Alternative Approaches to Measuring Physician Resource Use

Medicare/Medicaid Research and Demonstration
Task Order Contract (MRAD/TOC)
HHSM-500-2005-00027I, T.O. 4

Second Interim Report

December 2010

Submitted by

David Knutson, M.S.
Bryan Dowd, Ph.D.
Robert Kane, M.D.
Shriram Parashuram, M.P.H.
University of Minnesota

Robert F. Coulam, Ph.D.
Simmons College

CMS Project Officer: Craig Caplan, MA

Table of Contents

Executive Summary

This research project is designed to explore creative methods for measuring physician resource use that could be used to reward physicians in traditional fee-for-service (FFS) Medicare for efficient and high quality care. This project began in late September 2008. In the first year of work, the University of Minnesota (UMN) team focused on the fundamental issues of measuring and attributing the risk adjusted costs of care to physicians. In order to reward physicians for efficient and high quality care – one of the fundamental points of this entire effort – there first must be some way to attribute costs and quality to physicians who are in an ordinary sense "accountable" for those costs. Our attribution approach recognized there are two types of physicians – those who primarily provide care for hospitalized beneficiaries and those who primarily provide care in ambulatory settings. Our measures of attribution and subsequent use of resources for physicians providing inpatient care are based on the Medical Severity-Diagnosis Related Groups (MS-DRGs) and subsequent post-acute care, including readmissions. Our approach to ambulatory care (the Beneficiary analysis) holds a physician or physician group accountable for an entire year of care for a beneficiary. Since all care during the year is included, the physician will be held responsible for the cost of hospitalizations as well as the appropriateness of hospitalizations and emergency department utilization as part of the quality measures.

In the second year, the focus of this report, we continued work on attribution and cost profiling for both approaches and made substantial progress in developing and evaluating clinical performance ("quality") measures, as summarized below.

1. MS-DRG-Based Approach -- Our goal in the MS-DRG analysis is to create a prototype measurement approach for both cost and quality that encompasses risk adjusted cost of care for a selected list of conditions that are conceptually meaningful, acceptable to MDs, and easily interpreted. For the prototype measure development, we identified a set of 11 MS-DRGs in our first year of work on the basis of their high frequency and cost for the Medicare program (such as acute myocardial infarction (AMI) and congestive heart failure (CHF)). We used data from 10 States to build physician cost profiles for beneficiaries residing in the 10 States with at least one inpatient hospital admission for any of the 11 identified MS-DRGs in 2008 and part of 2009.

An episode starts with the hospitalization of a beneficiary for a particular MS-DRG and runs until 30- or 60- days after admission. An appropriate physician (attending or operating) is held accountable for all of the beneficiary's costs and outcomes within that 30- and 60- day episode window, except the cost of the initial hospitalization, which is not expected to be under that physician's control. The MS-DRG Technical Expert Panel (TEP, see below) agreed with this approach. In the second year, we identified two ways in which costs of 30- and 60- day MS-DRG episodes could be attributed to hospital physicians: 1) using NPI (National Provider

Identification number) and 2) using TINs (Tax Identification Number). We chose to use the latter approach as TINs address larger groups of physicians who share a tax ID and hence may collectively treat the same patient.

In our preliminary work with Colorado data, we used unstandardized payment amounts found on the Medicare claims. That approach was useful for initial exploratory purposes, but we believed could not be used in any final measure development. In the second year, we standardized costs for the both Part A and Part B claims. Using these standardization algorithms, we obtained the total standardized episode cost for each selected 30- and 60-day MS-DRG episode, by adding the standardized Part A and B costs for 30/60 days to the index standardized outlier payment. The *episodes* for a given MS-DRG were aggregated under each TIN to obtain the unadjusted standardized cost profile for that TIN.

The resulting amounts are unadjusted in that variations in the risk of *patients* have not been taken into account. For risk adjustment we adapted the Hierarchical Condition Category (HCC) model that is used for managed care payment under Medicare. We now have to complete the task of risk adjusting the physician profiles by running the regression at the physician level, with beneficiary-level risk adjustors, also adjusting for the number of episodes nested within each TIN.

2. Beneficiary-level Approach – There are four main strands of work in the project over the past year. The first is attribution. Originally, we had considered basing attribution in the Beneficiary analysis on PQRI reports, a conceptually attractive approach that would base attribution on an active, ex ante initiative by the physician (i.e., the PQRI report). However, PQRI reporting at the present level of participation (less than 10 percent of the physicians submitting an E&M claim) could not be the basis of an attribution system. Based on consultations with the Beneficiary TEP Panel and CMS, we moved to a "plurality of evaluation and management (E&M) visits" basis for attribution.

Our second strand of work following attribution has been to determine exactly what quality measures will be used to assess the care attributed to a physician. Usable sample sizes are maximized by claims-computable measures, which include "avoidables" like ambulatory care sensitive admissions and potentially preventable readmissions, as well as disease-specific quality measures endorsed by the National Quality Forum (NQF), such as:

1) Adult(s) with diabetes mellitus that had a serum creatinine in last 12 reported months.

2) Lower extremity amputations among patients with diabetes: This measure is used to assess the number of admissions for lower-extremity amputation among patients with diabetes.

3) Osteoporosis: Screening or therapy for women aged 65 years and older: “Percentage of female patients aged 65 years and older who have a central DXA measurement ordered or performed at least once since age 60 or pharmacologic therapy prescribed within 12 months.” [1]

4) Pneumonia vaccination status for older adults: “Percentage of patients 65 years of age and older who ever received a pneumococcal vaccination.”

5) Admissions for bacterial pneumonia: “This measure is used to assess the number of admissions for bacterial pneumonia.”

6) Chronic kidney disease - Lipid profile monitoring: "Percentage of patients with chronic kidney disease that have been screened for dyslipidemia with a lipid profile."

All of the measures on which we currently are working or have completed work were approved by the TEP panel. For all of the claims-computable measures, we have written programs to compute the measures at the TIN level and provided a descriptive analysis of the results. Some of the ways we are using these measures are novel – e.g., the use of a special algorithm developed elsewhere (Billings, 2000) to estimate the avoidable emergency department visits for patients of particular physicians. These admissions are extremely costly and, because possibly avoidable, may indicate poor quality care. Using that algorithm to analyze 2008 Colorado data, we found an average of 0.43 ED visits per beneficiary per (TIN) in 2008. The average probability of emergent ED admissions that were *not* potentially preventable/avoidable (and thus not likely to indicate poor care) was 0.59. The average probability for emergent but potentially preventable/avoidable admissions was 0.07. The average probability for ED admissions treatable through primary care, rather than the ED, was 0.15. The average probability for non-emergent care was 0.15.

Such results showed the practicality of using these claims-computable measures as performance measures. We also explored a number of special issues that arise in applying these measures: e.g., the small denominator problem that arises from a large proportion of cases with zero or few events.

The third strand of work in the Beneficiary analysis is the risk adjusting of costs to providers. The costs attributed to providers must be risk adjusted, for obvious reasons. But there are many ways this can be done. We have focused on resolving a series of technical choices in risk adjustment of costs (e.g., whether to risk adjust costs using residuals in a regression with cost or quality variables also on the right side; or to use a random effects model).

[1] The osteoporosis screening measure is “time-limited endorsed” by NQF.

The fourth task of the Beneficiary analysis in the second year was to develop methods to combine the risk adjusted cost measures and risk adjusted quality measures into a summary measure of physician performance. We have spent considerable time in the last year exploring options. Our preferred approach is to calculate frontier-based efficiency scores for the quality measures treated as *individual* measures. We can plot physician performance in two or more dimensions, with cost as the input and either a single or multiple summary measure of quality. When all physicians are thus plotted, their relative efficiency is determined by their spatial distance from a "frontier" defined by the physicians who produce the highest quality for the lowest cost. The Beneficiary TEP panel seemed favorably disposed toward this approach to combining cost and quality, and we have continued work to explore it.

3. <u>Technical Expert Panels to evaluate methodology and rate quality measures</u> – We developed an array of candidate quality measures for physicians, then looked to panels of experts to winnow those lists by rating the usefulness of each individual measure (which would also assist us in any attempts to develop composite measures of quality). For both the MS-DRG and Beneficiary approaches, the individual quality measures were a mix of broad constructs (e.g., emergency department visits and avoidable emergency department visits) and more specific measures (e.g., patient safety indicators for the MS-DRG approach and specific Patient Quality Reporting Initiative (PQRI) measures for the Beneficiary approach).

We assembled a panel of experts for each of the two approaches and met with them over two days in Minneapolis to review the status of the project and the rating exercise. We then had the panels perform the rating exercise. The ratings process did a good job of winnowing our candidate quality measures, the primary purpose for which it was used.

- In the MS-DRG ratings of the usefulness of measures, there was a mixture of results with regard to the seven overall constructs. Emergency Department visits received the lowest average score (26.7) and potentially preventable admissions received the highest (62.3). Only three overall constructs received average ratings of 50 or more: avoidable ED visits (53.3), potentially preventable hospital readmissions (62.3), and AHRQ Patient Safety Indicators (50.0). The ratings for individual items were generally higher than those for the general constructs: 33 of the 42 measures were rated ≥ 50, while 16 measures were rated ≥ 70, including aspirin at discharge for acute myocardial infarction patients (81.5), foreign body left during procedure (77.5), and accidental puncture or laceration (74.8).

- In the Beneficiary results on usefulness of measures, six of the nine general constructs received an average rating of 70 or more: avoidable emergency department visits, proper sequence of care measures, measures of overuse, PQRI measures, ambulatory care sensitive admissions, and avoidable readmissions. The exceptions were emergency department visits, fraud related measures, and all-cause mortality. In the ratings of the 180 individual PQRI measures, 47 received a score of 70 or better. The 47 measures tended to be:

— Prevention and screening measures, such as five highest rated measures: advance care plan (95.7), mammography screening (92.1), advising cessation of tobacco use (91.4), and pneumonia vaccination for patient ≥ 65 (90.7).

— Basic primary care measures, for common conditions such as diabetes (diabetes mellitus–foot exam, 92.1), osteoporosis (screening or therapy for osteoporosis for women aged 65 years and older, 84.2), heart disease (beta-blocker therapy for left ventricular systolic dysfunction (LVSD), 78.6), and hypertension (high blood pressure control for diabetic, 82.9).

The ratings of quality measures for the primary care Beneficiary approach were higher than those for the MS-DRG approach – a reasonable difference, given that the latter measures addressed overall hospital performance where the chain of causation is more tenuous than that for primary care providers. As anticipated, a number of specific measures were not judged useful for working with physicians.

Introduction

"Alternative Approaches to Measuring Physician Resource Use" is a research project, designed to explore how to create measures that could be used to reward physicians in traditional fee-for-service (FFS) Medicare for efficient and high quality care. Health care purchasers long have been aware of variation in the amount and the quality of care delivered in the United States that appear to be related to variations in provider efficiency rather than patient need.

Under value-based purchasing practices, physicians face payments that reflect not only the determined price of services but also the volume, intensity, and the quality of services. In the *Medicare Improvements for Patients and Providers Act (MIPPA) of 2008* Congress directed the Centers for Medicare and Medicaid Services (CMS) to: "develop a plan to transition to a value-based purchasing program for Medicare payment for physician and other professional services." In the *Affordable Care Act of 2010*, Congress went further and directed CMS to "establish a payment modifier that provides for differential payment to a physician or a group of physicians … based upon the quality of care furnished compared to cost." As a result, CMS is seeking to understand and develop mechanisms to assess the variation in resource use among practitioners, and the relationship of resource use to quality.

This project began in late September 2008 to explore creative methods for measuring physician resource use. In the first year of work, as summarized in our First Interim Report,[2] the University of Minnesota (UMN) team focused on the fundamental issues of measuring and attributing the risk adjusted costs of care to physicians. This Second Interim Report describes our progress through the second year of work, when we continued work on attribution and cost profiling and made substantial progress in developing and evaluating clinical performance measures. We developed several approaches to selecting, creating, and evaluating evidence-based quality (and some overuse) performance measures and – most important – worked with panels of national experts formally to assess the relative importance of such measures of performance. This Second Interim Report summarizes that second year of work as follows:

I. Progress at the End of the First Year (First Interim Report)

II. MS-DRG-Based Approach

III. Beneficiary-level Approach

IV. Clinical Performance Measures and the Process of Consulting Expert Panels

[2] David Knutson, Bryan Dowd, and Robert Kane, *Alternative Approaches to Measuring Physician Resource Use: Interim Report*, dated September 29, 2009, submitted to CMS pursuant to Medicare/Medicaid Research and Demonstration Task Order Contract HHSM-500-2005-000271, Task Order 0004

I. Progress at the End of the First Year (First Interim Report)

The activities addressed in this project are premised on the idea that physicians can be held accountable for the costs of care received by patients attributed to them and the quality associated with that care in an open access fee-for-service system. There is little point in developing elaborate performance measures of efficiency and quality if the care received (or not received) by a beneficiary cannot be attributed to a particular health care provider or set of providers. In a health insurance program like FFS Medicare, such attribution is difficult. Thus, the first problem that must be addressed is the attribution of beneficiaries to providers, and that problem was the focus of much of our initial work.

Our approach to the attribution challenge recognized that for our purposes there are two types of physicians – those who primarily provide care for hospitalized beneficiaries while in the hospital and those who primarily provide ambulatory care.

- MS-DRG approach – Our measures of attribution and subsequent use of resources for physicians providing inpatient care are based on the Medical Severity-Diagnosis Related Groups (MS-DRGs) and subsequent post-acute care, including readmissions. The MS-DRG analysis is conducted conditional on the patient having been admitted.[3] Attribution of inpatient medical care to the attending physician is relatively straightforward, but attribution of inpatient surgical care to the operating versus performing physician requires analyses that we currently are performing. We also addressed the key issues concerning which costs to attribute in a post discharge period and how long those costs should be recognized in the post-discharge period.

- Beneficiary approach – Our approach to physicians providing care primarily in ambulatory settings holds a physician accountable for an entire year of care for a beneficiary. Since all care during the year is included, the beneficiary level analysis incorporates measures of resource use and quality that involve the use of both inpatient services as well as ambulatory services (e.g., PQRI-based measures of appropriate outpatient care and measures of appropriate use of emergency departments and hospital admissions). Attribution in typically unstructured ambulatory care settings is difficult. We began our efforts to explore, for the first time, the usefulness of data from CMS's Physician Quality Reporting Initiative (PQRI) for attribution purposes.

[3] For admitted patients, we include all Medicare Part A and Part B costs within a 30- and 60- day window from the time of admission, except we exclude: 1) the MS-DRG Part A payment (although we retain the operational outlier payment amount associated with the initial MS-DRG excluding the capital component); 2) the operational disproportionate share and operational indirect medical education amounts from all hospital readmissions, to make hospital readmission costs independent of hospital type; and 3) costs associated with trauma readmissions, for which the physician cannot reasonably be held accountable.

In addition to our work on the attribution problem, we had begun in the first year to develop measures of physician resource use in the MS-DRG analysis. All measures of resource use must be risk adjusted, and the measures of resource use and quality must be combined to produce metrics that can be used to reward efficient, high-quality physician practices in a way that is viewed as fair. By the time of the First Interim Report, we had initial results from analyses of a resource use measure and a method for risk adjustment of resource use in the MS-DRG analysis.

Both the MS-DRG and Beneficiary analyses were in their initial stages at the time of the First Interim Report. The report focused to a large extent on descriptive data and anomalies, inconsistencies, and other features of the data encountered at that point in the analyses. This past year, we refined our work on attribution and cost, and continued development and expert review and critique of clinical performance ("quality") measures. We begin by presenting the work on the MS-DRG approach.

II. MS-DRG-Based Approach

A. Data and approach

Our goal in the MS-DRG analysis is to create a prototype measurement approach for both cost and quality that encompasses risk adjusted cost of care for a selected list of conditions that are conceptually meaningful, acceptable to MDs, and easily interpreted. For the prototype measure development, we identified a set of 11 MS-DRGs in our first year of work on the basis of their high frequency and cost for the Medicare program: acute myocardial infarction (AMI), congestive heart failure (CHF), chronic obstructive pulmonary disease, stroke, hip fracture, hip replacement, knee replacement, cholecystectomy, pneumonia/bronchitis, colon cancer and lung cancer.

To develop the measures, the MS-DRG approach uses data from 10 States chosen to represent variation in location, size and practice patterns – California, Colorado, Florida, Kansas, Louisiana, Minnesota, New Jersey, Virginia, Vermont, and Washington. This approach builds physician cost profiles using Medicare claims for 2008 and part of 2009 for beneficiaries residing in the 10 States with at least one inpatient hospital admission for any of the 11 identified MS-DRGs. Medicare claims from 2007 for these beneficiaries are used for the purpose of risk adjustment. To be included in the sample, beneficiaries were required to have continuous Part A and Part B enrollment and no Medicare Advantage enrollment during the entire period. Our preliminary work on the MS-DRG approach was done with the Colorado state data.

After examining the distribution of Medicare costs after hospitalization for an MS-DRG, we identified two time windows for which physician cost profiles would be created: 30 days and 60 days after admission. An episode starts with the hospitalization of a beneficiary for a particular MS-DRG and runs until 30- or 60- days after admission. An appropriate physician (attending or operating) is held accountable for all of the beneficiary's costs and outcomes within that 30- and 60- day episode window, except the cost of the initial hospitalization, which is not expected to be under that physician's control. The MS-DRG Technical Expert Panel (TEP, see Section IV) agreed with this approach.

B. Attribution issues and moving from NPIs to TINs

We identified two ways in which costs of 30- and 60- day MS-DRG episodes could be attributed to hospital physicians: 1) using NPI (National Provider Identification number) and 2) using TINs (Tax Identification Number). We chose to use the latter approach as TINs address larger groups of physicians who share a tax ID and hence may collectively treat the same patient.

Physician TINs are found only on Part B claims. To obtain TINs for NPIs on Part A claims, we created a NPI-TIN crosswalk from Part B claims and merged them with the Part A claims by NPI number. In our earlier work we had shown that there was poor agreement between the Part A attending and Part B operating NPIs. Attributing MS-DRGs to NPIs resulted in a low volume of episodes per NPI (20-40% of NPIs had only one attributable episode for an MS-DRG). We sought to attribute MS-DRGs to TINs both to improve the volume of episodes per physician unit, and closely to approximate the team of providers that cares for a patient. We also examined if aggregating to TINs improved the previously reported disagreement between Part A and Part B physicians for surgical MS-DRGs. Moving from NPIs to TINs increased the volume of episodes per physician unit (10%-25% of TINs had one attributable episode for an MS-DRG). However, the median number of episodes attributable to each TIN remained low (the median was 2 episodes per TIN for most MS-DRGs). There was improvement in matching between the Part A and Part B physicians for surgical MS-DRGs, supporting the case for TIN attribution. However, given that some disagreement persisted between Part A and Part B operating physicians – even after moving to TINs – we reached a decision with CMS to attribute costs of surgical MS-DRGs to using two alternative approaches that attribute care and costs two different physician-entities Part A and Part B TINs.

C. Moving from unstandardized to standardized costs

In our preliminary work with Colorado data we used unstandardized payment amounts found on the Medicare claims. That approach was useful for initial exploratory purposes, but could not be used in any final measure development, for three reasons: First, Medicare adjusts payment to providers to reflect differences in local input prices. We standardize costs to allow the fair comparison of treatment costs of individual physicians to those of their peers practicing in different geographic locations. Standardization removes geographic differences in reimbursement. Second, standardization also allows for proper risk adjustment of costs as risk adjustment models based on non-standardized costs are biased- physicians in hospitals/regions with relatively lower Medicare reimbursement rates would appear to be more efficient than their peers in hospitals/regions with higher rates. Lastly, in our preliminary work with the Colorado data we observed MS-DRG episodes with extremely low total episode costs that were often the result of some error in the payment process (e.g., errors in charges). Approximately 5 percent of claims had such erroneous charges. Physicians with such erroneous low-cost episodes might erroneously appear to be efficient. These low-cost episodes also skew the cost data and risk adjustment. Using standardized costs removes problems arising from erroneous dollar amounts on claims.

We standardize the following sources of variation in Medicare payment:

1. Geographic variation: e.g. hospitals or physicians in two different geographic locations.

2. Provider-specific variation: e.g. hospitals within the same geographic location that receive different payments for disproportionate share hospital (DSH) or indirect graduate medical education (IGME) for the same MS-DRG.

3. Payment errors: e.g., variation arising due to erroneous dollar amounts on claims.

Adapting previous work by Mathematica Policy Research (Mathematica Policy Research Inc., 2008) and MedPAC (MedPAC, 2006), we standardize costs for the both Part A and Part B claims. The approaches used to standardize the costs for specific services/providers on Part A and Part B claims are summarized below. The detail of this discussion not only will help to clarify how we are standardizing costs. It will also provide a more complete picture of precisely what cost elements we are attributing to physicians – and for which we are thus holding them accountable.

1) Part A Claims

a. Inpatient Hospital: We standardized costs to remove both geographic variation as well as hospital specific variation in inpatient hospital payment. There were two payment amounts that required standardization on inpatient hospital claims:

- *Operational outlier payment amount on the initial hospitalization claim:* For the initial hospitalization, we wished to hold physicians accountable only for the portion of the outlier payment that could be attributed to the care that the beneficiary received, since the initial admission per se was not in the physician's control. The operational outlier payment amount is attributed to the physician while the capital outlier payment amount is not, as the latter is more or less a function of the hospital where the care was provided. We standardized the operation outlier payment amount on the initial hospitalization claim by factoring out geographic differences in the wage index and COLA.

- *Total payment amount on subsequent hospital readmissions:* Hospital readmissions (excluding those for trauma) within 30- and 60- days after admission for an MS-DRG are assumed to be related to the care that the beneficiary received from the hospital physician during initial hospitalization. We standardize the total payment amount on hospital readmissions, for a given MS-DRG, by using the average payment amount for that MS-DRG across our sample of 2008 and 2009 claims as the standardized payment amount.

b. Skilled Nursing Facilities: To standardize SNF costs, we first calculated the average daily rate for each RUG using our entire beneficiary sample, and then multiplied the mean daily rate for each RUG by the length of stay on the claim.

c. Home Health Services: To standardize home health costs we calculated the average home health resource group payment (HHRG) for each HHRG over our entire beneficiary sample and then assigned the average payment as the standardized cost to home health claims, after matching on HHRG.

d. Hospital Outpatient Services: We obtained standardized costs for hospital outpatient services claims covered by the outpatient prospective payment system from their ambulatory payment classifications (APCs).

For HCPCS codes not covered by the outpatient prospective payment system, we calculated the average payment for each HCPCS code over the entire beneficiary sample and used this amount as the standardized payment for the HCPCS code appearing on the claim.

e. Hospice Care, Inpatient Rehabilitation Facilities, and Inpatient Psychiatric Facilities: To standardize cost of each of these Part A services, we computed appropriate averages based on claims over the entire sample of beneficiaries.

2) Part B Claims

a. Physician Services: To standardize costs for physician services we merged the 2008 Relative Value Unit (RVU) file to the carrier file claims data, matching on HCPCS and modifier codes. For each line item HCPCS, we calculated standardized allowed charge by removing the influence of Geographic Practice Cost Indices (GPCI) from allowed charges. In the standardization process we accounted for whether the care was provided in a facility or non-facility setting. All components of RVUs (work, practice expense and malpractice) were standardized.

b. Anesthesiology Services: Allowed charges for anesthesiology services were standardized by dividing the allowed charge appearing on a claim by the geographic conversion factor.

c. Professional Services Not Paid on an RVU Basis, Ambulance Services, Clinical Lab services, Part B Drugs and Durable Medical Equipment: These services were standardized by calculating their mean allowed charge for each HCPCS over the entire sample of beneficiaries (an approach used by Mathematica policy research).

d. Ambulatory Surgical Center Services: We used national APC (ambulatory payment classification) payment rates to standardize each ASC claim.

D. Creating physician (TIN) profiles using standardized costs

For each selected MS-DRG episode we obtained the total standardized episode cost, for 30- and 60-days by adding the standardized Part A and B costs for 30/60 days to the index standardized outlier payment.

The *episodes* for a given MS-DRG were aggregated under each TIN, using attribution rules discussed earlier, to obtain the unadjusted standardized cost profile for that TIN. The resulting amounts are unadjusted in that variations in the risk of patients have not been taken into account. However, using standardized costs, we were able to eliminate aberrant low-cost episodes due to erroneously low amounts on claims, replace them with higher standardized cost episodes, and the distribution of episodes was less skewed.

E. Risk adjusting physician cost profiles

We risk adjusted episode costs attributed to physicians to account for differences in beneficiary demographics and beneficiary health status at the time of hospitalization. For risk adjustment we adapted the Hierarchical Condition Category (HCC) model that is used for managed care payment under Medicare. We considered all diagnoses occurring 12 months prior to a given index IP admission for beneficiaries to count beneficiaries HCCs. Apart from the 70 HCC diagnostic categories, the independent variables used in our risk adjustment model were beneficiary demographic variables such as age, sex, Medicaid status, reason for Medicare eligibility, interaction terms of demographic variables, MS-DRG Status (Complication/Comorbidity or Major Complication/Comorbidity[4]), and variables which identified whether a beneficiary had prior hospitalizations or ER use for the same condition a year prior to the admission. We also eliminated costs due to rehospitalizations for trauma from the episode costs by identifying if the beneficiary had HCCs for trauma after the initial

[4] MS-DRGs on the basis of their severity are usually classified at the time of discharge as – those without complications/comorbidities (no CC/MCC), those with complications/comorbidities (CC), those with major complications/comorbidities (MCC).

hospitalization. These HCC flags were included as variables in the right hand side of the risk adjustment model. The HCCs that we assumed to be exogenous were:

1. HCC 67 Quadriplegia, Other Extensive Paralysis
2. HCC 68 Paraplegia
3. HCC 69 Spinal Cord Disorders / Injuries
4. HCC 75 Coma, Brain Compression / Anoxic Damage
5. HCC 154 Severe Head Injury
6. HCC 155 Major Head Injury
7. HCC 157 Vertebral Fractures w/o Spinal Cord Injury
8. HCC 158 Hip Fracture/Dislocation
9. HCC 161 Traumatic Amputation
10. HCC 164 Major Complications of Medical Care & Trauma
11. HCC 177 Amputation Status, Lower Limb/Amputation Complications

To estimate the model, we regressed episode costs on the 70 pre-episode HCCs, the post-episode trauma HCCs and beneficiary demographic characteristics.

III. Beneficiary-level Approach

A. Data and approach

The Beneficiary analysis makes use of two samples of claims: claims for all TINs and claims for a much smaller sample of physicians reporting PQRI measures. The first sample includes all 2008 and 2009 claims of beneficiaries using physicians in fifteen States: Arizona, Colorado, California, Florida, Kansas, Louisiana, Massachusetts, Minnesota, New Jersey, New York, North Dakota, Texas, Utah, Virginia, Vermont, and Washington. (These are the same ten states as for the MS-DRG analysis, plus five additional States to allow more detailed investigation of appropriate peer grouping in the later stages of our project.) The second sample is a national sample, including all 2008 and 2009 claims of beneficiaries using physicians who reported two or more PQRI measures. The latter sample will be used for the analyses of PQRI-based quality measures.

Originally, we had considered basing attribution in the Beneficiary analysis on PQRI reports. For reasons described below, we moved to a "plurality of evaluation and management (E&M) visits" basis for attribution. All cost and quality analyses are reported at the TIN level. The analyses to date involve the claims-computable quality measures.[5]

[5] PQRI reports by physicians usually are piggybacked on claims forms, but they are not "claims computable," as the

There are four main strands of work in the project over the past year.

1. To which physician(s) can the cost and quality of care reasonably be assigned (i.e., the question of attribution)?
2. How should the cost data be adjusted for variables that are exogenous to the physician?
3. What are the most appropriate measures of health care quality?
4. How should cost and quality measures be combined to provide a summary assessment of physician performance?

We discuss each of these strands of work below, along with certain important issues they raise.

B. Attributing beneficiaries to TINs

In order to reward physicians for efficient and high quality care – one of the fundamental points of this entire effort – there must be some way to attribute costs to physicians that are in an ordinary sense "accountable" for those costs. Our early work explored whether attribution could be *ex ante* and "active" – that is, a function of the physician taking some initiative that could be read as responsibility for the patient, through the mechanism of the Medicare Physician Quality Reporting Initiative (PQRI), a quality-reporting initiative that rewarded physicians for reporting on care in terms of a set of consensus-based measures. We knew in advance that PQRI reporting rates were low, but the purpose of the analysis was to compare PQRI-based attribution based on the "plurality of E&M visits." Our analyses of Colorado data showed that physicians who provided PQRI reports on their patients were in the minority: for example, of physicians submitting Medicare claims for non-hospital E&M services, only six percent filed a PQRI report on at least one beneficiary. Following the Beneficiary TEP panel meeting (described below in Section IV.C) and subsequent discussions with CMS, we determined that PQRI reporting at the present level of participation could not be the basis of an attribution system, and we turned our attention to the TINs with the plurality of a beneficiary's E&M visits.

C. Quality measures

Our second strand of work following attribution has been to determine exactly what quality measures will be used to assess the care attributed to a physician. Usable sample sizes are maximized by claims-computable measures as opposed to PQRI-reported measures. Those measures include avoidable emergency department visits, ambulatory care sensitive admissions

PQRI measure cannot be derived from other information on the form, in those cases (the large majority) in which physicians have not provided a PQRI report.

and potentially preventable readmissions, in addition to disease-specific quality measures endorsed by the National Quality Forum (NQF).

Section IVA below describes how we developed candidate quality measures for the Beneficiary analysis and then evaluated those measures through a panel of experts. All of the measures on which we currently are working or have completed work were approved by the TEP panel. For all of the claims-computable measures, we have written programs to compute the measures at the TIN level and provided a descriptive analysis of the results (e.g., computed means, analyzed distributions of different levels of quality, and so on).

There are two groups of measures – "avoidables" and other claims-computable measures:

1) The "avoidables" – indicators of potentially poor quality

We completed analyses of two of the three avoidables (ED visits and ACS admissions) based on the Colorado data and under the PQRI attribution system. Given the change in attribution rules described above, we reran all of these analyses and completed analyses of the third measure (potentially avoidable readmissions). The measures and our results are discussed below.

a. Avoidable emergency department (ED) admissions – These admissions are extremely costly and, because they possibly are avoidable, may be indicative of poor quality care. Using an algorithm developed by Billings, et al. (2000), we calculated two ED measures: (1) the total number of ED visits per beneficiary assigned to a provider; and (2) a measure of the average appropriateness (avoidability) of the ED visits by beneficiaries assigned to the provider. For each diagnosis code associated with an ED admission, the Billings algorithm assigns probabilities that the ED visit (a) truly was emergent (needed) and was not preventable/avoidable; (b) was emergent but preventable/avoidable; (c) was treatable through primary care, rather than the ED; or (d) was non-emergent. Each possibility (a through d) receives a diagnosis-specific probability.

   To obtain TIN-level scores we summed the diagnosis-specific probabilities for each category of ED admissions (a through d, above) across ED visits for all beneficiaries assigned to the TIN. Then we divided those sums by the total number of ED visits of beneficiaries attributed to the TIN. The result is a set of scores that represents the average probability score for each category of ED admission (a through d) for beneficiaries assigned to that TIN. When there were multiple diagnoses, we used the diagnosis that gave the greatest benefit of doubt (e.g., the highest probability that the ED admission was appropriate) to the TIN.

   To our knowledge, this is the first time the Billings algorithm has been applied at the individual provider (TIN) level. In our analysis of 2008 Colorado data, there was an average of 0.43 ED visits per beneficiary per TIN in 2008. The average probability score for emergent ED admissions that were not preventable/avoidable was 0.59. This means that among beneficiaries with ED admissions, the Billings algorithm assigned an average probability of 0.59 across the beneficiaries assigned to each TIN. The average probability score for emergent but preventable/avoidable admissions was 0.07. The average probability for primary care treatable admissions was 0.15 and the average probability for non-emergent care was 0.15. These scores can be used to rank the appropriateness of ED usage by TINs. Because the Billings algorithm is diagnosis specific, there is no need to risk adjust for patient casemix. Risk adjustment for sociodemographic variables is a possibility, but that is a complex problem as discussed in section 3, below.

b. Ambulatory care sensitive (ACS) hospitalizations –We are using the ACS admissions criteria identified by the Agency for Health Research and Quality (AHRQ; see e.g., Bindman, et al., 1994). The ACS algorithm is somewhat easier to apply than the ED algorithm – the algorithm simply flags the ACS diagnosis codes. There are five immunization-preventable conditions (e.g., polio and rheumatic fever), nine chronic conditions (e.g., asthma and diabetes with ketoacidosis or hypersmolar coma), seven acute conditions (e.g., bacterial pneumonia and gastroenteritis), and seven "untyped" conditions (e.g., angina and tuberculosis).

The results for the ACS admission analysis are shown in Table 1, below. Among the 261,802 beneficiaries represented in the 2008 Colorado data, 12,999 or five percent had an ACS admission. However, those beneficiaries with an ACS admission average 1.4 ACS admissions per beneficiary (18,228 / 12,999) so the TIN-level average number of ACS admissions per attributable beneficiary was slightly higher than five percent (0.07). ACS admissions also can be used rank TINs, although the vast majority of TINs had no ACS admissions, so the measure will be most useful in distinguishing TINs with more serious problems managing admissions. The large percentage of TINs with zero ACS admissions is due, in part, to TINs with a small number of attributable beneficiaries, as discussed below.

Table 1: Analysis of ACS admissions: 2008 Colorado data

Total attributable beneficiaries: 261,802

Total beneficiaries with ACS admission: 12,999

Total ACS admissions: 18,228

Total TINs: 7,755

[Total ACS admissions/total attributable beneficiaries]/TIN ± SEM: 0.07 (0.003)

c. Potentially avoidable readmissions – We are using the 3M algorithm (Goldfield, et al, 2008) to analyze "potentially preventable readmissions" (PPRs). Readmissions are evaluated to determine whether they are clinically related to a previous admission and could have been prevented by: (1) better care during the initial hospitalization; (2) better discharge planning; (3) better post-discharge follow-up; or (4) better coordination of inpatient and outpatient care. We have applied the algorithm to beneficiaries who had at least one hospitalization per TIN. The PPR measure is computed at 15, 30 and 60 days. Among the 7,755 TINs in the 2008 Colorado data, 46 percent had beneficiaries with at least one hospitalization. Among the TINs with hospitalized beneficiaries, 48 percent had a 15 day PPR, 53 percent had a 30 day PPR, and 58 percent had a 60 day PPR. This measure also can be used to rank TINs, and more TINs will have a computable score for PPRs than for ACS admissions.

2) Additional claims-computable quality measures

We have programmed a number of claims-computable quality measures based on measures endorsed by the National Quality Forum (NQF), a nonprofit organization whose membership includes a wide variety of public and private healthcare stakeholders that work under NQF leadership to reach consensus on the measures. The claims-computable measures appropriate to our work include:

1) Adult(s) with diabetes mellitus that had a serum creatinine in last 12 reported months.

2) Lower extremity amputations among patients with diabetes: This measure is used to assess the number of admissions for lower-extremity amputation among patients with diabetes.

3) Osteoporosis: Screening or therapy for women aged 65 years and older: "Percentage of female patients aged 65 years and older who have a central DXA measurement ordered or performed at least once since age 60 or pharmacologic therapy prescribed within 12 months." [6]

4) Pneumonia vaccination status for older adults: "Percentage of patients 65 years of age and older who ever received a pneumococcal vaccination."

5) Admissions for bacterial pneumonia: "This measure is used to assess the number of admissions for bacterial pneumonia."

[6] The osteoporosis screening measure is "time-limited endorsed" by NQF.

6) Chronic kidney disease - Lipid profile monitoring: "Percentage of patients with chronic kidney disease that have been screened for dyslipidemia with a lipid profile."

For these additional measures, we have completed preliminary descriptive analyses of the proportion of attributed beneficiaries to each TIN who experience these events. Many TINs have only a small number of at-risk beneficiaries which results in some of the proportions heavily distributed at 0, 0.5 and 1.0, as discussed below.

3) Risk adjusting quality measures

The rationale and consequences of risk adjusting cost data are reasonably well understood. However, the same is not true of risk adjusted quality data. The Affordable Care Act legislation requires risk adjustment of quality measures and notes correctly that failure to adjust quality data for beneficiary characteristics that might be associated with poorer quality outcomes could make physicians reluctant to take patients with those characteristics. However, if the quality measures are risk adjusted for characteristics of vulnerable populations that are associated with lower quality scores, then the risk adjustment process can result in an implicit acceptance of that lower level of quality for the vulnerable population. We have reviewed this issue at length and reached the decision: (1) to risk adjust for the HCC demographic factors available in CMS administrative data plus education and income from Census data, and (2) to determine how much difference that adjustment makes.

4) The problem of small denominators

In a number of our analyses we have observed a large number of zero events and a large proportion of both zeros and ones in proportion measures. That problem is illustrated in the analysis of potentially preventable readmissions, above. Both low and high levels could be due simply to the fact that very few beneficiaries have been assigned to the TIN, or few beneficiaries are at risk for the measure (e.g., few hospitalized beneficiaries leading to PPR proportions that frequently are zero or one). This problem is somewhat different from the usual small sample problem that results in wide confidence intervals for measures. If a TIN has only one beneficiary in the denominator of a binary measure (e.g., ACS admissions or PPRs) then a proportion measure can take on only the values of zero and one. If there are only two beneficiaries in the denominator, the measure can take on only the values 0, 0.5 or 1.0. If an event is rare, then the TIN might have to have a relatively large number of beneficiaries assigned to it in order for the expected value of the measure to be greater than one. (If the expected number of events is less than one, then we cannot distinguish zeros due to efficient management from zeros due to small sample sizes.) We have worked on ways to adjust for the expected number of events that will allow us to distinguish low and high proportions that represent high or low quality from results due to small denominators (the methods for doing this involve using the denominator sample size as a control variable). Those results are not ready for presentation at this time.

5) <u>Quality measures for small (PQRI-reporting) sample</u>

All of the discussion of quality measures above refers to the large-sample analysis, based on the most-frequently-seen attribution rule. As noted, however, there are certain attractive properties to a PQRI-based attribution rule. While the main analysis will not use that attribution rule, we reached a decision to do a separate analysis using a PQRI-based attribution rule – specifically, of TINs that reported at least two PQRI measures. We reached this decision in part to permit us to explore ways in which this different rule leads to different patterns of care or results. In addition, the analysis of quality measures for the PQRI sample also gives us a much more diverse set of quality measures – i.e., in addition to the avoidable and certain other claims-computable measures, we can use the extensive PQRI measures themselves.[7] We have completed some preliminary work on PQRI reporting as part of our earlier analysis of PQRI reporting as a potential attribution method.

D. Risk Adjusted Costs

The costs attributed to providers must be risk adjusted, for obvious reasons. But there are many ways this can be done. To begin with, there are two ways to do risk adjustment:

1) The approach we have used since early in the project, in which we look at residuals in a regression with cost or quality variables also on the right side; or

2) The approach used in certain other CMS projects, which, for example, use a random effects model to estimate the ratio of expected over predicted performance, times the mean of the hospital sample, to give a score for each hospital. (Krumholz, et al., undated).

We used the first method because we believe that it is most appropriate for our purposes.

A second cost issue is to determine what period will be used to calculate the risk adjustments: should we use concurrent (same year) risk adjustment (e.g. as is being done in the CMS PGP Demonstration Project) or prospective (prior year) adjustment (e.g. as in Medicare Advantage payment)? The TEP Panel recommended that we use the concurrent condition categories (CCs) to risk adjust the beneficiary-level data. The TEP panel's reasoning was that it would be both fair and advisable to hold the attributable physician responsible for the ambulatory care sensitive (ACS) CCs in the concurrent model. In other words, if a physician's patient is hospitalized for an ACS condition, the physician's risk adjusted costs should rise. We agreed. Thus, we have determined first to regress *inpatient* costs on the risk adjustment variables *minus* the ACS CCs, and then to regress all *remaining costs* (e.g., outpatient costs) on the risk adjustment variables *including* the ACS CCs. Since the second regression controls for the ACS

[7] We anticipate that small denominators will be a continuing problem in the condition-specific PQRI measures.

CCs physicians are not penalized for seeing those patients in outpatient settings.   These error terms can be considered separately or combined to obtain total risk adjusted costs for the physician.

A third issue is the problem of negative predictions (predicted costs below zero) that can result from the models we use. The approach we decided to use for now is to run simple ordinary least squares (OLS) models and see if we have a serious problem with negative predictions. The issue is not yet resolved.

E. Approaches to combining cost and quality

The fourth task in our analytic work is to combine the risk adjusted cost measures and risk adjusted quality measures into a summary measure of physician performance. This task is awaiting completion of the risk adjusted cost and quality data. We have spent considerable time in the last year exploring these options for combining the cost and quality data:

1. Assign weights to determine a single aggregate quality score – One method is to assign weights to the quality measures based on the TEP panel results (which in the Beneficiary analysis imply roughly equal weights) to derive a single aggregate quality score. The aggregate quality score can be compared to risk adjusted costs to derive a combined cost-quality performance measure. The ratio of the quality score to risk adjusted costs or efficiency measures based on stochastic or DEA frontier analysis are possible combination metrics.

2. Compute the proportion of costs attributable to avoidables – A simpler, less comprehensive combination would focus on the potentially avoidables – care that we have a strong reason to suspect is of poor quality because it involves potentially avoidable ED visits, avoidable hospitalizations and re-hospitalizations – and to compute the proportion of risk adjusted costs that were attributable to that care.

3. Frontier analysis – The final method is to calculate frontier-based efficiency scores for the quality measures treated as individual measures. Frontier analysis assumes that physicians use Medicare dollars (beneficiary and taxpayer dollars) to produce quality. We thus can plot physician performance in two or more dimensions, with cost as the input and either a single or multiple summary measure of quality. When all physicians are thus plotted, their relative efficiency is determined by their spatial distance from a "frontier" defined by the physicians who produce the highest quality for the lowest cost. The Beneficiary TEP panel seemed favorably disposed toward this approach to combining cost and quality, and we have continued work to explore it.

We have spent considerable time working through a number of conceptual issues. For example, some of the quality measures such as avoidable ED visits, avoidable hospitalizations, and rehospitalizations necessarily imply significantly higher cost (or lower cost if avoided). That implies that, when combining cost and quality measures into one metric, it would be necessary to distinguish those types of quality measures from those that do not have such dramatic deterministic relationships to costs. This work is not yet complete.

IV. Clinical Performance Measures and the Process of Consulting Technical Expert Panels

As suggested by the discussion to this point, the MS-DRG and Beneficiary analyses drew extensively upon the advice of panels of technical experts to evaluate the methodology of the project and, more important, to rate the usefulness of different quality measures the project was considering. It will be useful to begin by discussing why these ratings were so important.

Identifying quality measures that can be used as part of various incentive payment programs represents a major challenge. One faces several issues:

1. <u>Is the measure **fair** to physicians?</u> Can the physician <u>*reasonably*</u> be held accountable (even if not <u>*exclusively*</u> accountable) for this measure, in the sense that it legitimately reflects his/her care?

2. <u>Is the measure **actionable** by the individual physician?</u> How available are actions to improve the physician's performance on this measure? How easy is it to recognize how to take action to implement changes in clinical practice to improve performance on this measure?

3. <u>Is the measure **meaningful** to physicians?</u> Will it make sense to those who are affected by it? Are the differences it detects likely to be viewed as reasonable and important?

Few measures can achieve all of these criteria or achieve them equally well. That argues for the use of collections of measures to characterize performance. In any event, a composite must be computed when the measures in aggregate cover a number of different key clinical care domains (e.g. chronic illness, prevention, diagnosis).

Once we decide to combine individual measures, we must decide how to aggregate them, specifically what weights to give to each measure. Weighting all measures equally does not mitigate this dilemma; it simply assigns a value of "1", which may not reflect each measure's relative importance. Basically, two approaches are available to determine relative weighting: a statistical approach, such as principal components or factor analysis, or an organized judgment, based on the opinions of an expert panel regarding the relative importance of such measures in determining central issue or construct. The second approach, although more subjective than the first, can address directly the three criteria above if the expert panels are instructed to do so, and thus the results are likely to be more easily understood and more respected by clinicians. It also can provide legitimacy to the relative weightings, as it rests weightings on the professional judgment of peers to be governed by the weights. Given these advantages, we chose the second approach. Once we have ratings of the relative usefulness of each measure, we can test a variety of different methods for combining the measures – e.g., as discussed below, we can use the ratings to establish a cutoff value, below which the rating of the measure is effectively zero; or we can develop analytic methods to combine the measures.

To that end, we convened two technical expert panels (TEPs) – one for the Beneficiary analysis and one for the MS-DRG analysis. We describe the process of working with the expert panels, after first briefly reviewing how we chose the quality measures the panels would consider.

A. The selection of candidate quality measures

The selection of attributable quality measures was done in two stages and built heavily on work done by others. First, we reviewed the state of the art in quality measurement, as documented in the literature, in standards or measurement efforts of government agencies and professional associations, and in other projects akin to the current project (e.g., CMS' Professional Group Practice Demonstration). We developed sets of candidate measures from these sources, for both the MS-DRG approach and the Beneficiary approach. We did analyses of the frequency of events after we chose the candidate measures. Second, as discussed in greater detail below, we used the organized judgments of the TEP panel to refine and then formally to rate the relative usefulness of the measures.

The measures are organized into two groups: broad category or general construct measures, and then longer lists of more specific, subcategory measures of one or more of the broad categories:

- <u>MS-DRG measures</u> – There were seven **overall construct** measures in MS-DRG analysis, including:
  - five physician-level quality measures:
    - emergency department (ED) visits,
    - avoidable ED visits,
    - all cause readmissions (i.e., readmission for any reason),
    - potentially preventable hospital readmissions, and
    - all-cause case-mix adjusted mortality measures.

- two hospital-level measure sets:
  - provider-level patient safety indicators (PSIs) reported by the Agency for Healthcare Research and Quality (AHRQ)[8], and
  - Medicare Hospital Compare reported by CMS.[9]

A second group of **specific** measures was comprised of subsets of the CMS and AHRQ hospital-level sets: e.g., 15 specific component measures of the AHRQ patient safety indicators (such as death in low mortality DRGs, pressure ulcers, foreign body left during procedure, and postoperative sepsis), as well as components of each of five groups of Medicare Hospital Compare measures (for acute myocardial infarction, congestive heart failure, pneumonia, and surgical care improvement).

- Beneficiary measures – Nine **broad category** measures were selected for the Beneficiary analysis:
  - CMS' Physician Quality Reporting Initiative (PQRI) measures,
  - Emergency department visits,
  - Avoidable emergency department visits,
  - Ambulatory care sensitive admissions,
  - Avoidable readmissions,
  - Measures of overuse,
  - Proper sequence of care,
  - Fraud-related measures, and
  - All cause case-mix adjusted mortality.

The second group of **specific measures** was comprised of the 180 PQRI measures for 2010.

Some of these measures (e.g., avoidable ED use, avoidable admissions, and readmissions) represent applications of measures that, while in use for other purposes, have not to our knowledge been used for physician performance measurement in public programs. Certain hospital measures (e.g. from Medicare Hospital Compare) currently are used for hospital performance measurement, but have not been used to measure the performance of physicians who practice in these hospitals.

[8] The PSIs are a set of indicators providing information on potential in-hospital complications and adverse events following surgeries, procedures, and childbirth (AHRQ 2006). The PSIs were developed by AHRQ after a comprehensive literature review, analysis of ICD-9-CM codes, review by a clinician panel, implementation of risk adjustment, and empirical analyses.

[9] Hospital Compare is a consumer-oriented website that provides information on how well hospitals provide recommended care to their patients (CMS 2010). The performance rates for this website generally reflect care provided to all U.S. adults with the exception of the 30-Day Risk Adjusted Death and Readmission measures that only include Medicare beneficiaries hospitalized for heart attack, heart failure and pneumonia. This website was created through the efforts of CMS, along with the Hospital Quality Alliance (HQA). The HQA is a public-private collaboration established to promote reporting on hospital quality of care.

B. Creation of the Technical Expert Panels

To put together a panel of experts appropriate to this exercise in weighting quality measures, we sought clinical experts with expertise in relevant clinical specialties, familiarity with issues associated with performance measurement of physician practices, and an ability to characterize the attitudes of practicing physicians. To find physicians with this combination of attributes, we primarily considered physicians who were in management roles in physician practices across relevant specialties and were also involved with the activities of state or national professional associations in developing quality assessment and performance measurement tools and standards. We added one HMO medical director because of managed care's long history in producing physician performance reports. The final composition of the panels was as follows:

The MS-DRG Technical Expert Panel

1. Anne W. Alexandrov, PhD, RN, FAAN – Professor, University of Alabama, Birmingham; Comprehensive Stroke Center National Quality Forum; Consensus Standards Approval Committee and Chair, Stroke Committee.

2. Gail Amundson, MD – Quality Quest; National Quality Forum (NQF) Consensus Standards Approval Committee; American College of Physicians Subcommittee on Performance Measurement; NQF Composite Measures Steering Committee.

3. Christopher Bever, MD – American Academy of Neurology; American Neurological Association; Associate Chief of Staff for Research and Development at the VA Medical Center Baltimore.

4. Robert Bonow, MD – American Heart Association; American College of Cardiology; Chief of the Division of Cardiology at Northwestern Memorial Hospital.

5. David Jevsevar, MD, MBA – American Academy of Orthopedic Surgeons.

6. Martha Radford, MD – Chief Quality Officer, New York University, Langone Medical Center.

7. Joanna Sikkema, MSN, ARNP – American Nurses Association (Cardiac); Preventive Cardiovascular Nurses Association Board of Directors; Director of the Acute Care Nurse Practitioner Program - University of Miami.

The Beneficiary Technical Expert Panel

1. Douglas Carr, MD – Medical Director, Education & System Initiatives, Billings Clinic.

2. Charlie Fazio, MD – Senior Vice President and Chief Medical Officer, Medica Health Plans.

3. Alan Glaseroff, MD – Chief Medical Officer, Humboldt Del Norte IPA.

4. Peter Hollmann, MD – American Geriatrics Society; Associate Chief Medical Officer of BCBS Rhode Island; Member, AMA PI Committee.

5. Cheryl L. Phillips, MD – Geriatrician; President of the American Geriatric Society; Chief Medical Officer, OnLok Lifeways.

6. Michael van Duren, MD – Chief Medical Officer, Sutter Connect.

C. Process of Working with the TEPs

After the membership of the TEPs was established, we proceeded through a two-stage consultation process for each panel:

- TEP Panel Meetings – First, members of each panel met with the project team and project officers from CMS in Minneapolis, April 26 and April 27, 2010. Each meeting lasted one day. The format of the two meetings was similar:

    - Each meeting was moderated by University of Minnesota project leaders, but substantively led by senior researchers from the University: Robert Kane, MD, for the MS-DRG Panel, and Bryan Dowd, PhD, for the Beneficiary Panel, following a project overview by David Knutson, MS. Two project officials from CMS participated in both meetings.

    - The purpose of these meetings was to introduce TEP members to the details of each of the two respective approaches of the project, to discuss with them the quality measures the project was considering, and to reach some understanding of the rating process that was to follow. The work of the panels, and the structure of the discussion, was set by a series of slides laying out the quality measures the project was considering and a series of issues associated with using them. After preliminaries (overview of the project, etc.), the general format of the meeting was that the UMN senior researcher presented an issue, described some of the associated complexities, and asked a set of focused questions, which then became the basis of discussion.

- Rating process – After the TEP meetings, the UMN project team digested the discussions and worked with CMS to refine the list of quality measures and draft instructions and a form for the panel members to use in the rating process. We conducted cognitive interviews and spoke with several potential raters by phone to be sure that the form and the instructions were clear and interpreted as we intended. Each TEP reviewed the area it had worked on at the TEP meeting, but we invited all TEP members, including those who could not attend the April meetings, to participate.

D. Methods

The rating forms asked panel members to rate the usefulness of potential quality measures for each of the MS-DRGMS-DRG and Beneficiary approaches. Each TEP worked on its own measure set. We developed separate forms for each separate forms for each group, but used the same basic format. The task consisted of rating the usefulness of both general measures and specific measures, using a scale of 0-100, where 100 represented the greatest relationship to the aspect of quality being addressed and zero meant that a particular measure should not receive any consideration. In response to one panel's interest, we also asked the raters to rate each item's suitability as a stimulus for system change. We did not ask the raters to make judgments about the frequency with which measures would occur or the logistics of collecting such measures. We considered those to be empirical questions to be addressed in other parts of the project. After cognitive testing and final revisions, the rating forms were mailed to the panel members.

E. Results

In this section, we summarize the actual ratings by the two TEP panels, both with respect to general constructs and specific component measures.[10] In both the MS-DRG and Beneficiary results, we found the ratings of items for their "usefulness" were extremely informative, but the ratings of items for their suitability as stimuli for system change were difficult to interpret and not very informative.

In the MS-DRG ratings of the usefulness of measures, there was a mixture of results with regard to the seven overall constructs. Emergency Department visits received the lowest average score (26.7) and potentially preventable admissions received the highest (62.3). Only three overall constructs received average ratings of 50 or more: avoidable ED visits (53.3), potentially preventable hospital readmissions (62.3), and AHRQ Patient Safety Indicators (50.0). There was considerable variation in the ratings across raters. The ratings for individual items were generally higher than those for the general constructs: 33 of the 42 measures were rated ≥ 50, while 16 measures were rated ≥ 70, including aspirin at discharge for acute myocardial infarction patients (81.5), foreign body left during procedure (77.5), and accidental puncture or laceration (74.8).

In the Beneficiary results on usefulness of measures, the average ratings for the general constructs are much higher than for the MS-DRG ratings, and the variance is considerably

[10] See Section IV-A above ("The selection of candidate quality measures") for an explanation of the general and specific measures.

smaller. Six of the nine general constructs received an average rating of 70 or more: avoidable emergency department visits, proper sequence of care measures, measures of overuse, PQRI measures, ambulatory care sensitive admissions, and avoidable readmissions. The exceptions were emergency department visits, fraud related measures, and all-cause mortality. In the ratings of the 180 individual PQRI measures, 47 received a score of 70 or better. The 47 measures tended to be:

— Prevention and screening measures, such as five highest rated measures: advance care plan (95.7), mammography screening (92.1), advising cessation of tobacco use (91.4), and pneumonia vaccination for patient ≥ 65 (90.7).

— Basic primary care measures for common conditions such as diabetes (diabetes mellitus–foot exam, 92.1), osteoporosis (screening or therapy for osteoporosis for women aged 65 years and older, 84.2), heart disease (beta-blocker therapy for left ventricular systolic dysfunction (LVSD), 78.6), and hypertension (high blood pressure control for diabetic, 82.9).

Measures related to inpatient care received appropriately lower scores.

F. Discussion

Overall, the ratings of quality measures for the primary care Beneficiary approach were higher than those for the MS-DRG approach. This difference seems reasonable given that the latter measures addressed overall hospital performance and may involve the actions of people and organizations after discharge. While one can hold the initial attending physician accountable, the chain of causation is more tenuous than that for primary care providers. However, because the measures are intended to be used as *relative* performance indicators, even the more tenuous accountability can work as quality indicators for the purposes of this project.

As anticipated, a number of PQRI measures were not judged useful for working with primary care providers. One option would be to retain all measures – given that each has been designed and approved for use by some authoritative national vetting process – and allow the low rating to produce a low weight in the composite. Alternatively, we can choose a cut point in the data, which eliminates low-weighted measures from our quality measures, but preserves more highly rated measures for analysis. We chose the second strategy, in part because low-weighted measures would have little bearing on the combined performance measures in any event, but would complicate understanding of the full measure set. We thus have recommended dropping those measures rated less than 70. Seventy was a natural cut-point in the data that still left 47 measures for analysis.

The TEP process did a good job of winnowing our candidate quality measures, the primary purpose for which it was used. We now need to incorporate the results of this rating process into the analytic work of the project, which will be one focus of work in coming months.

References

1. Agency for Healthcare Research and Quality (2006). Patient Safety Indicators Overview. AHRQ Quality Indicators, February 2006, available at http://www.qualityindicators.ahrq.gov/psi_overview.htm, accessed October 19, 2010.

2. Billings, J., N. Parikh, and T. Mijanovich (2000). Emergency room use: the New York story. Issue Brief: Commonwealth Fund, Number 434 (November 2000) 1–12.

3. Bindman, A. B., K. Grumbach, D. Osmond, M. Komaromy, K. Vranizan, N. Lurie, J.Billings, and A. Stewart (1994). Preventable Hospitalizations and Access to Health Care. *Journal of the American Medical Association*, 275:4 (1994) 305-311.

4. Centers for Medicare and Medicaid Services (2010). Hospital Compare. October 13, 2010, available at https://www.cms.gov/HospitalQualityInits/11_HospitalCompare.asp#TopOfPage, accessed October 19, 2010.

5. Dowd, B. E., C. Li, D. Knutson, R. Kane, S. Parashuram, M. Karmarker, C. Caplan, and J. Levy (2010). The Physician Quality Reporting Initiative (PQRI): Quality Measurement and Beneficiary Attribution. Division of Health Policy and Management, University of Minnesota, prepared pursuant to Medicare/Medicaid Research and Demonstration Task Order Contract (MRAD/TOC) HHSM-500-2005-00027I, T.O. 4.

6. Goldfield, N. I., E. C. McCullough, J. S. Hughes, A. M. Tang, B. Eastman, L. K. Rawlins, and R.F. Averill (2007). Identifying Potentially Preventable Readmissions. *Health Care Financing Review*, 30:1 (June 2007) 75-91.

7. Krumholz, H. M., S. Normand, D. H. Galusha, J. A. Mattera, A. S. Rich, Yongfei Wang, and Yun Wang (undated). Risk-Adjustment Models for AMI and HF 30-Day Mortality Methodology. Paper submitted by Yale University to the Centers for Medicare & Medicaid Services, under Subcontract # 8908-03-02, available at http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1163010421830.

8. Mathematica Policy Research Inc. (2008). Standardized Unit Costs for Use in Resource Use Reports (RURs). Report Submitted to Centers for Medicare & Medicaid Services Washington, D.C (July 2008).

9. Medicare Payment Advisory Commission (2006). Report to the Congress: Increasing the Value of Medicare. Washington, D.C. (June 2006) 3-26.