 3 in Year 1 was small, 0.028 points, and by Year 4 it was even smaller. Also, the difference between the top and bottom deciles greatly decreased over time. In Year 1, it was 0.267 points. By Year 4, the difference between the top and bottom deciles was almost half that, only 0.154.

Table 45
Mean CQS scores by decile: Demonstration Years 1–4, clinical focus area PN

Decile	Mean CQS Year 1	Mean CQS Year 2	Mean CQS Year 3	Mean CQS Year 4
1	0.863	0.923	0.967	0.980
2	0.820	0.896	0.941	0.963
3	0.792	0.877	0.927	0.949
4	0.771	0.860	0.915	0.939
5	0.754	0.846	0.898	0.933
6	0.732	0.830	0.889	0.923
7	0.714	0.816	0.878	0.910
8	0.682	0.800	0.860	0.896
9	0.647	0.775	0.836	0.874
10	0.596	0.719	0.794	0.826

CQS = composite quality score; PN = pneumonia.

4.3 Rewards for Performance: Year 4

In this section we describe the results of the Pay for Performance in Year 4. Recall that for Year 4 of the Demonstration, there were 3 types of rewards for hospitals:

- **Median-Level Attainment Award:** Hospitals that attain or exceed the median-level CQS in a given clinical area, as measured 2 years prior for each clinical area receive an incentive payment. For example, a hospital will achieve median-level attainment in Year 4 if that hospital's CQS in Year 4 exceeds the median CQS for all participating hospitals in Year 2.

- **Top Performer Award:** Hospitals that have a CQS for a given clinical area that falls in the top 20% of scores receive an additional payment. These hospitals receive the median level attainment award as well.

- **Top Improvement Award:** Hospitals that attain median-level performance and are among the top 20% of hospitals with the largest percentage quality improvement in a clinical area receive an additional incentive payment. Improvements are calculated based on the performance year compared to 2 years prior. This group receives the median-level attainment award as well. Top performers are not eligible for the improvement award.

We describe the number of hospitals that received rewards, by each type of incentive that was rewarded in Year 4, and we report the total payments paid to the hospitals by each type of incentive reward. Note that in Years 1 through 4, incentive payment amounts and award amounts were the same for all but a few cases. However, there are instances where these amounts diverge.

This occurs when a hospital is eligible for more than one award for a particular case. In those instances, the hospital receives one incentive payment for the condition with the higher DRG weight and, thus, the higher payment amount. This prevents the hospital from receiving more than one incentive payment for a particular case.

4.3.1 Attainment Awards

As **Table 46** shows, the percentage of hospitals in each clinical area that achieved median attainment was extremely high—86% for both AMI and HK, 93% for HF, 95% for CABG, and 96% for PN. As results in the previous section showed, the CQS over time has increased greatly across the years. Because the attainment award for Year 4 is based on achieving the median set from Year 2, it is not surprising that a large percentage of hospitals received median attainment rewards in Year 4.

Table 46
N and percentage of hospitals receiving any award (attainment, top improver, or top performer) in Year 4, by clinical focus group

N (row percent)	No reward	Any reward	Total N hospitals
AMI	29 (14.01)	178 (85.99)	207
CABG	6 (5.17)	110 (94.83)	116
HF	16 (7.14)	208 (92.86)	224
HK	29 (14.15)	176 (85.85)	205
PN	8 (3.54)	218 (96.46)	226

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

4.3.2 Top Improver Awards

Hospitals that attain median-level performance and are among the top 20% of hospitals with the largest percentage quality improvement in each clinical area receive an additional incentive payment. Improvement is calculated based on the performance year compared to 2 years prior. This group receives the median-level attainment award as well. Top performers are not eligible for the improvement award. **Table 47** shows the number and percentage that received top improver awards, of those that were eligible for the top improver award: approximately one quarter of hospitals across each clinical focus area received top performer awards.

Table 47
N and percentage of hospitals receiving top improver awards in Year 4, by clinical focus group [a]

N (row percent)	No Top improver award	Yes Top improver award	Total hospitals eligible for top improver award
AMI	98 (73.68)	35 (26.32)	133
CABG	60 (74.07)	21 (25.93)	81
HF	119 (74.38)	41 (25.63)	160
HK	96 (74.42)	33 (25.58)	129
PN	128 (74.85)	43 (25.15)	171

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

[a] Of those hospitals eligible for top improver award.

4.3.3 Top Performer Awards

For the top performer awards, hospitals that have a CQS for a given clinical area that falls in the top 20% of scores receive an additional payment. These hospitals receive the median-level attainment award as well. Because these hospitals represent the top two deciles of hospitals within a given clinical focus area, as **Table 48** shows, approximately 20% hospitals in each focus area received top performer awards.

Furthermore, only 4% of AMI hospitals and 8% of CABG hospitals received a top performer award in Year 4 for both clinical focus areas.

4.4 Incentive Payments in Year 4

Tables 49 to **53** show for each clinical area, payments received for each award type available in Year 4. These tables show the mean, minimum, and maximum incentive payment received, by each clinical area. These results are discussed in terms of the mean, minimum, and maximum incentive payments received for each award type:

Attainment. The mean incentive payment for attainment was highest for the HF clinical area ($7,256.59). The mean incentive payment for attainment was lowest for the CABG clinical area ($1,969.64). Across all clinical focus areas, the minimum incentive paid for attainment was $224.68 (CABG), and the maximum incentive paid for attainment was $36,497.91 (HF).

Table 48
N and percentage of hospitals receiving top performer awards in Year 4, by clinical focus group

N (row percent)	No Top performer award	Yes Top performer award	Total N hospitals
AMI	166 (80.19)	41 (19.81)	207
CABG	93 (80.17)	23 (19.83)	116
HF	179 (79.91)	45 (20.09)	224
HK	164 (80.00)	41 (20.00)	205
PN	181 (80.09)	45 (19.91)	226

AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia.

Top Improver. Similar to the results for the mean attainment incentive payment, the mean incentive payment for top improvers was highest among hospitals that received rewards for the HF clinical area ($22,941.44) and the mean attainment incentive payment was lowest for the CABG clinical area ($5,193.44). Across all clinical focus areas, the minimum incentive paid for top improvers was $1,625.77 (PN), and the maximum incentive paid to top improvers was also in the PN clinical focus area, $62,863.16.

Top Performer. The mean incentive payment for top performers was highest for the CABG clinical area ($25,977.00) and the mean incentive payment for top performers was lowest for the AMI clinical area ($13,843.29). Across all clinical focus areas, the minimum incentive paid for top performers was $1,716.09 (AMI), and the maximum incentive paid to top performers was $123,648.95 (CABG).

Table 49
Demonstration: Incentive Payments for Year 4, AMI

Payment	N	Mean payment	Minimum payment	Maximum payment
Payment attainment	108	$5,306.78	$2,023.54	$16,751.09
Payment top improver	4	$12,057.81	$9,664.31	$15,353.15
Payment top performer	41	$13,843.29	$1,716.09	$68,372.72

AMI = acute myocardial infarction.

Table 50
Demonstration: Incentive Payments for Year 4, CABG

Payment	N	Mean payment	Minimum payment	Maximum payment
Payment attainment	108	$1,969.64	$224.68	$12,906.58
Payment top improver	4	$5,193.44	$3,161.22	$8,490.14
Payment top performer	23	$25,977.00	$3,251.54	$123,648.95

CABG = coronary artery bypass graft.

Table 51
Demonstration: Incentive Payments for Year 4, HF

Payment	N	Mean payment	Minimum payment	Maximum payment
Payment attainment	208	$7,256.59	$748.93	$36,497.91
Payment top improver	41	$22,941.44	$2,890.26	$49,947.31
Payment top performer	45	$23,949.02	$4,786.99	$75,779.02

HF = heart failure.

Table 52
Demonstration: Incentive Payments for Year 4, HK

Payment	N	Mean payment	Minimum payment	Maximum payment
Payment attainment	176	$6,539.11	$574.18	$29,108.46
Payment top improver	33	$20,776.48	$5,509.56	$50,669.88
Payment top performer	41	$24,866.81	$2,077.37	$71,985.55

HK = hip/knee replacement.

Table 53
Demonstration: Incentive Payments for Year 4, PN

Payment	N	Mean payment	Minimum payment	Maximum payment
Payment attainment	218	$5,532.59	$449.36	$28,958.67
Payment top improver	43	$19,173.18	$1,625.77	$62,863.16
Payment top performer	45	$15,191.93	$3,251.54	$41,005.57

PN = pneumonia.

SECTION 5
TEST MEASURES AND OTHER MEASURES

In the previous sections, we described the trends in the Demonstration for Years 1–4 for those measures for which we had full 16 quarters of data. In this section, we describe trend results for other measures for which there were not 16 quarters of data available. In the first section, we describe the results of test measures for Year 4, which were tested in the Demonstration but not used to compute the clinical quality score for Year 4. In the second section, we describe results of various other measures for which we had some quarters of data (but not full 16 quarters), including some measures new to Year 4.

As was true for the previous chapter, this chapter presents analyses that focus on changes in quality that may not be due strictly to the Demonstration alone. This chapter provides the trends in quality scores over time for hospitals that participated in the Demonstration.

5.1 Test Measures

Table 54 shows the results for the test measures for which we received data on all four quarters of Year 4. These measures were not included in the calculation of the clinical quality score for Year 4.

Table 54
Demonstration: Test measures Year 4, means

Clinical focus group	Measure	N hospitals reporting all four quarters	Q13 mean	Q14 mean	Q15 mean	Q16 mean	% change Q13–Q16
AMI	ACS	206	0.891	0.899	0.914	0.921	3.4%
AMI	PSIC	206	0.987	0.985	1.005	0.999	1.2%
CABG	ACS	115	0.903	0.919	0.919	0.933	3.4%
CABG	SI	115	1.001	1.001	1.000	1.006	0.5%
CABG	PSIC	115	0.980	0.977	0.971	1.002	2.3%
HF	ACS	223	0.766	0.797	0.821	0.837	9.2%
HF	SI	222	0.995	0.993	0.995	0.994	−0.1%
HF	PSIC	223	1.030	1.034	1.008	1.010	−2.0%
HK	ACS	204	0.870	0.832	0.846	0.864	−0.6%
HK	PSIC	204	1.080	1.093	1.089	1.082	0.2%
PN	ACS	225	0.803	0.808	0.850	0.856	6.6%
PN	SI	225	0.999	0.998	0.998	0.998	−0.1%
PN	PSIC	225	1.057	1.096	1.048	1.037	−1.9%

ACS = appropriate care score; AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia; PSIC = patient safety indicator composite; SI = survival index.

The appropriate care score (ACS) designates when a patient has received all possible care measures for which he is eligible. It is conceptualized as the "all or nothing" measure, meaning that a patient must receive all of the care measures for which he is eligible for the hospital to receive credit for the measure. For all clinical focus areas except HK, this measure had a positive percentage change between Q13 and Q16 of the Demonstration. The largest percentage increase was noted for HF, which increased from 0.766 to 0.837 over four quarters (9.2% increase). The second largest percentage increase was noted for PN, which increased from 0.803 to 0.856 over the four quarters (6.6% increase). The ACS measure for the HK clinical focus group remained essentially unchanged during the four quarters.

The patient safety composite (PSIC) is an AHRQ measure that is a composite of many other measures:

- Decubitus ulcer
- Iatrogenic pneumothorax
- Selected infections due to medical care
- Postoperative hemorrhage or hematoma
- Postoperative hip fracture
- Postoperative physiological and metabolic derangement
- Postoperative pulmonary embolism (PE) or deep vein thrombosis (DVT)
- Postoperative respiratory failure
- Postoperative sepsis
- Postoperative dehiscence
- Accidental puncture or laceration
- Transfusion reaction

This PSIC measure was also a test measure for each of the five clinical focus groups in Year 4 and was extremely high at the beginning of Year 4 for all clinical focus areas (.98 or higher starting in Q13). For this measure modest increases occurred over the quarters for AMI (1.2%, from 0.9) and CABG (2.3%). There was a modest decrease in this measure for HF and PN (2% decrease between quarters for both clinical focus areas), but this is essentially because the measure could only decrease over time because it was at its maximum in Q13. The PSIC measure for HK was essentially unchanged during the quarters.

The survival index (SI) was a test measure for CABG, HF, and PN during Year 4. For each of these clinical focus areas the SI measure started high in Q13 and remained high and essentially unchanged over the four quarters of Year 4.

Table 55 shows the changes in the standard deviation of each test measure during the four quarters of Year 4. Most of the standard deviations decreased during the four quarters, indicating that there is decreased variability in the measure over time. This is not surprising, given that similar findings have been noted for all of the other Demonstration measures (see Section 3). For the PSIC measure, there is no common trend in standard deviation changes across the clinical focus areas (some changes are positive and others are negative, but this is likely because the standard deviation for PSIC). Similarly, for the SI measure, all clinical focus areas show a percentage change that is positive, indicating that hospitals were becoming more variable over time. However, we believe this result occurred because the standard deviation was very low at the start of Year 4 (0.101 for CABG, 0.116 for HF); hence, the measure has little alternative but to increase slightly over time.

Table 55
Demonstration: Test measures Year 4, standard deviations

Clinical focus group	Measure	Standard deviation Q13	Standard deviation Q13	Standard deviation Q13	Standard deviation Q13	% change Q13–Q16
AMI	ACS	0.671	0.646	0.652	0.605	−9.82%
AMI	PSIC	0.705	0.794	0.674	0.680	−3.56%
CABG	ACS	0.759	0.600	0.773	0.694	−8.53%
CABG	SI	0.101	0.179	0.097	0.135	34.09%
CABG	PSIC	0.456	0.536	0.683	0.550	20.50%
HF	ACS	1.550	1.392	1.254	1.145	−26.13%
HF	SI	0.116	0.121	0.095	0.121	4.41%
HF	PSIC	1.895	1.815	1.901	1.643	−13.29%
HK	ACS	1.198	1.314	1.230	1.032	−13.84%
HK	PSIC	1.287	1.430	1.007	1.163	−9.66%
PN	ACS	1.182	1.277	0.946	0.923	−21.89%
PN	SI	0.137	0.155	0.150	0.157	15.08%

ACS = appropriate care score; AMI = acute myocardial infarction; CABG = coronary artery bypass graft; HF = heart failure; HK = hip/knee replacement; PN = pneumonia; PSIC = patient safety indicator composite; SI = survival index.

5.2 Other Measures

In addition to the core Demonstration measures for which we had 16 quarters of data for most hospitals, and in addition to the test measures described above, we also performed analyses on a few measures for which we had not full 16 quarters of data but were not considered test measures. These measures include

- PN: Initial antibiotic selection for community-acquired pneumonia (CAP) in immunocompetent patient: Year 4 data only.

- PN: Influenza screening and/or vaccination: Fall/Winter Quarters (Q1 and Q2) in Year 1 and Fall/Winter Quarters (Q13 and Q14) in Year 4

- CABG: Prophylactic antibiotic selection for surgical patients: Year 1 (Q1–Q4) and Year 4 (Q13–Q16)

- HF: Prophylactic antibiotic selection for surgical patients: Year 1 (Q1–Q4) and Year 4 (Q13–Q16)

In the following tables, we describe the trends in the means and standard deviations for these measures over time for the various quarters for which there are data.

5.2.1 PN: Initial Antibiotic Selection for CAP in Immunocompetent Patient: Year 4 Data Only

Table 56 shows the means over the four quarters in Year 4 for initial antibiotic selection in immunocompetent patient. In Q13 the mean score was 0.906 and in Q16 the mean score was 0.914, a slight increase of less than 1%.

Table 56
Demonstration: Initial antibiotic selection for CAP in immunocompetent patient, Year 4 means

Measure	Q13	Q14	Q15	Q16	% change Q13–Q16
Initial antibiotic selection for CAP in immunocompetent patient (n=225)	0.906	0.914	0.919	0.914	0.88%

CAP = community-acquired pneumonia.

Table 57 shows the standard deviations over the four quarters in Year 4 for initial antibiotic selection in immunocompetent patient. In Q13 the mean standard deviation was 0.723 and in Q16 the mean score was 0.667, a decrease of over 7%. There is less variation in this measure in Q16 than in Q13 among hospitals.

Table 57
Demonstration: Initial antibiotic selection for CAP in immunocompetent patient, Year 4 standard deviations

Measure	Q13	Q14	Q15	Q16	% change Q13–Q16
Initial antibiotic selection for CAP in immunocompetent patient (n=225)	0.723	0.696	0.629	0.667	−7.75%

CAP = community-acquired pneumonia.

5.2.2 PN: Influenza Screening and/or Vaccination: Q1 and Q2 in Year 1; Q13 and Q14 in Year 2

Table 58 shows the means over four quarters (two in Year 1 and two in Year 4) for influenza screening and/or vaccination. This measure has increased significantly over the Demonstration period. In Q1 the measure was 0.398 and, by Q14, the measure was 0.865, an improvement of over 100% between Q1 and Q14. Also, the rates of improvement during Year 1 and Year 4 are vastly different. During the two quarters of Year 1, the rate improved from 0.398 to 0.471, or 18.34%. During the two quarters of Year 4, the rate improved from 0.844 to 0.865, or 2.52%. As noted for many of the other Demonstration measures, this measure improved most during the early Demonstration years.

Table 58
Demonstration: Influenza screening and/or vaccination. Q1 and Q2 in Year 1; Q13 and Q14 in Year 2, means

Measure	Q1	Q2	Q13	Q14	% change Q1–Q2	% change Q13–Q14	% change Q1–Q14
Influenza screening and/or vaccination (n=221)	0.398	0.471	0.844	0.865	18.34%	2.52%	117.4%

Table 59 shows the standard deviations over the four quarters in Years 1 and 4 the influenza screening measure. In Q1 the mean standard deviation was 2.890 and in Q14 the mean standard deviation was 1.450, a decrease of almost 50%. There is less variation in this measure in Q14 than in Q1 among hospitals.

Table 59
Demonstration: Influenza screening and/or vaccination, Q1 and Q2 in Year 1; Q13 and Q14 in Year 2, standard deviations

Measure	Q1	Q2	Q13	Q14	% change Q1–Q14
Influenza screening and/or vaccination (n=221)	2.890	3.305	1.4578	1.450	−49.83%

5.2.3 Prophylactic Antibiotic Selection for Surgical Patients: Year 1 and Year 4: CABG and HK

Table 60 shows the means over eight quarters (Q1–Q4 in Year 1 and Q13–Q16 in Year 4) for prophylactic antibiotic selection for surgical patients for both CABG and HF clinical focus areas. For both CABG and HK, this measure has increased significantly between Q1 and Q16 (although the percent increase was greater for CABG than for HK). Also, the rates of improvement during Year 1 and Year 4 were vastly different for the CABG measure. During the

four quarters of Year 1, the rate improved from 0.942 to 0.960, or 1.9%. During the four quarters of Year 4, the rate improved from only 0.61%. As noted for many of the other Demonstration measures, this measure for CABG improved most during the early Demonstration years and has leveled off during Year 4. The rate of change for the HF measure was not different between Year 1 and Year 4.

Table 60
Demonstration: Prophylactic antibiotic selection for surgical patients, Year 1 and Year 4: CABG and HK, means

Clinical focus group	Measure	Q1	Q2	Q3	Q4	Q13	Q14	Q15	Q16	% change Q1 to Q4	% change Q13 to Q16	% change Q1–Q16
CABG	Prophylactic antibiotic selection for surgical patients (n=109)	0.942	0.941	0.951	0.960	0.981	0.990	0.977	0.987	1.91%	0.61%	4.88%
HK	Prophylactic antibiotic selection for surgical patients (n=167)	0.975	0.963	0.973	0.971	0.990	0.991	0.991	0.993	−0.41%	0.30%	1.81%

CABG = coronary artery bypass graft; HK =hip and knee replacement.

Table 61 shows the standard deviations in the Prophylactic antibiotic selection for surgical patients for both CABG and HF during Year 1 and Year 4. Between the first and last quarter of the Demonstration, the standard deviation has declined dramatically between (by 64.2% for CABG and by 43.3% for HK). As we have noted for many other measures, variation in this measure has decreased greatly for hospitals in Demonstration Year 4.

Table 62 describes the trends and rates of change between Year 1 and Year 4 for the measure. The percentage improvement during Year 1 (between Q1 and Q4) for CABG was 2% (from 0.942 to 0.960), and dropped to 0.7% increase during Year 4 (between Q13 and Q16). The rate of improvement for this measure for CABG has decreased since Year 1. On the other hand, the rate of change was not different between Year 1 and Year 4 for the HK measure. The measure actually decreased slightly during Year 1 (from 0.975 in Q1 to 0.971 in Q2), and increased slightly during Year 4 (from 0.990 to 0.993). Nevertheless, these changes were small and not different for HK.

Table 61
Demonstration: Prophylactic antibiotic selection for surgical patients, Year 1 and Year 4: CABG and HK, trends and rates of changes in standard deviations

Clinical focus group	Measure	Q1	Q2	Q3	Q4	Q13	Q14	Q15	Q16	% change Q1–Q16
CABG	Prophylactic antibiotic selection for surgical patients (n=109)	1.075	1.118	1.062	0.968	0.530	0.261	0.849	0.385	−64.2%
HK	Prophylactic antibiotic selection for surgical patients (n=167)	0.381	0.725	0.439	0.504	0.389	0.240	0.224	0.216	−43.3%

CABG = coronary artery bypass graft; HK =hip and knee replacement.

Table 62
Demonstration: Prophylactic antibiotic selection for surgical patients, Year 1 and Year 4: CABG and HK, trends and rates of changes in means

Clinical focus group	Measure	Q1	Q4	Q13	Q16	% change Q1–Q4	% change Q13–Q16
CABG	Prophylactic antibiotic selection for surgical patients (n=109)	0.942	0.960	0.981	0.987	2.0%	0.7%
HK	Prophylactic antibiotic selection for surgical patients (n=167)	0.975	0.971	0.990	0.993	-0.4%	0.3%

CABG = coronary artery bypass graft; HK =hip and knee replacement.

SECTION 6
CONCLUSIONS AND NEXT STEPS

In Phase I, we present results describing the trends in process, outcome, and composite quality score measures for Years 1–4 of the Demonstration. These analyses focused on trends in quality for a panel of Demonstration hospitals that remained in the Demonstration for all quarters of Years 1–4. We examined trends in quality means between Q1 (the start of the Demonstration) and Q16 (the end of Year 4 of the Demonstration), as well as trends in rates of change per year, and trend in quality measure standard deviations. We also examined trends in hospital deciles over time.

6.1 Trends in Quality

Overall, our results show that quality measures showed significant increases between Q1 and Q16 of the Demonstration. Quality is improving across all clinical focus areas for process and outcome measures as well as for composite quality scores. However, most of the increase in quality was seen during the first 2 years of the Demonstration while Year 4, in general, showed that early gains were sustained. Note that this analysis includes only Demonstration hospitals and does not examine a control group; therefore, we cannot definitively say that the observed quality trend was caused by the Demonstration and differs from a national trend among all hospitals. Regardless, it is reasonable to state that quality improved among these hospitals and that the gains did not diminish in Year 4.

There were a few measures that did not show increases between Q1 and Q16; these were measures that were very high at the start of the Demonstration (including mostly the outcome measures such as Survival index and Post-op physical/ metabolic derangement avoidance index). For these measures that are at or near the ceiling during Year 1, there is little room for improvement, however, performance is susceptible to decreases and these did not show decreases in performance over time. Measures that start out with such high performance have several implications; first, the near perfect performance among some measures results in a composite measure (CQS) mean that is strongly influenced by this high measure and does not reflect a low performance in other measures that make up the CQS score. However, it also means that on these measures nearly every hospitals performs at a high level so is rewarding performance necessary for clinical quality care measures that hospitals are already doing?

To assess whether quality was increasing at a constant rate across the Demonstration years, we examined the rates of change in quality measures during the first 4 years of the Demonstration and whether these rates were significantly different over time. These results show that for the majority of process, outcome, and CQS measures, improvements happened at a relatively fast rate during the early years (Years 1 or 2) of the Demonstration, and started to level off by Year 4. We found very few significant measure increases between Q13 and Q16. However, the CQS for both heart failure (HF) and pneumonia (PN) increased significantly during all years, including during Year 4 (although the rate of change during Year 4 was significantly smaller than the rate of change during Year 1, indicating that the rate of improvement has slowed). Because the Demonstration continued for 2 additional years after Year 4, our Phase II analysis will focus on whether hospitals are able to sustain this high level of

performance and whether the leveling off in quality improvement continued for the remainder of the Demonstration.

Beyond examining trends in mean quality measures over time, we also examined the variability in performance among hospitals over time. By looking at the standard deviation of the measures over time, we saw that measure variation among hospitals is decreasing. Standard deviations decreased significantly between Year 1 and Year 2, indicating that hospital performance for a given measure are beginning to "cluster" with smaller range between the high and low performers; everyone is getting better. This is a positive finding, in that less variation in measures over time indicates that hospitals for the most part are improving and that the lower limit is rising. Phase II analyses will continue to examine the trends in measure standard deviations over the course of the Demonstration to assess whether Demonstration hospitals reach a minimum standard deviation such that further decrease in variation is unlikely.

6.2 Trends in Deciles and Pay for Performance

Section 4 of the report describes our analyses of pay-for-performance incentives and the movement of hospitals through deciles over time. A key analysis examined all hospitals that reported a CQS measure in Year 1 (and thus were assigned a decile in Year 1). We examined the likelihood that hospitals in the top two deciles for Year 1 would be in the top two deciles in Year 4. Similarly, we examined the likelihood that hospitals in the bottom two deciles for Year 1 would remain in the bottom two deciles in Year 4 (as well as the likelihood that they would have dropped out of the Demonstration by Year 4).

These results showed that between 28% (HK) and 39% (PN) of hospitals in the top two deciles in Year 1 remained in the top 20% in Year 4. Moreover, hospitals in the top 20% in Year 1 were *unlikely* to drop to the bottom 20% by Year 4 and were also unlikely to be missing a CQS score in Year 4. In other words, a sizable fraction of high performing hospitals remained in the Demonstration and remained high performers. We also saw that hospitals performing at the bottom 20% for Year 1 *remained* in the bottom 20% of all hospitals in Year 4, e.g., 23% (HK) and 40% (CABG). Although hospital performance improved across the board, few hospitals improved enough to move from the bottom 20% in Year 1 to the top 20% in Year 4.

A key finding from this analysis is that hospitals that were in the bottom 20% in Year 1 were more likely to be missing from the Demonstration in Year 4, as compared with hospitals in the top 20% in Year 1. Between 25% (PN) and 35% (AMI) of hospitals that were in the bottom two deciles in Year 1 were missing CQS scores for those clinical focus areas by Year 4. This finding has implications for the evaluation of the Demonstration since those experiencing low performance are more likely to drop out of participation and not place themselves at risk for penalties associated with poor performance. The finding clearly indicates that self-selection is occurring: hospitals that were high performers early in the Demonstration are highly likely to remain in the Demonstration throughout (and thus push the quality scores higher over time), whereas hospitals that were low performers are likely to drop out.

We also used the decile thresholds provided by CMS to display graphically how the deciles are changing over time. For all clinical focus areas examined, the decile thresholds have increased over the Demonstration years, indicating that quality is improving. Furthermore, the

decile thresholds have compressed over time, such that in Year 4 there is often very little difference in the threshold that would put a hospital one decile or another. In particular, we found that the difference in decile threshold between decile 2 and decile 3 (i.e., the difference in receiving a top performer award) decreased dramatically in Year 4 compared with Year 2. We will continue to examine this phenomenon in Phase II of the evaluation, to determine if the trend continues (i.e., differences between deciles get smaller and smaller).

An intriguing result from the decile analysis revealed that the Demonstration extension's expansion of payment opportunities to include the attainment award resulted in substantially more hospitals receiving rewards in Year 4 compared with the previous years. Hospitals in Year 4 must have reached the median set in Year 2 to receive the attainment award, and our results reveal that the median in Year 2 was often close to the ninth or tenth decile in Year 4 (because quality scores had increased dramatically over time). Therefore, for all clinical focus areas, a large proportion of hospitals (often more than 90%) received Attainment Award bonuses. Phase II of the Demonstration evaluation will examine whether this trend continues. Because our trend analyses show that quality scores are improving at a much slower rate in Year 4 compared with previous years, we may find that not as many hospitals are rewarded for median attainment in Year 6 if the rates of improvement continue to slow between Year 4 and Year 6.

6.3 Next Steps

In Phase II of the analyses, we will continue to examine trends in quality measures for the full 6 years of the Demonstration and to examine whether the rates of improvement are indeed slowing, as they appear to be based on analysis of Years 1–4. Key analyses to be performed in Phase II will include comparison hospitals that did not participate in the Demonstration to examine whether Demonstration participants improved their quality measures during the period of the 6 Demonstration years significantly more so than the comparison hospitals. Other key analyses to be performed include incorporating Medicare claims data to examine whether improvements in quality of care, as measures by the process, outcome, and CQS measures in the Demonstration, affect Medicare beneficiary acute care inpatient hospital length of stay and mortality for the conditions included in the Demonstration. Future reports will also discuss the reported reasons that hospitals withdrew from participating during the Demonstration, including environmental reasons (hospital closures/mergers/acquisitions), hospital resource issues related to withdrawals, as well as low performance during Demonstration participation.

REFERENCES

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., Report to CMS, December 1, 2008.

Lindenauer, P., Remus, D., Roman, S., Rothberg, M., Benjamin, E., Ma, A., and Bratzler, D. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med,* 1:356(5): 486-496. 2007.

Ryan, A. Effects of the Premier Hospital Quality Incentive Demonstration on Medicare patient mortality and cost. *Health Serv Res,* 44(3): 821-842. 2009.