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

ALTERNATIVE APPROACHES TO MEASURING PHYSICIAN RESOURCE USE

FINAL REPORT

April 9, 2012

Submitted by

Bryan Dowd, PhD
Robert Kane, MD
Shriram Parashuram, MPH
Tami Swenson, MS
University of Minnesota

Robert F. Coulam, PhD
Simmons College

CMS Project Officer: Jesse Levy, PhD

TABLE OF CONTENTS

Figures

Tables

Executive Summary

Introduction

The Centers for Medicare and Medicaid Services (CMS) is working to design a payment system that can reward physicians in traditional fee-for-service (FFS) Medicare for efficient and high quality care. "Alternative Approaches to Measuring Physician Resource Use" is a research project designed to develop creative ways to measure cost and quality in physician services, to support CMS in this ambitious undertaking. Health care purchasers long have been aware of variation in the amount and quality of care delivered in the United States that appear to be related to variations in provider efficiency rather than patient need. This project is part of a growing body of work designed to improve the measurement of value in health care services, and thus make it possible to reward providers who offer greater value.

Under value-based purchasing practices, physicians face payments that reflect not only the price of services but also the volume, intensity, and 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" and in September 2008 CMS contracted with the University of Minnesota (UMN) team to begin work in that area.

However, 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.

In order to be successful, such a system must be considered fair, actionable and meaningful by both physicians and the Medicare program. The end product of this project is a set of recommendations designed to help CMS design a value based payment system for physicians serving Medicare patients. Our proposed approach is comprehensive, conceptually simple, transparent, reliable, and flexible. We have developed cost and quality of care measures that can be used to summarize physician performance and determine whether physicians will receive an incentive reward based on the quality and cost of services they provide to Medicare beneficiaries in the traditional fee-for-service Medicare program. We have proposed ways of combining cost and quality into a single metric of performance and examined the consistency and stability of that performance measure over time.

We are aware that several respected analysts do not believe that systems to measure physician cost and quality are "ready for prime time." These doubts are based on legitimate concerns about the sensitivity and reliability of attribution algorithms and cost and quality measures, and the problem of small sample sizes. We share these concerns, but our proposal is based on several countervailing observations including the fact that CMS is mandated by the 2010 ACA legislation to implement a value based payment system. Failure to implement a value based payment system is not a viable policy alternative.

The Essentials of Our System

We began our analysis with a focus on the broad policy objective – the development of reliable profiles of physician performance. We then developed a set of methods to achieve that goal. We did not begin our analysis with a predetermined format or "tool" such as episode groupers. Episode groupers focus on aggregating claims by feeding them through a predetermined software algorithm. As one of our TEP panels made clear, groupers are poorly suited to characterizing Medicare beneficiaries who often have multiple comorbidities.

Our proposed value based payment system is based on three observations regarding the traditional FFS Medicare program. The first, true of all health care systems, is that physicians practice in different settings that affect any consideration of how "value" will be measured. Some physicians practice primarily in the hospital, while others practice primarily in outpatient settings. Second, Medicare beneficiaries, whether disabled or elderly,[1] are characterized by multiple co-morbidities, which must be managed simultaneously, often by a variety of providers. Third, Medicare beneficiaries currently are not required to have a primary care physician who coordinates the care they receive in the hospital, or from medical or surgical specialists.

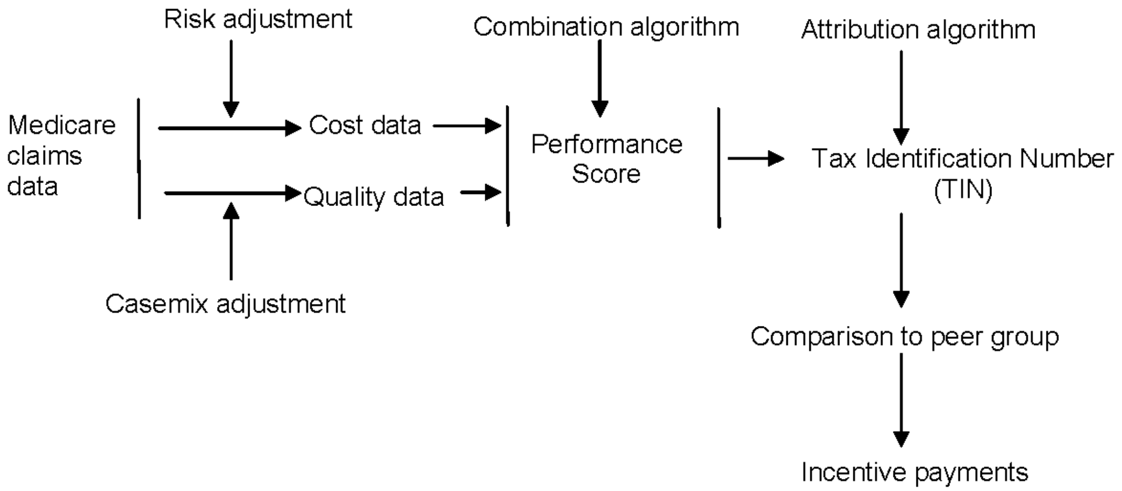
Our response to these observations has been to design a system that:

- Is tailored to the physician's practice setting (inpatient versus outpatient);
- Creates the rules for physicians to receive incentive payments, some of which are condition-specific, while not attempting to segment the beneficiary's co-morbidities into separate episodes of illness; and
- Balances a realistic approach to the lack of coordinated care in today's FFS Medicare program against the desire to move towards greater coordination of care in the future.

[1] Our proposed system does not cover beneficiaries with end-stage renal disease.

A simple diagram that summarizes the steps in our proposed approach is shown in Figure E.1:

Figure E.1: Our Overall Approach to Performance Assessment and Payment



We begin with Medicare claims data on the left-hand side of the diagram. The claims data contain all the information necessary to compute measures of cost and quality. The cost data are risk-adjusted to remove the effects of variables over which the physician has no control. The quality data are case mix adjusted by the definitions of the patients to whom the quality measures are applied (e.g., the measures of quality of care for diabetics are computed only for diabetics).

In order to tailor the system to the physician's practice setting, we designed two approaches to measuring physician performance. For physicians who practice in the hospital, the encounter with the patient – and the inpatient physician's responsibility – is triggered by a hospitalization, not an episode of illness. In FFS Medicare, hospitalizations are assigned a unique Medicare Severity – Diagnostic Related Group (MS-DRG) identifier. Thus, we refer to our approach for inpatient physicians as the "MS-DRG" approach.

Physicians practicing primarily in outpatient settings are responsible for an entire year of care for a beneficiary. Thus, we refer to our approach for outpatient physicians as the "beneficiary level" approach.

While recognizing that Medicare beneficiaries are not required to designate a primary care physician or to have their care coordinated by a single physician or physician group, we have designed both the MS-DRG and beneficiary level approaches to encourage greater accountability and coordination of care in three ways. First, we evaluate cost and quality, and recommend making incentive payments, at the tax identification number (TIN) level, rather than the individual physician. Second, in the MS-DRG approach we evaluate care over periods lasting 30 or 60 days post-discharge, thereby holding physicians responsible not only for cost and quality during the inpatient stay, but also for a "bundle" of post-acute care services. Third, the MS-DRG approach attributes beneficiaries to TINs based on the proportion of bills submitted by each TIN during the observation period. In the beneficiary approach, we hold the physician accountable for an entire year of expenditures for beneficiaries who are assigned to the TIN using a "plurality of non-hospital evaluation and management visits" attribution rule.

Although our proposed system is based on standard approaches to risk adjusted costs and nationally vetted quality measures, there are a few innovations in our system.

- We modified the algorithm developed by Billings, Parik, and Mijanovich (2000) to evaluate avoidable emergency department visits so that it could be used at the TIN level.
- Based on advice from our TEP panel, we used concurrent (same year) condition categories (CCs) to risk adjust the TINs cost data in the Beneficiary approach.
- We proposed and tested data envelopment analysis (DEA) as a way to combine cost and quality measures into a single performance score in the beneficiary approach.

We have given careful thought to the translation of performance measurement into incentive payments:

- The value based purchasing system would result in annual incentive payments for good performance during the previous year – not permanent increases in Medicare's physician fee schedule. The incentive payments are designed to reward good performance, not general increases in the cost of doing business, and the incentive payments would have to be earned each year.
- Incentive payments would be given only to physicians (TINs) who have enough attributed and eligible beneficiaries to produce a reliable estimate of cost and quality for each measure. Physicians with few attributed Medicare beneficiaries would not be eligible for incentive payments. Nor would physicians with few attributed beneficiaries in a specific disease category.
- Separate incentive payments would be made for each category (group) of performance measures. Thus, a physician could earn an incentive reward for appropriate monitoring of her diabetic patients, and another reward – perhaps based on many of the same beneficiaries – for appropriate breast cancer screening, or avoiding unnecessary visits to the emergency department. Combining different quality measures into one overall performance score runs counter to the literature suggesting that excellence in care for one medical condition is not necessarily indicative of excellence in the treatment of other conditions. In addition, combining multiple quality measures into one overall performance score would require dropping a physician from consideration of incentive payments if the physician lacked adequate sample size for even one quality measure.

Chapter 1

I. Introduction

The Centers for Medicare and Medicaid Services (CMS) is working to design a payment system that can reward physicians in traditional fee-for-service (FFS) Medicare for efficient and high quality care. "Alternative Approaches to Measuring Physician Resource Use" is a research project designed to develop creative ways to measure cost and quality in physician services, to support CMS in this ambitious undertaking. 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. This project is part of a growing body of work designed to improve the measurement of value in health care services, and thus make it possible to reward providers who offer greater value.

Under value-based purchasing practices, physicians face payments that reflect not only the 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" and in September, 2008 CMS contracted with the University of Minnesota (UMN) team to begin work in that area.

However, 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.

The UMN team included as well certain subcontractors and consultants who bring special methodological expertise[2] and management experience.[3] In the first year of work, the University of Minnesota (UMN) team focused on the fundamental issues of measuring and attributing risk-adjusted costs of care to physicians. Early in our work, we recognized that our approach needed to treat two groups of physicians differently –those who primarily provide care for hospitalized beneficiaries and those who primarily provide outpatient care. That decision framed the paths for much of the work that followed. In the second year of work, 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 measures and worked with panels of national experts formally to assess the relative importance of each measure, and our proposed approaches to combining cost and quality into a measure of value or performance. In the last phase of the project we obtained the final multi-state datasets on which the analyses in this report are based and applied our algorithms to those data to

[2] Professor Tom MaCurdy at Acumen LLC; Dr. Ateev Mehrota at RAND Corporation; and Professor William Greene at New York University.

[3] Professor Robert Coulam at Simmons College.

incorporate the analytic results into a comprehensive value based purchasing system for FFS Medicare.[4]

The result is a proposed payment system for value based purchasing of inpatient and outpatient physician services. This Final Report provides a comprehensive description and justification for that payment system. A set of Appendixes provide more technical support for specific elements of the payment system we propose. As will become clear in the material that follows, it is important to resolve some challenging technical issues in order to put in place a system of measurement and rewards for physicians that is fair, actionable, and meaningful to physicians, to Medicare, and to the wider community of stakeholders with less direct involvement but still much at stake.

Our report is organized as follows:

- Chapter 2 below begins by providing an overview of the complete system we propose. The intent here is to give the reader an understanding of the overall architecture of the system, as a guide for organizing the more detailed discussion that follows.

- Chapters 3 and 4 then lay out the details of our proposal for physicians practicing in an inpatient setting (the "MS-DRG approach") and those practicing in an outpatient setting (the "Beneficiary approach").
    - Both chapters detail and justify the methods we propose for attribution, measuring and risk-adjusting cost, measuring quality, and combining cost and quality measures into a working payment system.
    - In each case, as more technical discussion may be demanded by some readers, we refer as appropriate to specialized technical appendixes addressing these issues.

- The final four chapters (5-8) review certain implementation issues – e.g., the steps to follow to implement both MS-DRG and beneficiary approaches as well as what we would identify as the lessons of our work, beyond the specific proposals we are making .

[4] In addition, in fall 2011, we took on one large additional task not in the original scope of work. Congress instructed CMS, by January 2012, to choose an episode grouper to support value-based purchasing. We organized and ran a special panel of physician experts to evaluate and compare four prototype episode groupers. That work was not directly related to the scope of work under this contract, and is not a part of this report.

Chapter 2

General Overview

I. Introduction

The purpose of this project is to assist CMS with the design of a value-based payment system for physicians serving Medicare patients. The project has developed a set of cost and quality-of-care measures that can be used to summarize physician performance and determine whether physicians will receive an incentive reward based on the quality and cost of services they provide to Medicare beneficiaries in the traditional fee-for-service (FFS) Medicare program. In order to be successful, such a system must be considered fair, actionable and meaningful by both physicians and the Medicare program.

Our proposed approach is comprehensive, conceptually simple, transparent, reliable, and flexible.

- It is *comprehensive* in that our system represents a "turn-key" package that moves from Medicare claims data through the entire sequence of events that ends in incentive payment policy. We have developed performance measurement systems both for physicians who practice predominantly in outpatient settings and predominantly in inpatient settings. Calculation of valid performance measures is limited only by the same sample size restrictions that confront any other profiling system. We address the small sample size problem in several ways, discussed below.

- It is *conceptually simple* in that we combine cost and quality measures that are readily obtainable from Medicare claims data into a single measure of "value" or "performance" that addresses the common-sense question, "What are Medicare beneficiaries and taxpayers are getting for the money spent on physician care for beneficiaries?"

- It is *transparent in that* we have used publicly available software for every computation with only one exception (potentially preventable rehospitalizations).

- It is *reliable* in the sense that:

    - The majority of our quality measures are taken from those vetted by national organizations such as the National Quality Forum (NQF), the National C*ommittee on* Quality Assurance (NCQA), and the peer-reviewed literature on quality of care.

    - Where we have added novel measures of cost and quality (e.g., avoidable rehospitalizations and avoidable Emergency Department (ED) visits), it is the *application* of the particular measure that is innovative, not the measure itself.

    - Where we have added new approaches to combining the two types of measures, we do so by comparing physicians to each other to establish norms of practice, not to some external standard.

    - In addition, we convened two technical expert panels to review all aspects of our system and incorporated their advice into the final design.

- It is *flexible* in that our system can accommodate any measures of health care quality and cost, any attribution system, and any risk-adjustment system that CMS (in consultation with stakeholders and others) might choose.

Our approach avoids some common controversies. For example, we do not make any judgment regarding the cost-effectiveness of any test or treatment. Nor do we make any judgment regarding the monetary cost or monetary value of an incremental increase in the quality of care received by Medicare beneficiaries. We do compare the costs incurred by physicians who are producing the same level of quality, and we judge more expensive physicians (holding quality constant) to be less *efficient*, because they exhibit a lower level of "performance." We find that physicians with low performance scores generally tend to have both higher risk-adjusted costs and lower quality.

We are aware that several respected analysts do not believe that systems to measure physician cost and quality are "ready for prime time" (Mehrotra, et al., 2010). These doubts are based on legitimate concerns about the sensitivity and reliability of attribution algorithms, cost and quality measures, and small sample sizes. We share these concerns, but our proposal is based on several countervailing observations:

- CMS is mandated by the 2010 ACA legislation to implement a value based payment system. Failure to implement that system is not a viable policy alternative.
- The shortest route to improved measures of cost and quality may be to implement the measures we now have available. Feedback from the affected parties will come quickly, and if CMS is responsive, legitimate and constructive criticism will result in improved measures.
- A good example of measure improvement is the risk adjustment methods that are used to compute payments to Medicare Advantage (MA) plans. CMS has been applying various forms of risk adjustment to cost measures since the inception of the Medicare Advantage (as it is called now) program in the 1980s. Those risk adjustment methods have seen steady improvement and while there still are some issues to be addressed (GAO, 2012)[5] the system is working reasonably well.
- Many quality measures have been vetted by national organizations such NQF and NCQA, and many are computable from Medicare claims. More and better measures are being developed each year. As they become available, they can be incorporated into the proposed approach.
- The issue of small samples can be addressed in several ways:
    - By limiting eligibility for incentive payments to those physicians who demonstrate "meaningful participation" in the Medicare program. Meaningful participation is defined as having enough attributed beneficiaries in a particular

[5] The somewhat unsurprising finding from the GAO report is that when organizations' payments are dependent on diagnostic coding (i.e., MA plans), the organization codes diagnoses more carefully than when payments are not so dependent on diagnostic coding (i.e., FFS Medicare providers).

category (e.g., diabetics) for CMS to develop a reliable estimate of the physician's cost and quality.

    - By making separate awards for each quality measure, rather than combining measures with large and small numbers of eligible beneficiaries into a single performance metric. (The significance of this condition will become clearer in subsequent chapters.)
    - By noting that random variation in cost and quality measures becomes less important as the physician (or TIN) increases its participation in the Medicare program. Any individual measure will be subject to sampling variation, but the sampling variation in the *mean* of ten different quality measures (and thus ten different incentive payments) in one year will be smaller than the sampling variation in any one quality measure in one year.
    - By noting that in the presence of pure, random sampling variation in a physician's cost and quality measures, the mean incentive payment over multiple years of participation in the Medicare averages out to the true mean. The real problem is not small samples, per se, but systematic *bias* in cost and quality measures that would advantage or disadvantage a provider year after year. Of course, it is important to make the physician aware of the degree to which her performance measures in any year are affected by small samples (e.g., through presentation of confidence intervals).

II. The Essentials of Our System

We began our analysis with a focus on the broad policy objective – the development of reliable profiles of physician performance to serve as the basis of a value based payment system. We then developed a set of methods to achieve that goal. We did not begin our analysis with a predetermined format or "tool" such as episode groupers. Episode groupers focus on aggregating claims by feeding them through a predetermined software algorithm. Groupers are poorly suited to characterizing Medicare beneficiaries who often have multiple comorbidities.

Our proposed VBM system is based on three observations regarding the traditional FFS Medicare program. The first, true of all health care systems, is that physicians practice in different settings that affect any consideration of how "value" will be measured. Some physicians practice primarily in the hospital, while others practice primarily in outpatient settings. Second, Medicare beneficiaries, whether disabled or elderly,[6] are characterized by multiple co-morbidities, which must be managed simultaneously, often by a variety of providers. Third, Medicare beneficiaries currently are not required to have a primary care physician who coordinates the care they receive in the hospital, or from medical or surgical specialists.

Our response to these observations has been to design system that:

- Is tailored to the physician's practice setting (inpatient versus outpatient);

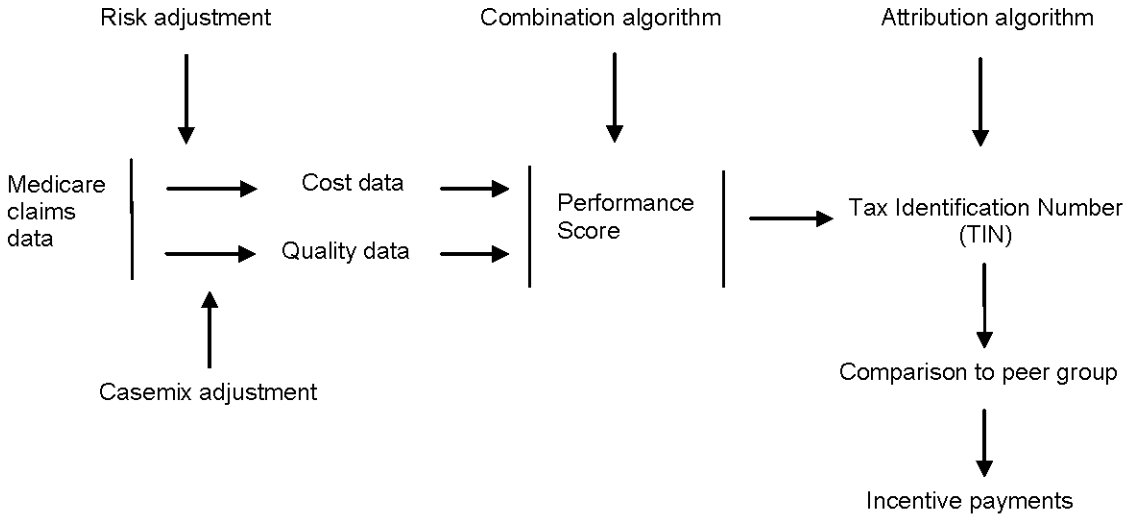
[6] Our proposed system does not cover beneficiaries with end-stage renal disease.

- Creates the rules for physicians to receive incentive payments, some of which are condition-specific, while not attempting to segment the beneficiary's co-morbidities into separate episodes of illness; and

- Balances a realistic approach to the lack of coordinated care in today's FFS Medicare program against the desire to move towards greater coordination of care in the future.

A simple diagram that summarizes the steps in our VBM system is shown in Figure 2.1:

Figure 2.1: Our Overall Approach to Performance[7] Assessment and Payment



We begin with Medicare claims data on the left-hand side of the diagram. The claims data contain all the information necessary to compute measures of cost and quality. The cost data are risk-adjusted to remove the effects of variables over which the physician has no control. The quality data are casemix adjusted by the definitions of the patients to whom the quality measures are applied (e.g., the measures of quality of care for diabetics are computed only for diabetics).

In order to tailor the incentive payment system to the physician's practice setting, we designed two approaches to measuring physician performance. For physicians who practice in the hospital, the encounter with the patient – and the inpatient physician's responsibility – is triggered by a hospitalization, not an episode of illness. In FFS Medicare, hospitalizations are assigned a unique Medicare Severity – Diagnostic Related Group (MS-DRG) identifier. Thus, we refer to our approach for inpatient physicians as the "MS-DRG" approach.

Physicians practicing primarily in outpatient settings are responsible for an entire year of care for a beneficiary. Thus, we refer to our approach for outpatient physicians as the "beneficiary level" approach.

[7] As noted above, we use the term "performance" to refer to the *combination* of cost and quality data that represents "value" in VBM systems.

While recognizing that Medicare beneficiaries are not required to designate a primary care physician or to have their care coordinated by a single physician or physician group, we have designed both the MS-DRG and beneficiary level approaches to encourage greater accountability and coordination of care. That emphasis begins with the operational definition of a "physician."

There are two types of coordinated care of interest. The first is coordination of care with other physicians and other health care professionals (if any) within the physician's own practice. The second is coordination of care with physicians and other health care professionals in the hospital and other practices. Our first step towards encouraging better care coordination is to define the "physician" in both the MS-DRG and beneficiary level approaches not as an individual physician, but as a tax identification number or TIN. Basing payment and accountability on activity at the TIN level encourages physicians to ensure that care coordination has a strong foundation in their own practices.

To encourage coordination of inpatient and outpatient care, the MS-DRG approach opens an observation "window" when the beneficiary enters the hospital. (We present results for both 30- and 60-day windows.) The cost for the inpatient stay is the total cost incurred on behalf of the beneficiary during the 30- or 60-day period. Quality measures are collected not only during the inpatient stay, but also for the post-discharge period within the observation window. Thus, even though Medicare has not moved to "bundling" inpatient and post-acute care for *standard payment* purposes, our *performance assessment and incentive payment system* is based on a bundled inpatient and post-discharge set of cost and quality measures, as a way to reward good coordination of care.

For physicians associated with the inpatient stay, we use a *proportionate* attribution rule, so that the beneficiary's cost and quality data are divided among TINs based on the proportion of each TIN's total Part B costs during the observation window. This approach is sensible for inpatient-triggered events because the specialist providing the most expensive treatment, often the surgeon, represents the physician who carries the greatest responsibility for the patient's immediate inpatient care. Under the current system, that physician may not have the greatest control over the patient's post-discharge care, but our attribution system encourages the highest billing physician – who often has the clearest idea of the patient's physiological prognosis – to play a greater role in the coordination of the patient's inpatient and post-discharge care.

Our encouragement of coordinated care is even more aggressive for physicians practicing primarily in outpatient settings. In addition to using the TIN as the unit of analysis, we encourage greater coordination of care through choice of the observation window. Rather than dividing the beneficiary with multiple comorbidities into many separate disease categories or episodes of illness, we use the beneficiary's entire year of cost and quality experience as the observation window.

Our beneficiary level attribution rule also encourages greater care coordination. Rather than parceling out the beneficiary's care proportionately to every provider who billed for the beneficiary – an approach that makes sense for inpatient physicians – we assign beneficiaries to the physician who provided the most (a plurality) of the beneficiary's non-hospital evaluation and management (NH-E&M) visits. We refer to this attribution rule as a *plurality* rule.

The cost and quality data then are combined to produce a "performance score."

- In the MS-DRG analysis, the data are combined by (1) forming peer groups of physicians who account for similar proportions of the patient's total cost; (2) creating tiers of quality; and (3) comparing the cost of care for physicians in each level of quality. Separate performance scores are computed for each MS-DRG

- In the beneficiary level analysis, the combining of cost and quality measures is done through data envelopment analysis (DEA) in which the beneficiary's annual total cost of care is the "input" and quality measures are the "outputs." DEA produces a measure of the cost incurred by a specific physician relative to other physicians who are producing the same level of quality.

In both cases, combined cost-quality scores are being compared to peers, not to arbitrary standards.

The MS-DRG and Beneficiary approaches interact in interesting ways. The same beneficiary can contribute to performance assessments for both inpatient physicians under the MS-DRG approach and outpatient physicians under the Beneficiary approach. Moreover, the same *physician* can earn incentive payments under both the MS-DRG and beneficiary-level system if the physician has provided a plurality of the beneficiaries NH-E&M visits and also has provided care within the 30- or 60-day observation window associated with an inpatient stay.

We have given careful thought to the effect of our system on incentives for physicians to improve their care of Medicare beneficiaries. In that respect, our system represents an important departure from current Medicare physician payment policy in the following respects:

- The value based purchasing system would result in annual incentive payments for good performance during the previous year – not permanent increases in Medicare's physician fee schedule. The incentive payments would have to be earned each year.
- Separate incentive payments would be made for each category (group) of performance measures. Thus, a physician could earn an incentive reward for appropriate monitoring of her diabetic patients,[8] and another reward – perhaps based on many of the same beneficiaries – for appropriate breast cancer screening, or avoiding unnecessary visits to the emergency department. Combining different quality measures into one overall performance score runs counter to the literature suggesting that excellence in care for one medical condition is not necessarily indicative of excellence in the treatment of other conditions. In addition, combining multiple quality measures into one overall performance score would require dropping a physician from consideration of incentive payments if the physician lacked adequate sample size for even one quality measure.

[8] There could, of course, be multiple measures of monitoring for diabetic patients (HbA1c, LDL-C, medical attention for nephropathy, etc.).

Chapter 3

The MS-DRG Approach

The objective of the Medicare Severity-Diagnosis Related Group (MS-DRG) approach is to develop measures that associate subsequent use of health services with physicians or physician groups who practice in hospital settings. For reasons we will discuss, physicians/physician groups are represented by their Tax Identification Number (TIN). The performance measures are a combination of risk adjusted cost and quality measures. The cost measures are computable from claims data, while the quality measures are either claims computable or are collected at the hospital level by CMS.

The essential features of the MS-DRG approach are:

I. Defining the MS-DRG episode -- To clarify what the MS-DRG approach is, we must begin by defining MS-DRG episodes precisely.

II. Data, Physician Entity and Attribution Methodology – There are complexities in attributing cost and quality measures to physicians in the hospital setting. This section addresses how we propose to deal with those problems.

  A. Episode Attribution to Single Physician Entity
  B. Episode Attribution to Multiple Physician Entities

III. Cost measures – The estimation of costs for hospital episodes involves a series of important technical questions, some of them obvious – e.g., risk adjustment – others less obvious – e.g., questions of aggregation across different conditions. This section addresses those questions

  A. Costs included/excluded in Episode
  B. Standardizing Costs
  C. Risk adjustment
  D. Problem with Low Ns and Low Ps
  E. Aggregation across conditions
  F. Determining TIN peer group for cost profiles

IV. Quality measures – The quality measures we use to appraise the physician's performance must be fair, actionable and meaningful to physicians. All of the measures we have selected, as well as our approach to combining them, were vetted by a panel of national technical experts. In addition, we have addressed certain technical issues in using those measures – risk adjustment, construction of composite quality measures, and aggregation across conditions.

  A. Selection of candidate individual measures
  B. Review by expert panel
  C. Rating process and results – including weighting of measures
  D. Attribution, framing, and risk adjustment of quality measures
  E. Results
    1. Episode level measures

2. Hospital level measures

V. Issues in constructing the composite quality measures – The weights provided by the TEP panel make it possible to combine quality measures in a meaningful way. This section reviews some of the most important issues in doing that, and presents our initial results.

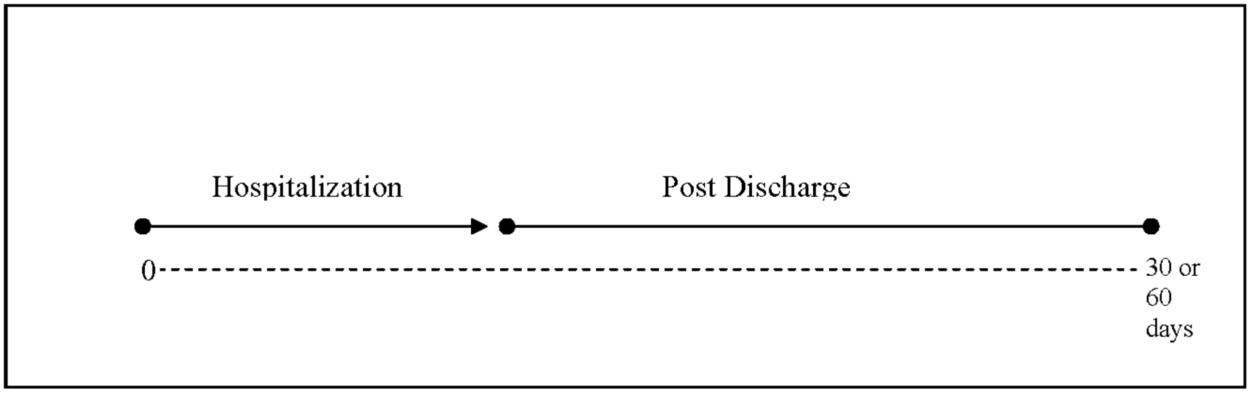
I. Defining the MS-DRG Episode

In fee-for service Medicare, a separate payment is made for each service. That payment system rewards healthcare providers for delivering more services, but provides no incentive to coordinate those services efficiently across time and settings.

The MS-DRG approach compares the cost and quality of hospital-based physicians across similar MS-DRG episodes, thus creating the basis for a value-based modifier system to reward hospital-based physicians for providing high quality and low cost care.

Figure 3.1 depicts an MS-DRG episode. An MS-DRG episode is initiated when a Medicare beneficiary is hospitalized for a particular condition. The hospitalization that starts an MS-DRG episode is known as the index hospitalization. The physician(s) providing care during the index hospitalization are then attributed all costs of care and quality (processes and outcomes) associated with the care incurred within a predefined period of time – known as the episode window (e.g. 30 days or 60 days). Because the actual hospital admission is not under the control of the target physician, the basic DRG costs are not included in this calculation, but any additional costs (e.g., outlier payments for long stays) are counted, as is the Part B re-imbursement for physician services. The cost and quality measures, when adjusted appropriately for beneficiary case mix, facilitate comparison of hospital -based physicians' performance across similar MS-DRG episodes.

Figure 3.1: MS-DRG Episode



While the MS-DRG approach can theoretically be applied to all MS-DRGs, allowing for an exhaustive comparison of all hospital-based physicians, we demonstrate how this approach works for a selected set of 11 conditions, chosen on the basis of their high frequency and cost for the Medicare program:

1. Acute Myocardial Infarction (AMI)

2. Congestive Heart Failure (CHF)
3. Chronic Obstructive Pulmonary Disease
4. Pneumonia
5. Bronchitis
6. Stroke
7. Hip fracture
8. Hip replacement
9. Knee replacement
10. Cholecystectomy
11. Back pain

Some of these conditions are represented by different types of medical and/or surgical procedures with multiple MS-DRG codes – like AMI, that can be managed medically (Medical AMI), with percutaneous coronary intervention (AMI with PTCA), or surgery (AMI with CABG). The list of the selected conditions and their corresponding MS-DRG codes are given in Appendix 1 (Tables 1 through 11).

To choose the appropriate length of the MS-DRG episode for our analyses, we compared the proportion of a beneficiary's annual Medicare expenditure captured by 30-day, 60-day and 90-day episodes, for our selected set of MS-DRG episodes. 30-day episodes accounted for 45-65 percent of the beneficiaries' average annual Medicare costs, 60-day episodes accounted for 60-80 percent of the beneficiaries' average annual Medicare costs, while 90-day episodes accounted for 70-85 percent of the Medicare costs (data not shown). The 90-day episodes offered only a 5 percent gain over the 60-day episodes in capturing costs. We therefore chose 30- and 60-day episode windows for the MS-DRG approach. The MS-DRG physician technical expert panel (TEP)[9] agreed that 30-day and 60-day episode windows were clinically meaningful for hospital based physicians.

The costs included in the MS-DRG episodes are shown in Figure 3.2. We assumed that hospital-based physicians are not in control of the initial hospitalization, and hence excluded the hospital DRG payment for the initial hospitalization from the MS-DRG episode costs. However we include in the episode costs:

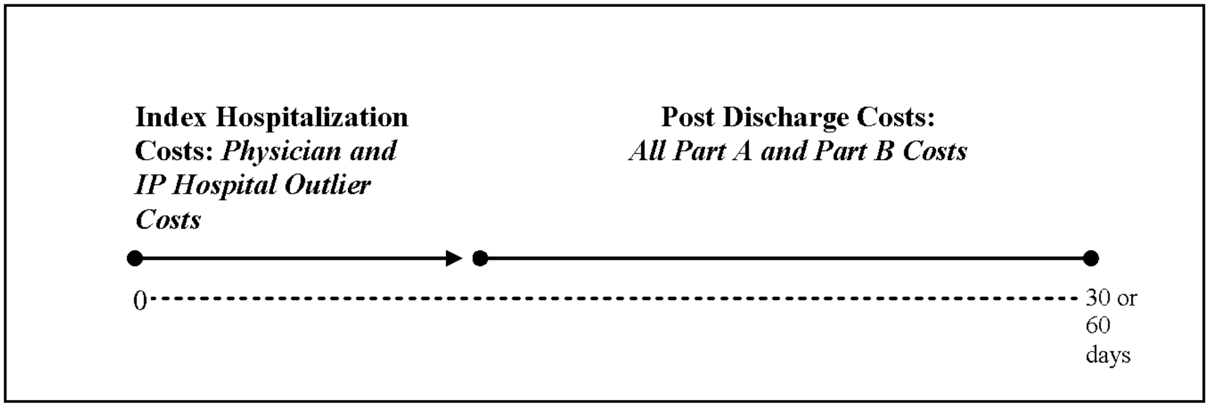
- The part B payments for the initial hospitalization and the outlier payment for the initial hospitalization (if any), as we assume these to be determined by treatment choices made by the physicians.

[9] Our work with the MS-DRG TEP is discussed in detail in Section IV.

- All post discharge Part A and Part B costs incurred by the beneficiary up to 30- or 60- days.

We use risk adjustment to exclude post-discharge costs associated with trauma within the episode windows. The MS-DRG TEP agreed with this approach.

Figure 3.2: MS-DRG Episode Costs


Index Hospitalization Costs: Physician and IP Hospital Outlier Costs
Post Discharge Costs: All Part A and Part B Costs
0
30 or 60 days

II. Data, Physician Entity and Attribution Methodology

The MS-DRG approach used data from ten states chosen to represent variation in location, size, and practice patterns – Colorado, California, Florida, Kansas, Louisiana, Minnesota, New Jersey, Virginia, Vermont, and Washington. Our beneficiary sample consisted of beneficiaries residing in these 10 States with at least one inpatient hospital admission for any of the 11 identified MS-DRGs in 2008. 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. We used Medicare claims in 2008 through the first quarter of 2009 for capturing the entire episode experience for these beneficiaries. Medicare claims from 2007 for these beneficiaries were used for the purpose of risk adjustment. Medicare Part D claims were not used. *Index* hospitalizations outside the beneficiary's state of residence were excluded, as our goal was to specifically profile hospital-based physicians in the beneficiary's state. But we did include *rehospitalizations* outside a beneficiary's state of residence. Our physician sample included all hospital based physicians who treated these beneficiaries during their index hospitalization for the selected MS-DRGs.

To profile hospital based physicians, we need to attribute a beneficiary's care and costs of care associated with an MS-DRG episode to an appropriate physician entity. Physician entities that can be identified using beneficiary claims are unique physician NPIs (National Provider Identifiers) and physician TINs (Tax Identification Numbers). Physician TINs are not directly available on Medicare claims. They can be obtained by linking claims to Provider Enrollment, Chain and Ownership (PECOS) data. Unlike NPIs, Physician TINs are not unique – NPIs can bill under multiple TINs. TINs might also represent larger group practices with multiple NPIs. However TINs represent the level at which physicians are reimbursed and have current incentive payments structured.

A separate question is whether we attribute all care to one physician entity or somehow divide responsibility for the care among multiple physician entities. We examined both approaches, as discussed in the next section.

A. Episode Attribution to Single Physician Entity

We first attempted to attribute all care and costs of care incurred in the MS-DRG episode to a single physician entity responsible for coordinating care during the initial hospitalization and making post-acute care decisions. Episodes could be attributed to different physician entities depending on whether the MS-DRGs were medical or surgical.

Medical MS-DRGs. For medical MS-DRGs the episode could be attributed to either:

- the attending NPI listed on the hospitalization Part A claim or
- the NPI primarily billing on the Part B claims associated with the initial hospitalization.

Using data from Colorado, we examined how often the listed attending Part A NPI was the primary part B biller for medical MS-DRGs. Table 3.1 shows the ranges for the percentages of matches and of Part B payments for each of the three groups of conditions (AMI, CHF, and COPD). As seen in Table 3.1, column A, the match between the attending physician and the operating NPI is quite high – as high as 92 percent, and at least 57 percent in the lowest case. But as shown in column B, between 22 percent and 37 percent of costs from part B bills for the initial hospitalization went to the listed attending Part A NPI. Across conditions, the proportion of the attending physician's payment decreased with increasing severity of the hospitalization, indicating that more physicians were involved in the care of complex cases.[10] The listed attending Part A NPI was the highest billing Part B NPI only for 38-52 percent of the medical MS-DRG cases (column C in the table).

In view of these results for medical MS-DRGs (AMI, COPD and Pneumonia/Bronchitis), attributing all care and cost of care to the attending physician identified from Part A claims may not be appropriate.

Surgical MS-DRGs. For surgical MS-DRGs, the appropriate physician entity to which to attribute the episode could be one of three possibilities:

- the attending NPI listed on the Part A hospitalization claim,
- the operating NPI listed on the Part A claim, or
- the operating NPI primarily billing on the Part B claims associated with the hospitalization.

[10] See Appendix 2, Table 1 for detail by type of case.

Medical MS-DRGs: Match Rate Between Attending and Operating Physician and Percent of Part B Payment for initial Hospitalization billed to Attending Physician (Percentage range across MS-DRGs within each condition)

Condition/ Treatment	MS-DRGs included	COLUMN A Percent MATCH B/W OPERATING AND ATTENDING NPI (PART A)	COLUMN B percent INDEX PART B PAYMENT BILLED BY PART A ATTENDING NPI	COLUMN C percent MATCH B/W ATTENDING NPI (PART A) AND HIGHEST BILLING NPI (PARTB)
AMI	280-282	80 percent-85 percent	22 percent-29 percent	42 percent-51 percent
CHF	291-293	83 percent-92 percent	27 percent-34 percent	38 percent-46 percent
COPD	190-192	57 percent-62 percent	29 percent-37 percent	46 percent-52 percent

Source: See Appendix 2 Table 1

Table 3.2: Surgical MS-DRGs: Match rate between Part A attending, Part A Operating, and Part B Primary Operating Physician (Percentage range across MS-DRGs within each condition)

Condition/Treatment	MS-DRGs Included	COLUMN A percent; MATCH B/W OPERATING AND ATTENDING NPI (Part A)	COLUMN B percent; MACTH B/W OPERATING NPI (PART A) AND PRIMARY OPERATING NPI (Part B)	COLUMN C percent; MATCH B/W ATTENDING NPI (PART A) AND PRIMARY OPERATING NPI (Part B)
AMI with CABG	231,232, 233, 234	73 percent-83 percent	33 percent-57 percent	0 percent-50 percent
AMI with PTCA	246, 247, 248, 249	77 percent-86 percent	55 percent-68 percent	36 percent-54 percent
Hip Replacement	469, 470	99 percent-100 percent	66 percent-72 percent	64 percent-72 percent
Hip Fracture	535,536	69 percent-71 percent	7 percent-10 percent	11 percent-13 percent

Source: appendix 2 Table 2

For surgical MS-DRGs we examined the level of congruence between the listed attending NPI, the listed operating NPI, and the primary operating NPI on Part B claims. The results are summarized in Table 3.2 above. The agreement between the listed attending and listed operating NPIs on the hospitalization claim (shown in Column A) ranged from 70 percent to 85 percent, with the exception of hip replacement. The match between listed attending NPI and primary operating Part B NPI (shown in Column C) was poorer, with the match rate as low as 0 percent-50 percent for AMI with CABG and 11-13 percent for hip replacement. Equally poor was the congruence between the listed operating NPI on the hospitalization claims and the primary operating NPI on the Part B claims (Column B), with the match rate of 7 percent-10 percent for hip replacement and 33 percent-57 percent for AMI with CABG.

On moving from NPIs to TINs, the match between the listed operating Part A physician and primary operating Part B physician improved, as seen in Column B of Table 3.3 below. After moving to TINs the match rate between Part A and Part B physicians for surgical MS-DRGs (shown in Column B) ranged from a high of 85 percent for hip replacement to a low of 22 percent for hip fracture. Moving from NPIs to TINs marginally improved the rate of agreement between the listed attending Part A physician entity and the highest billing part B physician entity from 25 percent-35 percent to approximately 60 percent (data not shown).

Table 3.3: Match Rate between Part A and Part B Operating NPIs and TINs for Surgical MS-DRGs

Condition/Treatment	MS-DRG	COLUMN A Percent match B/W operating NPI (Part A) and primary operating (Part B)	COLUMN B Percent match B/W operating TIN (Part A) and primary operating TIN (Part B)
AMI with CABG	231,232,233, 234	56 percent	75 percent
AMI with PTCA	246, 247,248, 249	57 percent	78 percent
Hip Replacement	469, 470	69 percent	85 percent
Hip Fracture	535, 536	8 percent	22 percent

Problems in moving from NPIs to TINs. While moving from NPIs to TINs helps to improve certain matches, it creates a new challenge for profiling medical MS-DRGs. As shown in Tables 3.4 (a) and 3.4 (b) below using multi-state data , 7 percent of the listed attending Part A NPIs for Medical AMI and 18 percent of the listed attending Part A NPIs for Medical COPD could not be linked to TINs.

The problem here is easily explained. TINs for physician entities are available on linked-claims data only when the physician entity directly bills Medicare for a service (billed on Part B claims). This suggests that attending Part A NPIs for Medical MS-DRGs that did not have a matching TIN (as a consequence of not billing Medicare on Part B claims) were likely to be employed by hospitals. The rate of such hospital employed attending physicians who could not be linked to TINs varied by state, with the highest rates in California and New Jersey, and the lowest in Kansas and Virginia.

Table 3.4 (a): AMI MS-DRGs: Nature of Part A attending TINs across states after moving from NPIs to TINs

AMI	CA	FL	KS	LA	MN	NJ	VA	VT	WA	Total
# of AMI Index Admissions	23,087	25,988	4,632	6,345	5,176	14,738	1,187	9,236	5,843	96,232
# of Part A Attending TINs	4,424	3,887	435	778	324	2,774	126	887	538	14,173
Average Admissions per TIN	5	7	11	8	16	5	9	10	11	7
Median Admissions per TIN	2	2	2	2	2	2	2	2	2	2
Actual TINs (percent)	88	92	98	94	98	87	98	96	97	93
NPIs (percent)	11	6	1	5	1	12	1	3	2	7

Table 3.4 (b). COPD MS-DRGs: Nature of Part A Attending TINs across states after moving from NPIs to TINs

COPD	CA	FL	KS	LA	MN	NJ	VA	VT	WA	Total
# of COPD Index Admissions	19,131	23,502	4,072	6,630	3,433	12,188	760	9,270	4,336	83,322
# of Part A Attending TINs	4,422	3,609	470	940	261	2,710	111	871	546	13,940
Average Admissions per TIN	4	7	9	7	13	4	7	11	8	6
Median Admissions per TIN	2	2	3	3	2	2	2	2	2	2
Actual TINs (percent)	72	83	96	86	88	69	94	84	81	82
NPIs (percent)	27	16	3	14	11	30	6	16	19	18

Given the disagreement between the different approaches of identifying the attending and operating physician for Medical and Surgical MS-DRGs, and the challenge of assuming, ex-post, that a single physician entity was responsible for a beneficiary's hospitalization and post-discharge experience, *we pursued an alternative methodology of attributing MS-DRG episodes to multiple physician entities*.

B. Episode Attribution to Multiple Physician Entities

There are two ways of attributing MS-DRG episode costs to multiple physician entities:

i). Attributing the entire episode cost to all physician entities who provide care during the index hospitalization (all or none rule).

ii). Attributing episode costs to physician entities that provide care during the index hospitalization based on the proportion of care provided by the physician entity (proportionate rule).

Because the purpose of our MS-DRG approach is to develop performance measures for hospital based physicians, we wanted an attribution system that held individual physicians accountable for the treatment choices that they make during the beneficiary's hospitalization, and did so in a way that reflected how beneficiaries actually receive care. A proportionate rule of attribution, where a beneficiary's episode costs attributed to a TIN are a function of the proportion of the TIN's Part B billing during the beneficiary's index hospitalization, is best suited to achieve this objective.

We chose TINs over NPIs as the physician entity of choice for two reasons. Our preliminary work where we attributed MS-DRGs to NPIs resulted in a low volume of episodes per NPI (20-40 percent of NPIs had only one attributable episode for an MS-DRG), which created difficulties for any kind of performance attribution system for that large proportion of cases. We sought to attribute MS-DRGs to TINs both to improve the volume of episodes per physician unit, and more closely to approximate the team of providers that cares for a patient.

III. Cost Measures

A. Costs included and excluded in the MS-DRG Episode

In calculating the costs of the MS-DRG episode, the initial cost of the DRG is excluded from the MS-DRG episode. The following costs are included:

- Part B costs for the index hospitalization,
- outlier costs associated with the hospitalization and
- all post discharge Part A and Part B costs within the 30- or 60- day episode window.

We do not include prescription drug costs as we did not have Medicare Part D data. We also controlled for costs associated with trauma after the index hospitalization, assuming that trauma is exogenous. For instance if a beneficiary had severe head injury or traumatic amputation after the index hospitalization, we controlled for costs incurred in the episode due to such trauma using risk adjustment. More details of costs included and excluded in MS-DRG episodes are presented are presented below in our discussion of standardizing costs and risk adjustment.

B. Standardizing Costs

In our preliminary work with Colorado data we used unstandardized payment amounts found on the Medicare claims. Moving to the multi-state data we used standardized payment amounts for three reasons. First, Medicare adjusts payments to providers to reflect differences in local input prices. We standardized for those 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 low 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 skewed the cost data and risk adjustment. Using standardized costs removed problems arising from erroneous dollar amounts on claims.

We standardized 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 (2008) and the Medicare Payment Advisory Commission (MedPAC, 2006), we standardized 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 standardized costs. It will also provide a more complete picture of precisely what cost elements we are attributing to TINs – and for which we are thus holding them accountable.*

i. Part A Claims

- 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 operational 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 the 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.

- 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.

- 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.

- 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.

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

ii. Part B Claims

- 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 Healthcare Common Procedure Coding System (HCPCS) and modifier codes. For each line item HCPCS, we calculated the standardized allowed charge by removing the influence of Geographic Practice Cost Indices (GPCI) from allowed charges. In the standardization process we allowed for differences in costs if the care was provided in a facility versus a non-facility setting. All components of RVUs (work, practice expense, and malpractice) were standardized.

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

- 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 code over the entire sample of beneficiaries

- Ambulatory surgical center services: We used national APC (ambulatory payment classification) payment rates to standardize each ASC claim.

Appendix 3 summarizes the costs of the MS-DRG episodes from the multi-state data for two conditions (AMI and COPD), in Tables 1 and 2 respectively. The interstate variation is lower for 30 and 60 day standardized episode costs, compared with the unstandardized costs. Cost standardization has removed the geographic variation in payment. The observed interstate variation in the standardized episode costs is due to the variation in intensity of service utilization across states.

C. Risk Adjustment

In order to maintain fairness while attributing costs to physician TINs, costs should be adjusted for factors beyond a TIN’s control. The purpose of risk adjusting costs is to control for these exogenous factors – usually beneficiary level covariates – and thus allowing a fair comparison of episode costs for TINs that treat different beneficiaries. Risk adjustment is usually performed by regressing standardized episode costs on a set of beneficiary level variables that affect costs but are beyond the physician’s control. The residuals from such a regression are the true risk adjusted costs for the beneficiaries’ episodes. Risk adjusted costs are best presented as such residuals because they represent actual dollar amounts that are more or less than the expected costs of the episodes.

The standardized 30 or 60 day episode cost for an episode for a selected MS-DRG for the *i*th beneficiary is:

$$Y_{it} = X_{it}\beta + \varepsilon_{it} \tag{3.1}$$

where:

t is the episode duration of interest (30 or 60 days)

X_{it} is a vector of exogenous beneficiary variables used for risk adjustment

β is a vector of coefficients, and

ε_{it} is the error term, representing the costs for which the physician TINs treating the episode are responsible.

The risk adjusted 30- or 60-day episode cost for an episode for a selected MS-DRG for the *i*th beneficiary is simply the residual from the above regression:

$$\hat{\varepsilon}_{it} = Y_{it} - X_{it}\hat{\beta} \tag{3.2}$$

$\hat{\varepsilon}_{it}$ can be positive or negative and represents the amount by which the beneficiary's actual cost for that episode is above/below the cost that would be expected, given the beneficiary 's value of X variables for the episode.

iii. Beneficiary Level Covariates

For risk adjustment, we adapted the CMS Hierarchical Condition Category (HCC) model (Pope, Ellis, Ash et al, 2000). A beneficiary with COPD could have multiple episodes for COPD for the year, so we risk adjusted costs for each unique beneficiary episode. We used a modified prospective model for risk adjustment where we considered all diagnoses occurring 12 months prior to a given index hospitalization to count HCCs for each beneficiary- episode.

Our risk adjustment model had the following beneficiary/beneficiary-episode level covariates:

a) 70 HCC diagnostic categories.

b) 24 Age/Sex groups

c) Medicaid status

d) Reason for Medicare eligibility

e) Dummy variables identifying whether the MS-DRG episode had major complications/comorbidities (MCC), complications/comorbidities (CC) or no comorbidities. MS-DRGs with MCCs and CCs have higher costs associated with them than those without any MCCs or CCs. In using these variables for risk adjustment we assumed that MCCs and CCs in an episode were *not endogenous*, that is, they did not result from care provided by physicians during the index hospitalization for the episode

f) Variable identifying if the beneficiary died within the episode window. Episodes where the beneficiary died within the episode window do not have claims until the end of the episode, and hence may have lower costs than expected

g) A severity variable which identified the counts of prior hospitalizations for the same condition a year prior to the beneficiary-episode

h) Another severity variable which identified the number of prior emergency department visits (not resulting in hospitalization) for the same condition a year prior to the beneficiary-episode

i) We also eliminated costs due to rehospitalizations for trauma from the episode costs by identifying if the beneficiary-episode had HCCs for trauma after the initial hospitalization. These HCC flags were included as variables in the right hand side of the risk adjustment model.[11]

We did not use disease interactions present in the original CMS-HCC model in our risk adjustment work. We did not consider it important given that our model is akin to a CMS-HCC model on a population that has already been stratified by diagnosis. Also, HCC diagnostic categories that were not present for a given condition were dropped from the risk adjustment model for that condition to ensure model parsimony.

[11] The HCCs that we assumed to be exogenous were: HCC 67- Quadriplegia, Other Extensive Paralysis; HCC 68- Paraplegia; HCC 69- Spinal Cord Disorders/ Injuries; HCC 75 - Coma, Brain Compression / Anoxic Damage; HCC 154 - Severe Head Injury; HCC 155- Major Head Injury; HCC 157- Vertebral Fractures w/o Spinal Cord Injury; HCC 161- Traumatic Amputation; HCC 164- Major Complications of Medical Care & Trauma; HCC 177- Amputation Status, Lower Limb/Amputation Complications.

iv. Running Risk Adjustment Models

Using the multi-state data we ran separate risk adjustment models with 30- and 60- day standardized episode costs as the dependent variable for each of the eleven conditions on which our work focuses (section I of this chapter explains our choice of these 11 conditions):

1) AMI: (a) Medical AMI (b)AMI with Percutaneous Transluminal Coronary Angioplasty (c) AMI with Coronary Artery Bypass Graft

2) CHF

3) COPD

4) Pneumonia

5) Bronchitis

6) Stroke: (a)Acute Ischemic Stroke (b)Stroke with Cerebral Infarction

7) Hip Fracture

8) Hip Replacement

9) knee replacement

10) Cholecystectomy: Separate models for (a) Laparoscopic Cholecystectomy (b)Non-Laparoscopic Cholecystectomy

11) Back Pain: (a) Medical Back Pain (b) Back Pain with Spinal Fusion (c) Back Pain with Other Back Procedures.

For conditions that are managed by different treatments – like AMI, Stroke, Cholecystectomy and Back Pain – we ran separate risk adjustment models for each treatment. As our dependent variable was non-zero and skewed, we ran generalized linear models with gamma regression[12] and log linked dependent variable.

The risk adjusted 30 or 60 day episode cost for an episode for a selected condition for the *i*th beneficiary was obtained from equations 3.1 and 3.2 above. The results from the regressions for all the selected conditions are presented in Appendix 19.

[12] The coefficient from the modified Park Test estimated with raw-scale residuals squared supported the choice of gamma family for the GLM models.

TIN's risk-adjusted 30- or 60 day episode cost for a selected condition is:

$$= \frac{\sum_{i=1}^{n_j} \hat{\varepsilon}_{it} \, p_{ij}}{n_j} \qquad (3.3)$$

Where:

$\hat{\varepsilon}_{it}$ is the risk adjusted 30- or 60- day episode cost for the *i*th beneficiary attributed to the *j*th TIN,

p_{ij} is the proportion of the beneficiary's episode attributed to the TIN (based on the proportion of the TINs part B billing for the MS-DRG), and n_j is the number of episodes of that condition attributed to the TIN.

In essence, a TIN's true responsibility for n_j episodes can be expressed as $\sum_{i=1}^{n_j} p_{ij}$, and can be considered to be equivalent to the number of "full time episodes" (FTEs) attributed to the TIN.

Tables 3.5(a) and 3.5 (b) below, show results before and after risk adjustment from the multi-state data for all selected MS-DRG conditions. Mean standardized and risk adjusted episode costs, as well as the mean standardized and risk-adjusted episode costs at the TIN level for all selected conditions are summarized in the table. The mean risk adjusted episode costs are residuals from equation (3.2) and hence are smaller in magnitude than the standardized costs. The mean risk adjusted costs at the episode level are negative when the standardized costs are positively skewed by high spenders. The number of episodes is usually more than the number of TINs managing the episode, except for AMI with CABG, ischemic stroke, hip fracture and non- laparoscopic cholecystectomy. For the aforementioned conditions the number of TINs managing the episode exceeds the number of episodes. There are three main conclusions to draw from the table:

- The mean risk-adjusted episode costs don't vary much by condition. We would ideally expect the mean risk-adjusted episode costs to be zero as they are residuals from our GLM model, but they are negative due to positively-skewed thick-tailed unadjusted episode costs.

- There is a large variation in both standardized and risj adjusted episode costs at the episode and TIN levels. The pattern of variance in risk adjusted episode costs is similar to that for unadjusted costs. The variance is higher for the more expensive conditions like Medical AMI, followed by CABG and PTCA. While the mean costs for 30- and 60- day episodes are not very different, the variance is much higher for 60-day episode costs.

- Variance in TIN risk adjusted episode costs is lower for CABG compared to PTCA and Medical AMI, as there are more TINs on average managing an episode of the former compared to the latter two. In summary, if more TINs on an average treat a condition during initial hospitalization, the variance in TIN risk-adjusted episode costs will be lower.

In summary, the variance in risk adjusted episode costs, at both the episode level and the TIN level, is similar to the variance in standardized episode costs. We can therefore safely assume that risk adjustment has served its purpose of preserving the variance in TIN costs, while controlling for beneficiary-episode covariates that are beyond the control of the TIN.

Table 3.5a: Standardized and Risk Adjusted Episode Costs and TIN's Standardized and Risk Adjusted Episode Costs for all selected MS-DRG Conditions from the Multi State Data

Condition	Number of Episodes	Number of TINs	Episode Length	Standardized Episode Cost ($): Mean (Std)	Risk Adjusted Episode Cost ($): Mean (Std)	TIN's Standardized Episode Cost ($): Mean (Std)	TIN's Risk Adjusted Episode Cost ($): Mean (Std)
Medical AMI	34,194	15,732	30-day	13,281 (17,449)	22 below expected (16,528)	3,257 (4,104)	84 (3,333)
Medical AMI	34,194	15,732	60-day	18,032 (21,824)	12 below expected (21,139)	4,450 (5,179)	132 (4,267)
AMI w CABG	2,559	4,463	30-day	16,000 (15,475)	83 below expected (13,299)	1,922 (2,638)	97 (1,720)
AMI w CABG	2,559	4,463	60-day	20,594 (21,978)	246 below expected (19,742)	2,479 (3,452)	100 (2,298)
AMI w PTCA	13,679	7,991	30-day	6,910 (10,520)	119 below expected (9,816)	1,670 (2,874)	148 (2,474)
AMI w PTCA	13,679	7,991	60-day	9,927 (15,052)	189 below expected (14,087)	2,374 (4,110)	193 (3,551)
CHF	107,185	21,120	30-day	8,974 (13,788)	67 below expected (13,313)	2,627 (3,547)	65 (3,064)
CHF	107,185	21,120	60-day	14,208 (19,034)	126 (17,989)	4,101 (4,875)	65 (3,883)
COPD	78,760	18,525	30-day	6,594 (9,564)	165 below expected (9,191)	2,270 (3,174)	13 (2,792)
COPD	78,760	18,525	60-day	10,654 (14,501)	342 below expected (13,831)	3,617 (4,548)	24 below expected (3,841)
Pneumonia	86,869	20,056	30-day	7,380 (10,558)	177 below expected (10,079)	2,470 (2,803)	20 below expected (2,284)
Pneumonia	86,869	20,056	60-day	11,119 (15,305)	307 below expected (14,436)	3,721 (4,183)	35 below expected (3,338)
Bronchitis	13,780	9,577	30-day	4,866 (7,739)	185 below expected (7,409)	1,984 (2,932)	56 below expected (2,610)
Bronchitis	13,780	9,577	60-day	7,582 (11,742)	390 below expected (11,463)	3,083 (4,550)	133 below expected (3,990)
Acute Ischemic Stroke	1,458	3,135	30-day	15,204 (15,160)	1,068 below expected (15,837)	3,059 (3,833)	147 below expected (3,344)
Acute Ischemic Stroke	1,458	3,135	60-day	19,920 (20,193)	1,230 below expected (20,617)	4,026 (5,199)	109 below expected (4,133)
Stroke w Cerebral Infarct	43,246	17,397	30-day	12,887 (13,491)	255 below expected (13,096)	3,298 (3,593)	20 below expected (2,937)
Stroke w Cerebral Infarct	43,246	17,397	60-day	17,404 (18,510)	242 below expected (17,491)	4,505 (5,028)	33 (4,029)

Table 3.5(b): Standardized and Risk Adjusted Episode Costs and TIN's Standardized and Risk Adjusted Episode Costs for all selected MS-DRG Conditions from the Multi State Data

Condition	Number of Episodes	Number # of TINs	Episode Length	Standardized Episode Cost ($): Mean (Std)	Risk Adjusted Episode Cost ($): Mean (Std)	TINs Standardized Episode Cost ($): Mean (Std)	TINs Risk Adjusted Episode Cost ($): Mean (Std)
Hip Replacement	24,603	10,083	30-day	10,513 (8,456)	101 (7,433)	2,615 (3,205)	73 (1,816)
Hip Replacement	24,603	10,083	60-day	12,395 (11,232)	162 below expected (9,655)	3,184 (4,049)	87 (2,432)
Knee Replacement	53,647	12,189	30-day	9,243 (7,244)	41 below expected (6,545)	2,231 (2,752)	159 (1,516)
Knee Replacement	53,647	12,189	60-day	10,738 (9,193)	56 below expected (8,232)	2,642 (3,326)	192 (1,896)
Hip Fracture	3,098	12,189	30-day	13,468 (13,468)	421 below expected (13,473)	4,213 (5,956)	137 (4,659)
Hip Fracture	3,098	12,189	60-day	18,078 (18,672)	517 below expected (17,616)	5,579 (7,712)	165 (6,005)
Laparoscopic Cholecystectomy	15,851	13,132	30-day	5,459 (9,183)	231 below expected (8,619)	1,291 (1,881)	32 (1,670)
Laparoscopic Cholecystectomy	15,851	13,132	60-day	7,591 (12,379)	471 below expected (11,983)	1,795 (2,515)	1 below expected (2,316)
Non-Laparoscopic Cholecystectomy	3,701	7,316	30-day	8,497 (12,242)	387 below expected (11,439)	1,791 (3,734)	22 (3,377)
Non-Laparoscopic Cholecystectomy	3,701	7,316	60-day	11,528 (17,034)	708 below expected (16,068)	2,410 (4,748)	14 below expected (4,180)
Medical Back Pain	9,904	9,558	30-day	10,032 (11,210)	51 below expected (10,831)	3,328 (4,380)	74 (3,635)
Medical Back Pain	9,904	9,558	60-day	14,713 (16,292)	95 below expected (15,490)	4,923 (6,302)	130 (5,152)
Back Pain w Spinal Fusion	6,332	4,390	30-day	12,463 (11,659)	36 below expected (11,052)	2,463 (4,385)	136 (2,690)
Back Pain w Spinal Fusion	6,332	4,390	60-day	14,452 (14,594)	60 below expected (13,658)	2,917 (5,387)	184 (3,564)
Back Pain w Other Back Procedure	9,025	7,932	30-day	5,963 (8,454)	-83 below expected (7,932)	1,794 (2,976)	234 (2,348)
Back Pain w Other Back Procedure	9,025	7,932	60-day	2,976 (11,385)	157 below expected (10,854)	2,317 (3,863)	287 (3,099)

Table 3.6(a): Episodes per TIN and Total Episode Proportion per TIN and Average Episode Proportion per TIN for all selected MS-DRG Conditions from the Multi State Data

Condition	Number of Episodes	Number of TINs	Column A Episodes per TIN[13]: Mean (Std)	Column B Total Episode Proportion Per TIN[14]: Mean (Std)	Column C Average Episode Proportion Per TIN[15]: Mean (Std)
Medical AMI	34,194	15,732	8.8 (20.1)	2.2 (6.6)	0.24 (0.29)
AMI w CABG	2,559	4,463	3.6 (5.1)	0.5 (1.7)	0.11 (0.13)
AMI w PTCA	13,679	7,991	5.7 (12.1)	1.7 (6.8)	0.19 (0.21)
CHF	107,185	21,120	19.0 (50.0)	5.1 (15.1)	0.28 (0.21)
COPD	78,760	18,525	14.0 (35.0)	4.2 (11.9)	0.32 (0.24)
Pneumonia	86,869	20,056	14.0 (36.0)	4.3 (13.0)	0.32 (0.24)
Bronchitis	13,780	9,577	4.0 (8.0)	1.4 (2.8)	0.39 (0.29)
Acute Ischemic Stroke	1,458	3,135	2.2 (2.6)	0.5 (1.0)	0.19 (0.17)
Stroke with Cerebral Infarct	43,246	17,397	10.2 (25.4)	2.5 (7.7)	0.24 (0.19)
Hip Replacement	24,603	10,083	9.0 (25.0)	2.4 (10.6)	0.20 (0.23)
Knee Replacement	53,647	12,189	16.0 (46.0)	4.4 (19.7)	0.20 (0.23)
Hip Fracture	3,098	12,189	2.4 (7.9)	0.7 (3.9)	0.28 (0.23)
Laparoscopic Cholecystectomy	15,851	13,132	6.1 (13.1)	1.2 (2.9)	0.20 (0.18)
Non-Laparoscopic Cholecystectomy	3,701	7,316	2.7 (3.8)	0.5 (1.1)	0.19 (0.20)
Medical Back Pain	9,904	9,558	3.3 (6.0)	1.0 (2.3)	0.32 (0.24)
Back Pain w Spinal Fusion	6,332	4,390	5.4 (10.4)	1.4 (5.0)	0.18 (0.27)
Back Pain w Other Back Procedure	9,025	7,932	5.9 (12.3)	1.9 (6.0)	0.23 (0.26)

[13] n_j is the number of episodes of that condition attributed to the TIN.

[14] Episode proportion per TIN is computed as the sum of all the p_{ij} , or $\sum_{i=1}^{n_j} p_{ij}$

[15] p_{ij} is the proportion of the beneficiary's episode attributed to the TIN.

D. Problem with Low Ns (Number of Episodes) and Low Ps (Proportion of Episode)

An entirely different problem is raised by an issue noted only in passing to this point: the challenge to any performance measurement system when physicians have few episodes or very low proportions of episodes. For all the selected conditions from the multi-state data, Table 3.6(a) above. shows:

- the mean and standard deviation of total episodes attributed per TIN (column A)
- total episode proportion per TIN (full time episodes or FTEs), in column B, which is the sum of the average episode proportion attributed to the TINs (column C)
- There is large standard deviation from the mean, implying that the distribution of episodes per TIN as well as FTEs per TIN is highly skewed – most TINs have zero or few episodes, and only a few TINs have very many. Figure 3.3(a) below, is a bar graph showing the distribution of the number of per TIN for Medical AMI. Almost 70 percent of TINs have 0 - 6 episodes attributed to them. In data not shown:
- The median number of episodes per TIN for Medical AMI is 3 (mean: 8.8). Thirty one percent of Medical AMI TINs have only one episode attributed to them (data not shown).
- The median number of episodes per TIN is skewed even more for other conditions. For example, it is 2 for PTCA (mean: 5.7) and 2 for CABG (mean: 3.6).

These results point to a problem of low Ns, as discussed below.

Figure 3.3a: Distribution of Total Episodes per TIN for Medical AMI

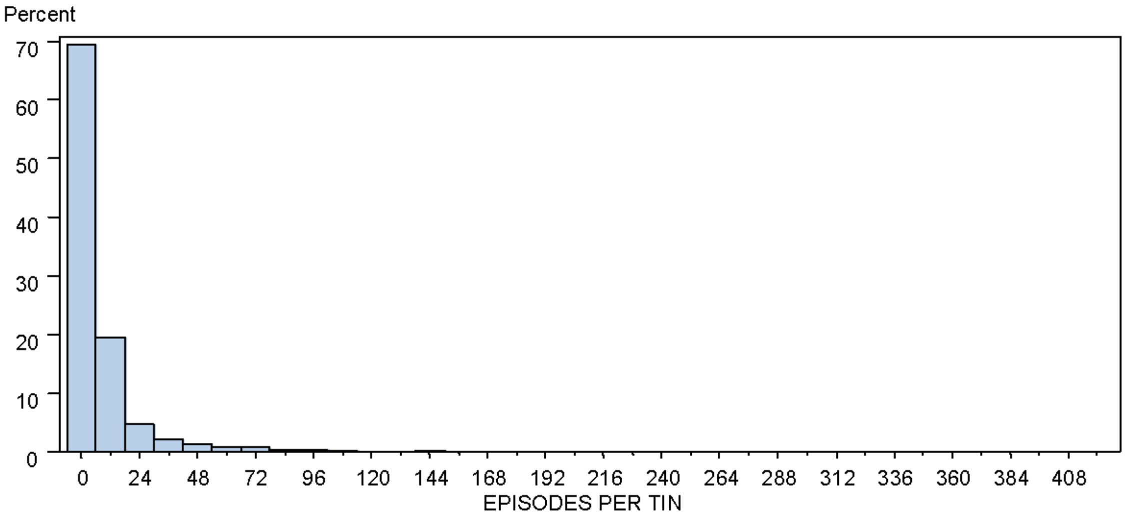


Figure 3.3.b: Distribution of Total Episode Proportion per TIN for Medical AMI TINs

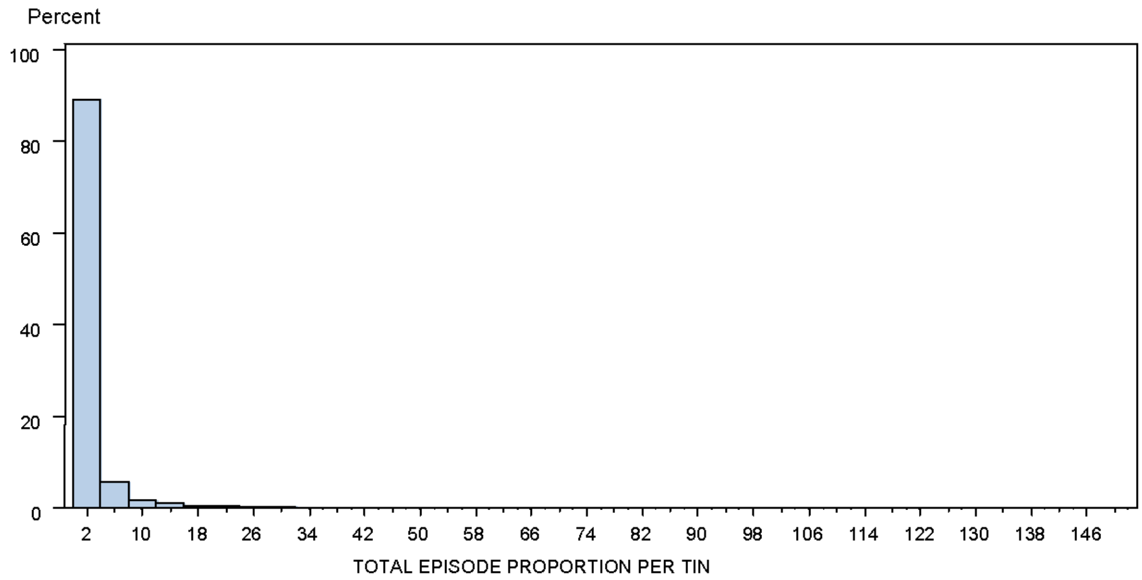


Figure 3.3(b) above is also a bar graph, in this case showing the distribution of total episode *proportion* per TIN for AMI. Almost 90 percent of TINs have 2 or full time episode (FTE) equivalents attributed to them, and (data not shown) 65 percent of AMI TINs have less than 1 FTE of AMI attributed to them. These results mean that the care for AMI episodes provided by these TINs (in terms of proportion of episodes) is less than the care provided one full episode of AMI. In data not shown, 89 percent of CABG TINs and 82 percent of PTCA TINs have less than one FTE attributed to them, meaning that more TINs are involved in sharing the responsibility of an episode for these conditions.

Figure 3.3 (c) below is a bar graph showing the distribution of average episode proportion per TIN for medical AMI. Table 3.6(b) below shows the distribution of AMI TINs across different levels of average episode proportion . 26 percent of AMI TINs are responsible for less than 10 percent of the proportion of AMI episodes. 51 percent of medical AMI TINs, 72 percent of PTCA TINs and 88 percent of CABG TINs are responsible for less than 20 percent of episodes. Only 5 percent of medical AMI TINs, 8 percent of PTCA TINs and 1.2 percent of CABG TINs are responsible for more than 60 percent of episode proportion for the said conditions. This is the problem of low Ps.

Figure 3.3.c: Distribution of Average Episode Proportion per TIN for Medical AMI TINs

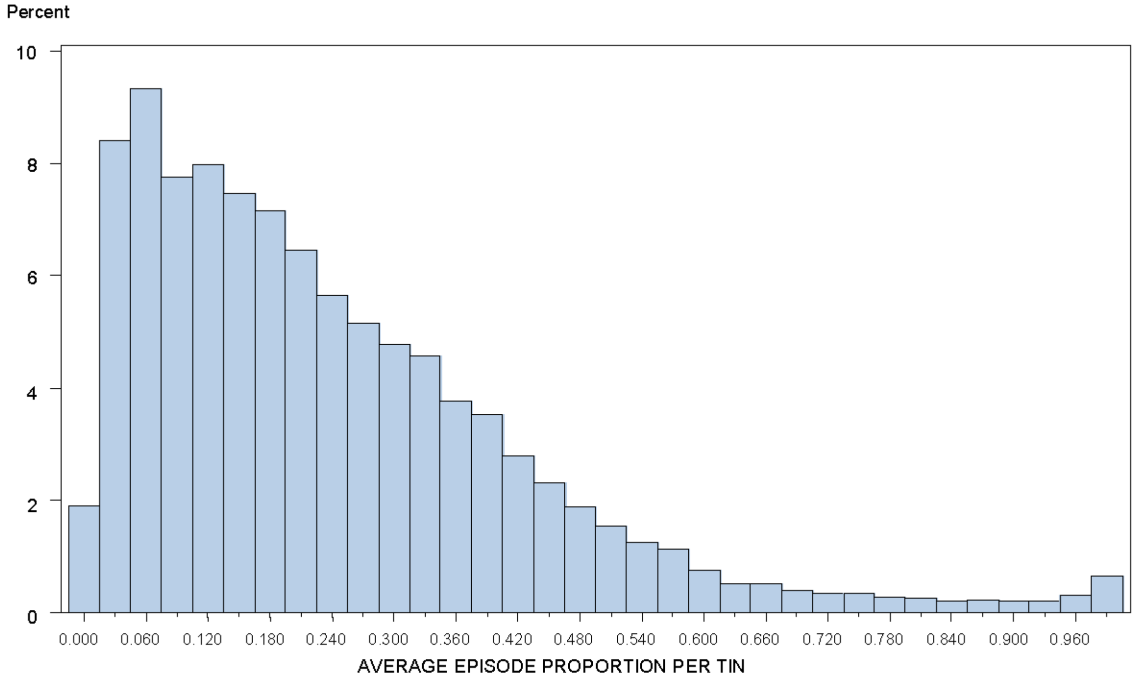


Table 3.6 (b): Distribution of AMI TINs across Levels of Mean Episode Proportion

TIN's Mean Episode Proportion	Medical AMI TINs (N=15,732)	AMI with PTCA TINs (N=7,991)	AMI with CABG (N=4,463)
0%-20%	51%	72%	88%
20%-40%	32%	15%	8%
40%-60%	12%	5%	3%
60%-80%	3%	5%	0.5%
80%-100%	2%	3%	0.7%

To examine the nature of TINs with low episode volume, low episode proportion and low Full Time Episodes, we prepared a TIN master-file from Part B claims detailing the TIN type (single/multi specialty), TIN practice area (Rural/Non-rural), number of NPIs within each TIN, and their sub-specialties. The TIN master-file was linked to the TIN episode profiles and we ran a series of logistic regressions to study the characteristics of TINs with (i) Low episode volume (n=1), (ii) Low episode proportion (p ≤10 percent) and (iii) Low Full Time Episodes (FTE≤1).

The results are summarized in Table 1 in Appendix 5. TINs with low episode volume/proportion/FTE for AMI were most likely to be single specialty TINs, smaller physician groups, rurally located (with the exception of medical AMI), *not* have cardiologists, internists, critical medical, or cardiac surgeons. TINs with surgeons, surgical specialists and anesthesiologists were likely to be associated with higher episode volumes/proportions/ FTEs for CABG, but this association changed direction for PTCA. TINs with radiologists and anesthesiologists were likely to be associated with higher episode volumes but lower episode proportions and lower FTEs.

So, while there are variations across TIN types, the problems of low proportion and low FTEs remain. The next section explores one possible solution: pooling across conditions which, as we shall discuss, raises problems of its own.

E. Pooling Across Conditions

One way to address the low episode volume/episode proportion problem at the TIN level is to pool TIN risk-adjusted episode costs across different conditions. However, such pooling creates a potentially misleading case mix. Tables 2 and 3 in Appendix 5 show the results of pooling TIN risk adjusted cost profiles for medical AMI, CABG and PTCA into a single category: pooled AMI. The total number of TINs for AMI is 17,538 instead of 28,486, as the three conditions share 38 percent of common TINs. Pooling thereby increases the episode volume, episode proportion and full time episodes across TINs.

After pooling, 68 percent of the TINs have less than one FTE. TINs that are non-rural, and TIN's with anesthesiologists, radiologists, and general surgeons are associated with lower FTEs. Pooling marginally increases the number of FTEs for TINs. It also increases the episode volume for TINs with low episode proportions. However for TIN's with anesthesiologists, radiologists and general surgeons- pooling does not appreciably increase the total episode proportion attributed to them. Pooling also does not change the average episode proportion per TIN by much.

A solution to increase full time episode volumes for TINs with low FTEs is to pool TIN profiles across all of the 11 conditions selected for the MS-DRG approach, but this would create an average value that would be hard to interpret and open to criticism about case mix. Profiles for specific conditions are more actionable for quality improvement among specialist providers.

Pooling profiles across multiple conditions may allow for better profiling of certain types of physicians: possibly, internists/general surgeons/radiologists/anesthesiologists. But even with these physician types for whom pooling across multiple conditions makes more conceptual sense, pooling creates at the same time an additional challenge of combining quality measures that are unique to particular clinical conditions, often resulting in missing values for conditions where costs and quality may not be observed for some TINs.

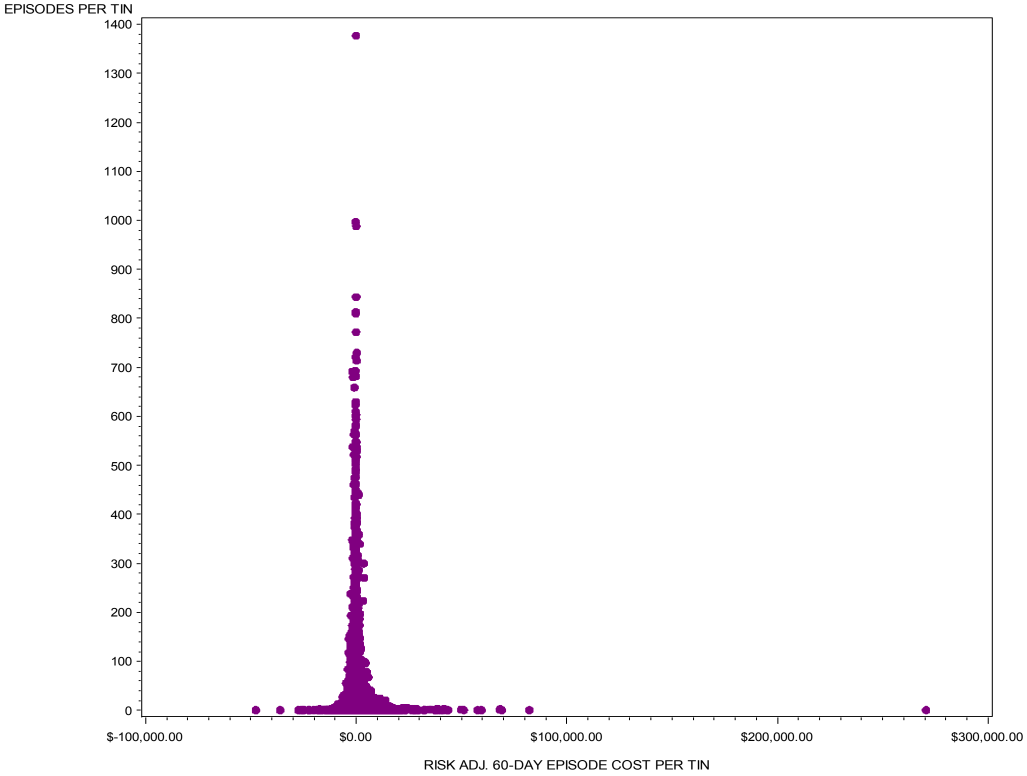
The implications of this problem can best be understood by reiterating how the multiple attribution approach discussed here attributes responsibility. Specifically, this approach holds physician groups providing care during the initial hospitalization responsible for the treatment choices made by them and the consequential costs of care. For instance, if a cardiologist attempts a failed PTCA on a patient with AMI who subsequently receives a CABG, the cardiologist's TIN is responsible for the Part B cost associated with the failed PTCA and a proportional amount of post-discharge costs (the cardiac surgeon's TIN is absolved of the costs associated with the cardiologist's choice). The approach makes sound sense for hospital based physician-group payment policy.

However one of the challenges faced by this approach is that of profiling TINs with low episode volume/episode proportion, as discussed above. While pooling of TIN profiles across different conditions can solve the low P/low N problem, we choose to avoid pooling, because the benefits are more than offset by the artificial aggregation created by combining very different diseases and treatments. Leaving the conditions intact and separate allows users to provide their own weights for any subsequent pooling they wish to do.

F. Determining TIN Peer Group for Cost Profiles

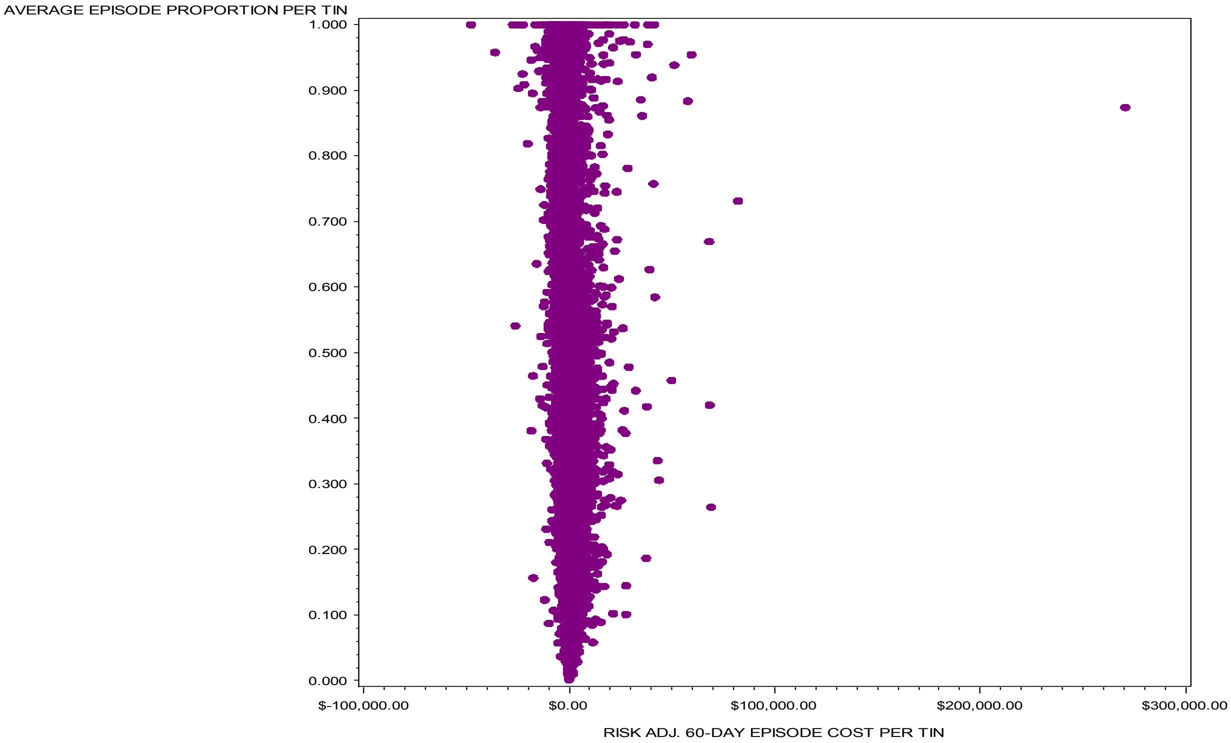
The high number of TINs with low episode volume and low total episode proportions dissuades us from using episode volume or total episode proportion as the determinant of peer groups for TINs. Figure 3.4 (a) is a scatterplot showing the association between TIN's risk adjusted 60-day CHF episode costs and the number of CHF episodes attributed to the TIN. We can conclude that most of the variation in TIN risk adjusted costs occurs among low episode volume TINs – note in the figure how almost all of the spread in cost occurs near zero episodes per TIN. There much lesser variation among high episode volume/proportion TINs simply because there are fewer such TINs.

Figure 3.4 (a): Association between TIN's Average Risk Adjusted 60-day CHF Episode Costs and Number of Episodes per TIN for CHF TINs



The mean episode proportion for a TIN, which is simply the TIN's total episode proportion divided by the number of episodes allows a more consistent measurement of variation in risk adjusted episode costs for TINs over its entire range from 0 to 1. TINs' mean episode proportion for an MS-DRG condition depends on the TIN specialty- TINs with cardiologist are more likely to have higher proportion of AMI episodes attributed to them on an average, while generalists are likely to have a lower proportion of AMI episodes attributed to them on an average. Hence *the mean episode proportion attributed to TINs serves as the best determinant of TIN peer groups*. 3.4 (b) is a scatterplot showing that the association between the mean episode proportion attributed to TINs and TINs' residual episode costs is consistent across all levels of the TIN's mean episode proportion – note how tightly bunched the cost plot is across all levels of episode proportion. This has an important implication for our measurement system: *TINs can hence be* *grouped into quartiles or quintiles on the basis of their mean episode proportion. Costs and quality for TINs then can be compared with those in their peer groups.*

Figure 3.4 (b): Association between TIN's Average Risk Adjusted 60-day CHF Episode Costs and Average Episode Proportion per TIN for CHF TINs



It will be useful to consider how these relationships vary across the individual states in our multi-state data. If the relationships vary by state, we might need to take state fixed effects into account in any scoring system. Table 3.7 (a) below, shows the variation in CHF episodes, TINs, standardized and risk adjusted 60-day CHF episode costs attributed to TINs, across states in the multi-state sample. New Jersey and Florida have the highest standardized 60-day CHF episode costs among states, but the lowest standardized average 60-day episode costs per TIN. In these states more TINs manage an average episode of CHF. Minnesota and Kansas have the lowest standardized 60-day CHF episode costs among states, but the highest standardized average 60-day episode costs per TIN as fewer TINs in these states TINs manage an average episode of CHF.

Table 3.7a: State Variation in CHF Episodes, TINs, Standardized 60-day CHF Episode Costs, Standardized 60-day CHF Episode Costs per TIN and Risk Adjusted 60-day CHF Episode Costs per TIN

State	CHF: Number of Episodes	CHF: percent Episodes	CHF: Number of TINs	CHF: percent TINs	Standardized 60-day episode cost: Mean (Std)	Standardized 60 day episode cost per TIN: Mean (Std)	Risk adjusted 60-day episode per TIN: Mean (Std)
California	25,792	24 percent	10,508	33 percent	$14,475 ($20,519)	$4,187 ($4,925)	-1 (4,052)
Florida	26,519	25 percent	8,754	28 percent	$15,048 ($20,287)	$3,874 ($5,594)	27 (4,640)
Kansas	4,557	4 percent	808	3 percent	$12,148 ($15,337)	$5,642 ($5,707)	128 (4,203)
Louisiana	9,534	9 percent	1,760	6 percent	$14,103 ($18,345)	$4,524 ($4,502)	203 below expected (3,239)
Minnesota	4,853	5 percent	446	1 percent	$11,728 ($16,851)	$5,186 ($4,669)	-487 (3,013)
New Jersey	18,221	17 percent	6,122	19 percent	$16,055 ($19,032)	$3,804 ($3,450)	287 (2,497)
Virginia	731	1 percent	1,950	6 percent	$12,654 ($19,266)	$3,814 ($4,202)	352 below expected (2,903)
Vermont	11,314	11 percent	206	1 percent	$12,474 ($16,717)	$4,487 ($5,130)	48 (4,521)
Washington	6,087	6 percent	910	3 percent	$10,976 ($14,562)	$4,490 ($5,459)	533 below expected (4,624)

Table 3.7 (b): State Variation in Average CHF Episode Proportion per TIN and TIN Distribution across Average CHF Episode Proportion Peer Groups

State	AVG. Episode Prop. Per TIN: Mean (Std)	Percent of State's TINs in Peer Group 1 (Mean Episode Prop.: 0-0.20)	Percent of State's TINs in Peer Group 2 (Mean Episode Prop. : 0.20-0.40)	Percent of State's TINs in Peer Group 3 (Mean Episode Prop.: 0.40-0.60)	Percent of State's TINs in Peer Group 4 (Mean Episode Prop.:0.60-0.80)	Percent of State's TINs in Peer Group 5 (Mean Episode Prop.:0.80-1)
California	0.29 (0.2)	40 percent	35 percent	16 percent	6 percent	3 percent
Florida	0.25 (0.17)	46 percent	38 percent	13 percent	2 percent	1 percent
Kansas	0.44 (0.31)	32 percent	23 percent	13 percent	12 percent	20 percent
Louisiana	0.32 (0.24)	40 percent	31 percent	16 percent	8 percent	6 percent
Minnesota	0.48 (0.35)	35 percent	11 percent	13 percent	13 percent	28 percent
New Jersey	0.24 (0.16)	49 percent	36 percent	13 percent	2 percent	1 percent
Virginia	0.29 (0.21)	42 percent	30 percent	18 percent	7 percent	2 percent
Vermont	0.34 (0.24)	39 percent	22 percent	25 percent	8 percent	5 percent
Washington	0.37 (0.28)	37 percent	24 percent	14 percent	16 percent	10 percent
Pooled	0.28 (0.21)	43 percent	34 percent	15 percent	5 percent	3 percent

Table 3.7(b) above, shows the state variation in average episode proportion per TIN, our variable of choice for peer grouping TINs. We group TINs into five episode proportion peer groups, wherein TINs with mean episode proportions of 0-0.20, 0.20-0.40, 0.40-0.60, 0.60-0.80, and 0.80-1 grouped together. The table shows the percent of CHF TINs in every state in each of the episode proportion peer groups. The number of TINs across the peer groups varies by states. All states have more than a third of their CHF TINs in lowest episode proportion peer group. TINs in this peer group have 0%-20% of the episode attributed to them. Kansas and Minnesota have about a third of their CHF TINs in this peer group, while Florida and New Jersey have almost half their CHF TINs in this group. Florida and New Jersey also have only 1 percent of their CHF TINs in the highest episode proportion group; while Minnesota and Kansas have 20 percent and 28 percent of CHF TINs in this peer group respectively. The way TINs are integrated and deliver care within hospitals hence varies considerably by state.

We did not adjust for state fixed effects during risk adjustment but adjusted for Medicare variation in state payment by cost standardization, because we wanted to capture the variation in physician practice patterns across states. However, it could be argued that some of the state variation in costs comes from factors exogenous to the physician – like rurality (we observe fewer TINs managing an episode of CHF in rural states).

In summary, for each selected MS-DRG condition, we can group TINs into peer-groups based on their mean episode proportion, and compare the performance of TINs within these peer-groups. The mean episode proportion peer-group of a TIN reflects the TIN's level of responsibility for the episode, and is more or less determined by TIN's specialty, size and location. We discuss more implications of peer grouping TINs thus in Chapter V.

IV. Quality Measures

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

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

2. Is the measure *actionable* by the individual physician? 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. Is the measure *meaningful* to physicians? 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, and more specifically, what weights to give to each measure. Weighting all measures equally does not mitigate this dilemma. It simply assigns a weight of "1.0" to each, 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 issues or constructs.

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. The results are accordingly more likely to be more understood and respected by clinicians. Organized judgment also can provide legitimacy to the relative weightings, as it rests weightings on the professional judgment of peers of the physicians to be governed by the weights. Given these advantages, we chose the second approach and convened an MS-DRG TEP.

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.

A. The Selection of Candidate Quality Measures

Before convening the MS-DRG TEP, we did extensive work to select quality measures attributable to hospital based TINs for the selected MS-DRG conditions. 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 the Beneficiary Approach as discussed in Chapter 4, well as the MS-DRG 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:

- There were seven global measures in MS-DRG analysis, applicable to all selected MS-DRGs, including:
    - five episode-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:
        - Medicare Hospital Compare reported by CMS(see Appendix 5, Tables 5.1 (a), (b), (c) and (d) for measures)[16]
        - Hospital-level patient safety indicators (PSIs) reported by the Agency for Healthcare Research and Quality (AHRQ) (See Appendix 5, Tables 5.2 for measures)[17], and
- There was a second group of specific measures 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) The measures in Medicare Hospital Compare process measure sets for acute myocardial infarction, congestive heart failure, pneumonia, and surgical care improvement are listed in Appendix 5, Tables 5.1 (a), (b), (c) and (d), respectively. The measures in AHRQ patient safety indicators are listed in Appendix 5, Table 5.2.

Some of these measures (e.g., avoidable ED use and potentially preventable hospital 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.

B. Review by Expert Panel

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 hospital 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. The final composition of the MS-DRG expert panel was as follows:

[16] 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.

[17] 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.

The MS-DRG Technical Expert Panel

Panel Member	Affiliation
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
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
Christopher Bever, MD	American Academy of Neurology, American Neurological Association, Associate Chief of Staff for Research and Development at the VA Medical Center Baltimore
Robert Bonow, MD	American Heart Association, American College of Cardiology, Chief of the Division of Cardiology at Northwestern Memorial Hospital
David Jevsevar, MD, MBA	American Academy of Orthopedic Surgeons
Martha Radford, MD	Chief Quality Officer, New York University, Langone Medical Center
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

C. Rating Process and Results

- <u>TEP Panel Meetings</u> Members of the MS-DRG panel met with the project team and project officers from CMS in Minneapolis on April 26 2010. The meeting was moderated by UMN project leaders. Two project officials from CMS participated in both meetings.

- The purpose this meeting was to introduce TEP members to the details of the MS-DRG approaches, to discuss with them the quality measures the approach 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.

- <u>Rating process</u> –After the TEP meetings, with suitable preparation, TEP members then were asked to rate the usefulness of potential quality measures for the MS-DRG approach. 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 the MS-DRG ratings of the usefulness of measures, there was a mixture of results with regard to the seven overall constructs. The summary of weights from the TEP rating exercise are presented in Appendix 5, Table 5.3. 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).[18]

The TEP process did a good job of winnowing our candidate quality measures, the primary purpose for which it was used.

[18] Note that, overall, the ratings of quality measures for the primary care Beneficiary approach were higher than those for the MS-DRG approach (the results for the Beneficiary approach are presented in Chapter 4). This difference seems reasonable given that the MS-DRG 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.

Following our review of literature and the results from our TEP discussion, we were prepared to select measures to be employed in data analyses for profiling hospital TINs. We selected *representatives* of two types of quality measures for profiling hospital TINs: episode level quality measures and hospital level quality measures.

All of our episode level measures are computable from Medicare claims data. With the exception of potentially preventable rehospitalizations, all can be computed using free, publicly available software. All of our hospital level quality measures are available from the Medicare Hospital Compare dataset. The representative episode level and hospital level quality measures that we selected are:

1. Episode Level Quality Measures:
   a. All cause case-mix adjusted mortality rate. This measure is applicable to all MS-DRGs.
   b. Case mix- adjusted potentially preventable rehospitalization per episode at end of 30-day and 60- day episode window (Goldfeld, et al., 2007[19]). This measure is applicable to all MS-DRGs.
   c. Avoidable emergency department (ED) visits per episode computed at the at end of 30-day and 60- day episode window (Billings, Parikh and Mijanovich, 2000[20]). This measure is applicable to all MS-DRGs.
2. Hospital Level Quality Measures:
   a. Hospital Level Process Measures- Medicare Hospital Compare Measures
      - Acute Myocardial Infarction Measure Set (Appendix 5,Table 5.1 (a))
      - Congestive Heart Failure Measure Set (Appendix 5, Table 5.1 (b))
      - Pneumonia Measure Set (Appendix 5, Table 5.1 (c))
      - Surgical Care Improvement Measure Set (Appendix 5, Table 5.1 (d)) The measures for AMI, CHF and Pneumonia are applicable to MS-DRGs for those respective conditions. The surgical care measures are applicable to all surgical MS-DRGs.
   b. Hospital Level Outcome Measures- AHRQ Patient Safety Indicators (PSI): These measures are listed in Appendix 5, Table 5.2. We selected those measures rated >60 by the TEP for usefulness and available on Hospital Compare, viz.:
      - PSI 5: Foreign body left during procedure
      - PSI 6: Iatrogenic Pneumothorax
      - PSI 7: CV Catheter related blood infections
      - PSI 11: Postoperative respiratory failure
      - PSI 12: Postoperative sepsis
      - PSI 14: Postoperative wound dehiscence
      - PSI 15: Accidental puncture or laceration

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

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

PSI 6 and 7 are applicable to all MS-DRGs while the other PSIs are applicable to all surgical MS-DRGs.

Table 3.8 (a1) and Table 3.8 (b) identify all the quality measures applicable from above for the selected medical and surgical MS-DRGs respectively. All medical and surgical MS-DRG episodes are denominators for the three episode-level measures. The surgical care process measure set from Hospital Compare is appropriate for all surgical MS-DRGs. AMI, Pneumonia and CHF measure sets are applicable to the MS-DRGs for the respective conditions. Among the hospital level PSI outcome measures, iatrogenic pneumothorax is applicable for all MS-DRGs except AMI with CABG and AMI with PTCA; catheter related blood infections are not applicable to pneumonia, while the other PSIs are appropriate for all surgical MS-DRGs.

Table 3.8 (a1): Quality measures for medical MS-DRG conditions (1) Episode Level Measures

Condition	Risk Adj. Mortality 30&60 day	Risk Adj. PPR 30&60 day	Avoidable ED Visits 30&60 day
Medical AMI	X	X	X
CHF	X	X	X
Pneumonia	X	X	X
COPD	X	X	X
Bronchitis	X	X	X
Acute Ischemic Stroke	X	X	X
Stroke with Cerebral Infarct	X	X	X
Medical Back Pain	X	X	X

Table 3.8 (a2) Hospital compare : Hospital process measures

Condition	AMI Measure Set	CHF Measure Set	Pneumonia Measure Set	Surgical Care Improvement Measure Set
Medical AMI	X			
CHF		X		
Pneumonia			X	
COPD				
Bronchitis				
Acute Ischemic Stroke				
Stroke with Cerebral Infarct				
Medical Back Pain				

3.8(a3) AHRQ Patient safety indicators-hospital level outcome measures

Condition	Foreign body left during proc.	Iatrogenic Pneumothorax	CV Cath. related blood inf.	Postop. resp. failure	Postop. sepsis	Postop. wound dehiscence	Accidental puncture/ laceration
Medical AMI		X	X				
CHF		X	X				
Pneumonia		X					
COPD		X	X				
Bronchitis		X	X				
Acute Ischemic Stroke		X	X				
Stroke with Cerebral Infarct		X	X				
Medical Back Pain		X	X				

Table 3.8 (b1): Quality measures for surgical MS-DRG conditions (1) episode level measures

Condition	Risk Adj. Mortality 30&60 day	Risk Adj. PPR 30&60 day	Avoidable ED Visits 30&60 day
AMI with CABG	X	X	X
AMI with PTCA	X	X	X
Hip Repl.	X	X	X
Knee Repl.	X	X	X
Hip Frac.	X	X	X
Lap. Chole.	X	X	X
Non-Lap. Chole	X	X	X
Back Pain w Spi. Fu.	X	X	X
Back Pain w Other Back Proc	X	X	X

Table 3.8(b2) Hospital compare: Hospital process measures

Condition	AMI Measure Set	CHF Measure Set	Pneumonia Measure Set	Surgical Care Improvement Measure Set
AMI with CABG	X			X
AMI with PTCA	X			X
Hip Repl.				X
Knee Repl.				X
Hip Frac.				X
Lap. Chole.				X
Non-Lap. Chole				X
Back Pain w Spi. Fu.				X
Back Pain w Other Back Proc				X

3.8 (b3) AHRQ Patient Safety Indicators-Hospital Level Outcome Measures

Condition	Foreign body left during proc.	Iatrogenic Pneumothorax	CV Cath. related blood inf.	Postop. resp. failure	Postop. sepsis	Postop. wound dehiscence	Accidental puncture/ laceration
AMI with CABG	X		X	X	X	X	X
AMI with PTCA	X		X	X	X	X	X
Hip Repl.	X	X	X	X	X	X	X
Knee Repl.	X	X	X	X	X	X	X
Hip Fracture	X	X	X	X	X	X	X
Lap. Chole.	X	X	X	X	X	X	X
Non-Lap. Chole	X	X	X	X	X	X	X
Back Pain with Spi. Fu.	X	X	X	X	X	X	X
Back Pain with Other Back Proc	X	X	X	X	X	X	X

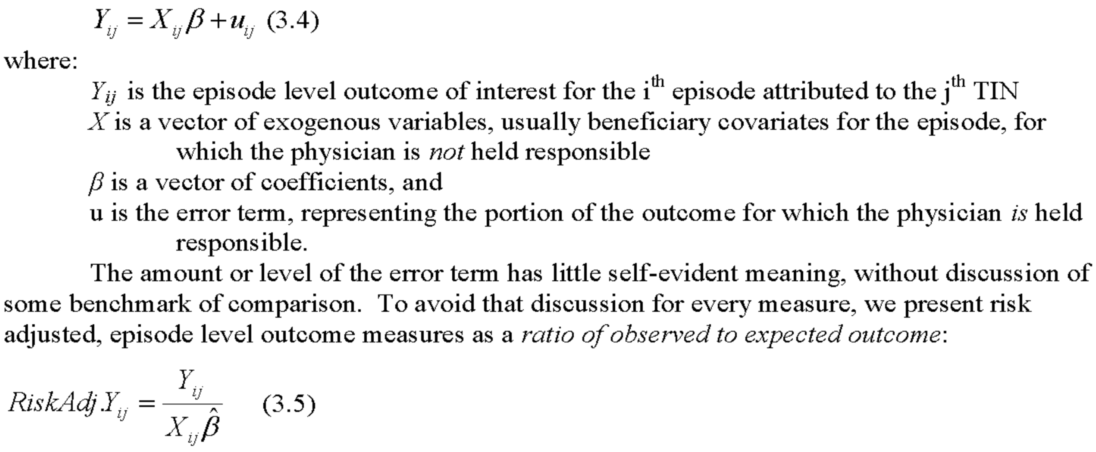
D. Attribution, Framing and Risk Adjustment of Quality Measures

As described in Section II of this chapter, we used a multiple attribution approach for attributing *costs* of care for an MS-DRG episode based on each TIN's proportion of the Part B billing for the episode.

For attributing *quality* for the care provided during an MS-DRG episode, we use a multiple attribution approach based on an all-or-none rule. If a TIN was responsible for *any* care during the index hospitalization associated with an MS-DRG episode, *all* the quality measures associated with that episode –during the 30- and 60-day windows are attributed to the TIN. Analogously, keeping with the recommendations of the TEP, all hospital level process and outcome measures associated with the episode are attributed to the TIN. The purpose of such an attribution rule for quality is to motivate all physician entities involved in the care of the beneficiary to better coordinate care during the beneficiary's hospitalization and post-discharge.

Given Medicare's goal of identifying best performers and rewarding them for value, we deem it important to present the quality measures in a positive frame (Tversky and Kahneman, 1981[21]). For instance, we can report mortality at the TIN level as 30- or 60- day likelihood of survival for that TIN, or readmissions as risk-adjusted 30- or 60-day PPR-free likelihood for the TIN. Reporting in this way helps us to easily identify TINs that are better performers. Providers are more comfortable with positively framed quality measures as it highlights their strengths, and are more willing to accept quality reporting when framed positively (AHRQ[22]).

As in the case of costs, we risk-adjust quality measures by regressing the episode-level outcome of interest on exogenous episode level covariates that are beyond the physician's control.



$$Y_{ij} = X_{ij}\beta + u_{ij} \quad (3.4)$$

where:

Y_{ij} is the episode level outcome of interest for the i$^{\text{th}}$ episode attributed to the j$^{\text{th}}$ TIN

X is a vector of exogenous variables, usually beneficiary covariates for the episode, for which the physician is *not* held responsible

β is a vector of coefficients, and

u is the error term, representing the portion of the outcome for which the physician *is* held responsible.

The amount or level of the error term has little self-evident meaning, without discussion of some benchmark of comparison. To avoid that discussion for every measure, we present risk adjusted, episode level outcome measures as a *ratio of observed to expected outcome*:

$$RiskAdj.Y_{ij} = \frac{Y_{ij}}{X_{ij}\hat{\beta}} \quad (3.5)$$

The risk adjusted value of the outcome Y is expressed in equation 3.5 as a percent of the expected value of Y, given the beneficiary's values of Xs for the episode. For instance if a TIN's $RiskAdj.Y$ =1.5, it means that the Y for the TIN was 50 percent higher than expected, after controlling for beneficiary characteristics. Clearly, presenting risk adjusted quality measures as a

[21] Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science 1981 Jan 30;211(4481):453-8.

[22] Agency for Healthcare Research and Quality. "Talking Quality: Framing Scores as Positive or Negative" available at https://www.talkingquality.ahrq.gov/content/create/scores/framing.aspx#_ftnref3

ratio rather than as residuals makes them more easily interpretable to physicians and for that reason more actionable.

Normally, in practice, while aggregating Y_{ij} from an episode level to a TIN level, Y_j is subjected to further adjustment using shrinkage estimators (or random effects). The shrinkage estimator 'corrects' for small sample sizes at the provider level, which often result in large standard errors for the estimate Y_j , by producing a weighted average estimate of Y_j by combining it with the mean $\bar{Y}$ across all the providers in the sample. More recently, the use of shrinkage estimators has come under criticism in the health services literature, for incorrectly moving the estimates for low volume providers to a sample mean – which in itself can never be truly known (Silber, et al., 2010[23]). In our analysis we avoid using shrinkage estimators while aggregating risk adjusted quality measures from the episode to the provider level, even though we present them as ratios for easier interpretation and making them more actionable to providers.

A. Results for risk-adjusted quality measures

The specific quality measures, their framing, risk adjustment and descriptive statistics are presented below.

i. Episode Level Measures

a. Risk-Adjusted 30 and 60 day Survival Likelihood

Our primary outcome for each of the selected conditions was 30-day and 60-day all-cause mortality, following the index hospitalization for an episode. Claims from the index hospitalization and 12 months prior were used to obtain the diagnoses that were classified into HCCs based on the CMS HCC model. We obtained demographic variables from the beneficiary summary file. We ran the analysis by beneficiary episode. The majority of the variables in our risk adjustment model are used in CMS's validated model for 30-day all-cause hospital-specific mortality measures. The covariates for our risk adjustment model included:

a) 70 HCC diagnostic categories.

b) Beneficiary's age and sex

c) Medicaid status

d) Reason for Medicare eligibility

e) Dummy variables identifying whether the index hospitalization DRG had major complications/comorbidities (MCC), complications/comorbidities (CC) or no comorbidities

[23] Silber, Jeffrey H., Rosenbaum, Paul R., Brachet, Tanguy J., Ross, Richard N., Bressler, Laura J., Even-Shoshan, Orit, Lorch, Scott A. and Kevin G. Volpp. "The Hospital Compare Mortality Model and the Volume-Outcome Relationship," *Health Services Research* 45:5 Part 1 (2010) 1148-1167.

f) A severity variable which identified the counts of prior hospitalizations for the same condition a year prior to the beneficiary-episode

g) Another severity variable which identified the number of prior emergency department visits (not resulting in hospitalization) for the same condition a year prior to the beneficiary-episode

We modeled the probability of *not dying or survival* in the 30- and 60- day period using logistic regression at the beneficiary episode level. For beneficiary episodes where the beneficiary survived at the end the episode window- we calculated the ratio of the observed to expected probability of survival. If a beneficiary's expected probability of survival is 0.5, the ratio of observed to expected probability of survival is 2, meaning that the beneficiary's probability of survival is twice what is expected given the beneficiary's covariates. This ratio shows how successful a TIN is at managing the episode of a morbid beneficiary (whose expected probability of survival is less than 1). We set the maximum for the ratio of observed to expected survival for a beneficiary episode to 2, so as to not give undue credit to TINs for treating really sick beneficiaries (<5% of the beneficiary episodes had the ratio of observed to expected survival probability greater than 2). We then subtracted 1 from the ratio so as to allow it to be interpreted as percentage risk-adjusted survival likelihood. If a beneficiary had an observed to expected ratio of 2, the risk adjusted survival likelihood was 100%. If the beneficiary did not survive at the end of the beneficiary episode- the risk-adjusted survival likelihood was set to 0. The TINs Risk adjusted survival likelihood is the average risk-adjusted survival likelihood for all.

Beneficiary episodes attributed to the TIN:

$$\text{TIN's Risk adjusted 30- and 60- day Survival Likelihood} = \frac{1}{N_j}\sum_{i=1}^{N_j} RiskAdj.Y_{ij}$$

$$RiskAdj.Y_{ij} = 0 \text{ ; for } Y_{ij} = 0,$$

$$RiskAdj.Y_{ij} = \frac{Y_{ij}}{P_{ij}} - 1 \text{ ; for } Y_{ij} = 1,$$

$$\text{where } P_{ij} = \frac{\exp(X_{ij}\beta)}{1+\exp(X_{ij}\beta)}$$

Where, N_j is the number of beneficiary episodes attributed to the jth TIN, P_{ij} is the expected probability of survival given the beneficiary covariates (obtained from logistic regression), while Y_{ij} is the observed probability of survival.

The risk adjusted survival likelihood for a TIN tells us how good the TIN is at keeping "sicker" beneficiaries alive across 30- or 60-day episodes. This measure is applicable to all selected MS-DRG conditions. The descriptive statistics for CHF, pneumonia, COPD, and hip replacement & knee replacement TINs are presented in Table 3.9, while the results of the risk adjustment are summarized in Appendix 20.

Table 3.9: Descriptive statistics for 30 and 60 day Episode Mortality Rate and TIN's Risk-Adjusted Survival Likelihood for CHF, Pneumonia, COPD, Hip Replacement & Knee Replacement

Condition	Number of Episodes	Number of Tins	30 day Episode Mortality Rate	60 day episode Mortality Rate	TIN's Risk-Adjusted 30-day Survival Likelihood: Mean (Std)	TIN's Risk-Adjusted 60-day Survival Likelihood: Mean (Std)
CHF	107,185	21,120	9.5 percent	15.4 percent	0.11 (0.1)	0.15 (0.12)
Pneumonia	86.869	20.056	12.0 percent	13.9 percent	0.13 (0.1)	0.15 (0.14)
COPD	78,760	18.525	4.9 percent	8.4 percent	0.05 (0.06)	0.08 (0.09)
Hip Replacement	24,603	10,083	0.7 percent	1.0 percent	0.02 (0.1)	0.03 (0.1)
Knee Replacement	52,647	12,189	0.3 percent	0.4 percent	0.01 (0.04)	0.01 (0.05)

Among the three medical MS-DRGs shown in table 3.9, pneumonia has the highest 30-day episode mortality rate and CHF the highest 60-day episode mortality rate. Among the two surgical MS-DRGs the 30- and 60-day episode mortality rates are higher for hip replacement than for knee replacement. The mean for the TIN's risk adjusted 30-day survival likelihood should be interpreted as follows: the average TIN's 30-day probability of survival for CHF episodes is 11 percent higher than expected given the beneficiary covariates for the episodes. The TIN's risk adjusted 60-day survival likelihood is interpreted similarly: the average TIN's 60-day probability of survival for CHF episodes is 15 percent higher than expected, given the beneficiary covariates for episodes. A TIN's mean risk adjusted survival likelihood is higher if the TIN manages sicker episodes (identified by their CCs) *and* the beneficiaries are alive at the end of the 30- or 60- day period.

The mean risk-adjusted survival likelihood for TINs varies by conditions and episode lengths. Conditions/episode lengths with higher mortality rates, like CHF and Pneumonia have higher mean risk adjusted survival likelihood, due to TINs who keep sicker beneficiaries alive for such conditions/episode lengths having higher survival likelihood scores.

Figure 3.5 (a): Distribution of Average Unadjusted 30-day Survival Rate for CHF TINs

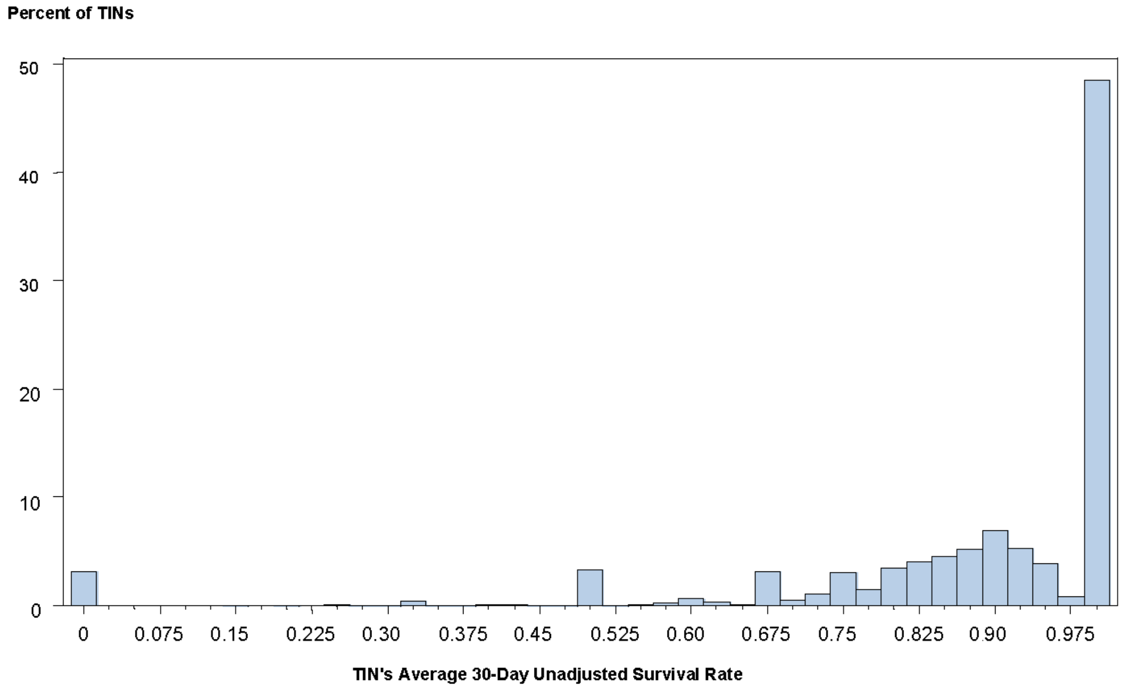


Figure 3.5 (b): Distribution of TIN's Risk Adjusted 30-day Survival Likelihood for CHF TINs

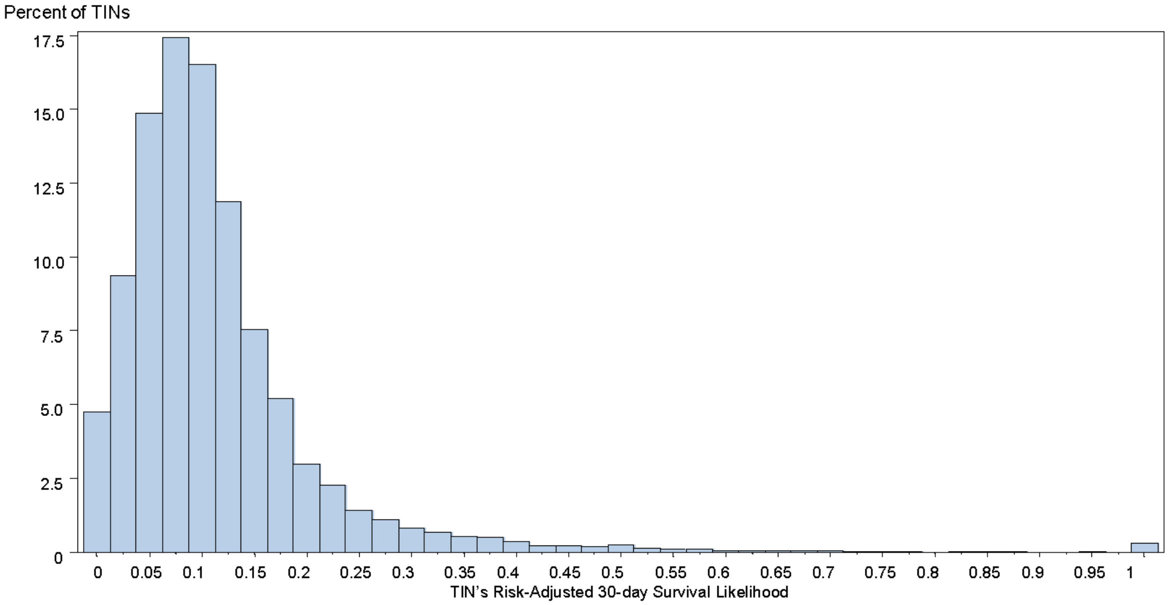


Figure 3.5(a) above is a bar graph showing the distribution of average unadjusted 30-day survival rates for CHF TINs – 48 percent of the CHF TINs have all of their episodes surviving for 30-days. The distribution of unadjusted survival rates for other TINs varies from zero to 1, with higher values for TINs that have more episodes without death.

Figure 3.5(b) is a bar graph showing the distribution of risk adjusted 30-day survival likelihood for CHF TINs, which also varies from 0 to 1. TINs with episodes of CHF where the beneficiary died have risk adjusted 30-day survival likelihood close to zero. TINs that 'successfully' manage sicker episodes of CHF (success here is defined in this case as having no episode mortality) are higher on the distribution.

The MS-DRG TEP identified case mix adjusted mortality as an important quality measure for measuring quality of care provided by physician entities in hospitals, giving it a weight of 33 out of 100. We positively framed this measure as 30- and 60- day risk adjusted survival likelihood for TINs. This measure is closer to zero when TINs have more episodes associated with mortality in the episode window, and closer to 1 when TINs successfully manage sicker episodes keeping beneficiaries alive at the end of the 30- or 60- day period. We use this measure for calculating a composite quality score for 30- and 60 day episodes for TINs, to be discussed later.

b. Risk-adjusted 30 and 60 day potentially preventable readmission free likelihood

One source of costly inpatient care is potentially preventable readmissions. According to MedPAC, potentially preventable hospital readmissions cost Medicare $12 billion in 2005. Readmissions are especially interesting from a quality point of view because they may be a result of errors of omission or commission during the index hospitalization or post-discharge. They may reflect poor discharge planning, poor coordination of services during the transition from acute to post-acute care, or poor choice of post-acute care setting (Goldfield, et al., 2008).

Goldfield and colleagues (2008) at 3M developed an algorithm to evaluate readmissions to determine whether they are "potentially preventable." MS-DRGs associated with a rehospitalization are mapped into All Patient Refined DRGs (APR DRGs). The APR DRGs have four severity of illness levels that are used to assign to each discharge a probability that the discharge was a PPR. The probabilities are then used to calculate the expected rate of PPR readmissions to which an actual rate is compared. 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.

Using the 3M PPR algorithm we identified potentially preventable readmissions within 30- and 60- day episode windows. We excluded episodes where the beneficiary died outside a hospital setting. We then used Medicare claims 12 months prior to the index hospitalization to obtain the diagnoses that were classified into CCs based on the CMS HCC model. We obtained demographic variables from the beneficiary summary file. We ran the analysis by beneficiary episode. The majority of the variables in our risk adjustment model are used in CMS's validated model for 30-day all-cause hospital-specific readmission measures. The covariates for our risk adjustment model included:

1. 70 HCC diagnostic categories.
2. Beneficiary's age and sex
3. Medicaid status
4. Reason for Medicare eligibility
5. Dummy variables identifying whether the index hospitalization DRG had major complications/comorbidities (MCC), complications/comorbidities (CC) or no comorbidities

1. A severity variable which identified the counts of prior hospitalizations for the same condition a year prior to the beneficiary-episode
2. Another severity variable which identified the number of prior emergency department visits (not resulting in hospitalization) for the same condition a year prior to the beneficiary-episode
3. Variables for history of AMI, CABG, PTCA and CABG in the last 12 months.

Majority of the beneficiary-episodes do not have any PPRs (approximately 80 percent for 30-day episode and 70 percent for 60-day episodes). When beneficiary episodes do have PPRs within the episode window, they sometimes have more than one readmission. Hence the variance for PPR counts is much greater than the mean. This is known –over dispersion, a problem commonly encountered when modeling count data. To avoid concerns of over dispersion due to excess zeroes, we did not model PPRs as counts in the 30- and 60- day window.

We instead chose to model *not having any PPRs* in the 30 and 60 day window using logistic regression at the beneficiary episode level. For beneficiary episodes where the beneficiary had no PPRs at the end of the episode window- we calculated the ratio of the observed to expected probability no PPRs. If a beneficiary's expected probability having no PPRs is 0.5, the ratio of observed to expected probability of no PPRs is 2, meaning that the beneficiary's probability of no PPRs is twice what is expected given the beneficiary's covariates. This ratio shows how successful a TIN is at managing the episode of a morbid beneficiary (whose expected probability of PPRs is less than 1). We set the maximum for the ratio of observed to expected probability of no PPRs for a beneficiary episode to 2, so as to not give undue credit to TINs for treating really sick beneficiaries (<5% of the beneficiary episodes had the ratio of observed to expected probability of no PPRs greater than 2). We then subtracted 1 from the ratio so as to allow it to be interpreted as percentage risk-adjusted PPR-free likelihood. If a beneficiary had an observed to expected ratio of 2, the risk adjusted PPR-free likelihood was 100%. If the beneficiary had any PPRs during the beneficiary episode- the risk-adjusted PPR-free likelihood was set to 0. The TINs Risk adjusted PPR-free likelihood is the average risk-adjusted PPR-free likelihood for all beneficiary episodes attributed to the TIN:

$$\text{TIN's Risk adjusted 30 and 60 day PPR-free likelihood} = \frac{1}{N_j}\sum_{i=1}^{N_j} RiskAdj\,Y_{ij}$$

$$RiskAdj\,Y_{ij} = 0 \text{ ; for } Y_{ij} = 0 \text{ ,}$$

$$RiskAdj\,Y_{ij} = \frac{Y_{ij}}{P_{ij}} - 1 \text{ ; for } Y_{ij} = 1 \text{ ,}$$

$$\text{where} \quad P_{ij} = \frac{\exp(X_{ij}\beta)}{1+\exp(X_{ij}\beta)}$$

Where, N_j is the number of beneficiary episodes attributed to the jth TIN, P_{ij} is the expected probability no PPRs given the beneficiary covariates (obtained from logistic regression), while Y_{ij} is the observed probability of no PPRs.

The risk adjusted PPR-free likelihood for a TIN tells us how good the TIN is at keeping ' sicker' beneficiaries free from potentially preventable readmissions across 30- or 60-day episodes. This measure is applicable to all selected MS-DRG conditions. The descriptive statistics for CHF, pneumonia, COPD, and hip replacement and knee replacement TINs are presented in Table 3.10, while the results of the risk adjustment are summarized in Appendix 20.

Table 3.10: Descriptive statistics for 30 and 60 day Episode Mortality Rate and TIN's Risk-Adjusted Potentially Preventable Readmission-Free (PPR-free) Likelihood for CHF, Pneumonia, COPD, Hip Replacement & Knee Replacement

Condition	Episodes	TINs	30 day Episode PPR Rate	60 day Episode PPR Rate	TIN's Risk Adjusted 30 day Adjusted PPR-Free Likelihood: Mean (Std)	TIN's Risk Adjusted 60 day Adjusted PPR-Free Likelihood: Mean (Std)
CHF	107,185	21, 120	24.6 percent	37.6 percent	0.21 (0.17)	0.29 (0.18)
Pneumonia	86,869	20,056	17.2 percent	26.1 percent	0.18 (0.17)	0.2 (0.19)
COPD	78,760	18,525	20.5 percent	32.2 percent	0.18 (0.08)	0.27 (0.15)
Hip Replacement	24,603	10,083	17.0 percent	19.8 percent	0.2 (0.22)	0.21 (0.22)
Knee Replacement	53,647	12,189	14.9 percent	17.3 percent	0.17 (0.18)	0.18 (0.19)

Among the three medical MS-DRGs shown in table 3.10, CHF has the highest 30- and 60- day episode PPR rate. Among the two surgical MS-DRGs, the 30- and 60-day episode PPR rate is marginally higher for hip replacement than for knee replacement. The mean for the TIN's risk adjusted 30-day PPR-free likelihood should be interpreted as follows: the average TIN's 30-day probability of having no potentially preventable readmissions for CHF episodes is 21 percent higher than expected given the beneficiary covariates for episodes. The TIN's risk adjusted 60-day PPR-free likelihood is interpreted similarly: the average TIN's 60-day probability of having no potentially preventable readmissions for CHF episodes is 29 percent higher than expected, given the beneficiary covariates for episodes. A TIN's mean risk adjusted PPR-free likelihood is higher if the TIN manages sicker episodes (identified by their CCs) *and* the beneficiaries do not have any potentially preventable readmissions in the 30- or 60- day period.

The mean risk-adjusted PPR-free likelihood for TINs varies by conditions and episode lengths. Conditions/episode lengths with higher PPR rates, like CHF and COPD have higher mean risk adjusted PPR-free likelihood, due to TINs who keep sicker beneficiaries out of the hospital for such conditions/episode lengths having higher PPR-free likelihood scores.

Figure 3.6 (a): Distribution of Average Unadjusted 30 day Potentially Preventable Rehospitalization Free rate for CHF TINs

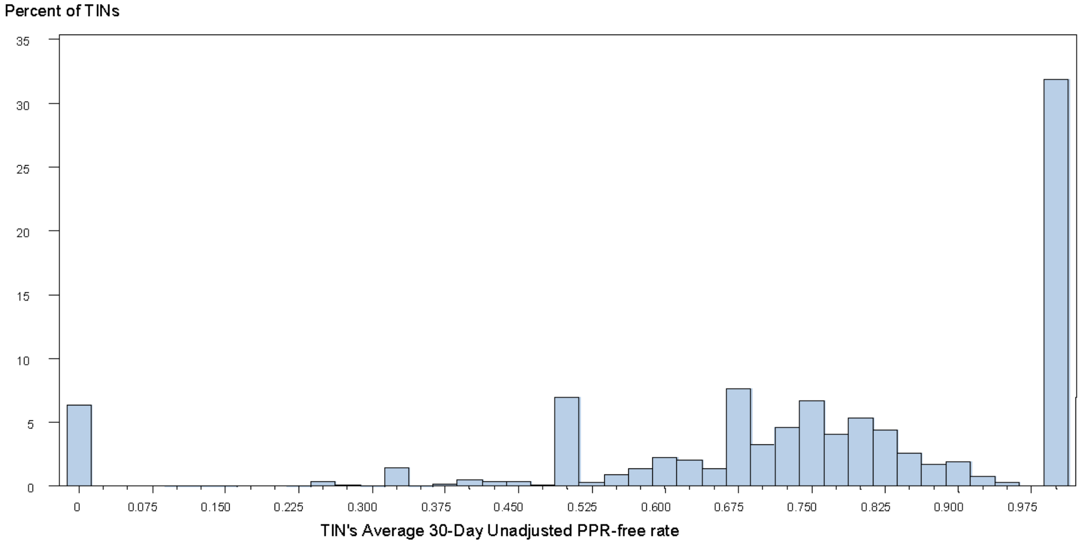
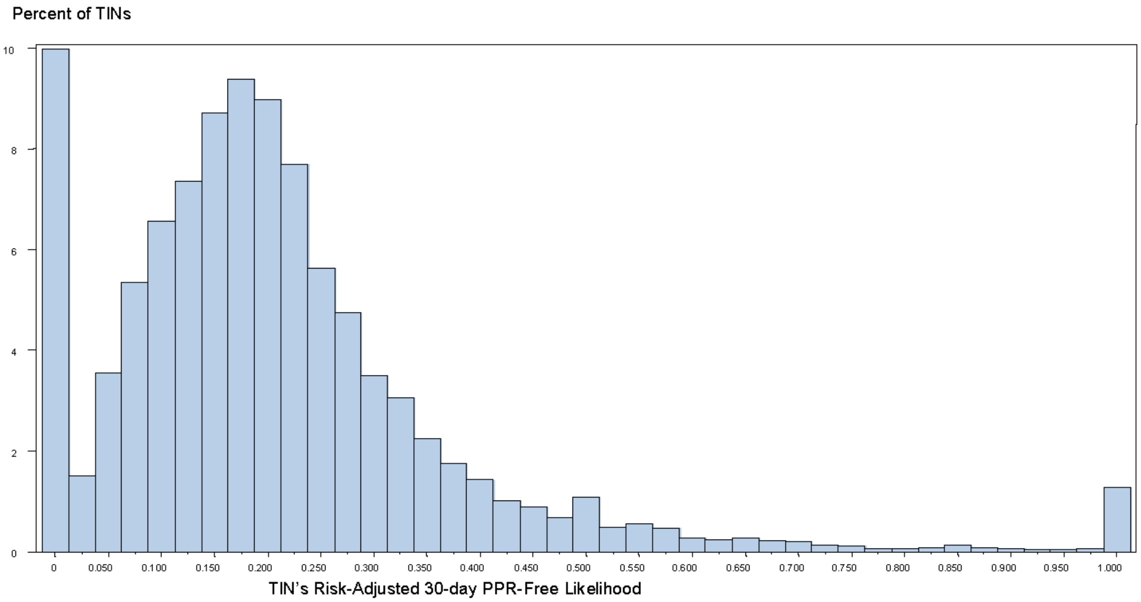


Figure 3.6(a) is a bar graph showing the distribution of average unadjusted 30-day PPR-free rate for CHF TINs – 32 percent of the CHF TINs have no PPRs in all of their 30-day episodes. The distribution of unadjusted PPR-Free rate for other TIN varies from zero to 1, with higher values for TINs that have more episodes without PPRs. Figure 3.6(b) below is a bar graph showing the distribution of risk adjusted 30-day PPR-free likelihood for CHF TINs, which also varies from 0 to 1. Approximately10 percent of CHF TINs that had PPRs for all their 30-day episodes of CHF have a PPR-free likelihood of zero. TINs with some episodes of CHF where their beneficiaries had PPRs have their risk adjusted 30-day PPR-free likelihood close to zero. A TIN's PPR-free likelihood would be relatively lower on the distribution if it managed episodes of CHF where the beneficiaries were relatively healthier. TINs that successfully manage sicker episodes of CHF, without having any PPRs are higher on the distribution.

Figure 3.6 (b): Distribution of TINs Risk Adjusted 30 Day Potentially Preventable Rehospitalization Free Likelihood for CHF TINs



The MS-DRG TEP identified PPR rate as an important quality measure for measuring quality of care provided by physician entities in hospitals, giving it a weight of 62 out of 100. We positively framed this measure as 30- and 60- day risk adjusted PPR free likelihood for TINs. This measure is closer to zero when TINs have more episodes associated with PPRs in the episode window, and closer to 1 when TINs successfully manage sicker episodes- keeping beneficiaries out of the hospital at the end of the 30- or 60- day period. We use this PPR-free likelihood measure for calculating a composite quality score for 30- and 60 day episodes for TINs, to be discussed later.

c. Thirty and 60 day Probability of No Avoidable Emergency Department (ED) Utilization

The purpose of the avoidable ED visit quality measure is to assess the appropriateness of ED use, post-discharge, within the 30- and 60- day episode window. While ED visits can be indicators of problems with access to primary care, our TEP panel was of the view that ED visits are also indicators with problems in continuity of care after discharge from the hospital. Physician entities providing inpatient hospital care should thus be held accountable for ensuring that beneficiaries are in an appropriate post-acute care setting where continuity of care is maintained. One way to do that is by including avoidable ED visits as a quality measure in profiling hospital physician performance. According to the MS-DRG TEP, avoidable ED visits were more useful than all-cause ED visits for comparing how hospital based physicians perform in providing better continuity of care. The TEP gave a score of 53/100 for avoidable ED visits versus 27/100 for all cause ED visits. We hence chose to focus only on avoidable ED visits for profiling

We use the algorithm developed by Billings et al. (2000) to calculate avoidable ED visits within 30- and 60-days of the index hospitalization for the selected MS-DRG TINs using post-discharge Medicare claims. In Medicare claims data, each ED visit has one or more diagnosis codes. The Billings algorithm considers the 640 diagnosis codes that include most of the high-volume codes for ED visits. Based on clinical judgment, the Billings algorithm then assigns each of the 640 codes a probability of four different categories of appropriateness.[24] The assigned probabilities sum to 1.0 across the four categories for any single diagnosis.

The four categories of appropriateness in the ED algorithm are:

I. Truly emergent – ED care needed, and not preventable/avoidable – Immediate care in an ED setting is needed and the condition could not have been prevented/avoided with ambulatory care.

II. Emergent – primary care treatable – Care is needed within 12 hours, but care could be provided in a typical primary care setting. If care was needed within 12 hours but could have been treated in a primary care setting, it falls into a category of a possibly inappropriate visit.

III. Emergent – ED care needed, but preventable /avoidable – Immediate care in an ED setting is needed, but the condition potentially could have been prevented or avoided with timely and effective ambulatory care. Here, the issue is not whether the patient should have gone to the ED, but rather whether better ambulatory care would have obviated the necessity for the ED visit.

IV. Non-emergent – Cases where immediate care is not required within 12 hours. ED care may have been unnecessary if the patient could have waited more than 12 hours and sought care in the ambulatory care system.

We apply the Billings algorithm to each ED visit's diagnosis codes to determine the probabilities of each of the four categories of appropriateness for each diagnosis. The diagnosis with the highest score for "emergent-ED care needed: not preventable/avoidable (injuries included)" – i.e., the code that implies the likelihood of an *appropriate* ED visit – is identified as the diagnosis for the remainder of the analysis and is given that probability. This approach minimizes the likelihood that a TIN will be penalized for an ED visit that truly was appropriate. Using the single diagnosis derived in this manner, each visit is assigned probabilities for each level of appropriateness. This approach generates four probability "scores" for each ED visit. For any single visit the scores (probabilities) sum to one.

From this data, we develop an avoidable ED visit score for each 30 and 60-day episode and aggregate to get a score for the TIN. This score was computed separately for all selected MS-DRGs. This score was the average probability score per episode attributed to the TIN for all three types of *avoidable* ED utilization: non-emergent, emergent but primary care treatable, and emergent but preventable or avoidable. To arrive at this set of scores, we added the probability scores in each category of appropriateness across all the episodes attributed to the TIN and divided by the total number of episodes attributed to the TIN.[25] We thus obtained the TIN's

[24] Diagnosis codes related to injuries, mental health, alcohol abuse, and substance abuse are not classified by the Billings algorithm and are excluded from our performance measure calculations.

[25] If we had divided by the number of ED visits rather than the number of attributed beneficiaries then the probabilities across all types of ED visits including truly emergent visits would sum to one. However, in that case, a zero could have been interpreted two ways: either then TIN had no episodes with ED visits, or the TIN had no avoidable ED visits. Using our method avoids that problem and also acknowledges that every attributed episode is at risk for both an ED visit and an avoidable ED visit.

average probability of avoidable ED utilization per episode for 30 and 60 days. More than 95 percent of the TINs across conditions had average probability of avoidable ED utilization per episode less than or equal to 1[26]. We winsorized the probability to 1 for less than 5 percent of TINs and positively reframed it to get the TINs average *probability of no avoidable ED utilization per episode* for 30 and 60 days. The descriptive statistics for ED visits for 30-day episodes of CHF, pneumonia, COPD, hip replacement, and knee replacement are presented in Table 3.11

Table 3.11: Descriptive Statistics for ED Visits for 30 day Episodes and TINs for CHF, Pneumonia, COPD, Hip Replacement, and Knee Replacement

Condition	Episodes	TINs	ED Visits per 30 day Episode: Mean (Std)	TIN'S Average ED Visits per 30 day Episode: Mean (Std)	TIN'S Average Probability of no Avoidable ED Utilization per 30 day Episode: Mean (Std)
CHF	107,185	21, 120	0.3 (0.59)	0.57 (0.53)	0.78 (0.26)
Pneumonia	86,869	20,056	0.23 (0.52)	0.22 (0.32)	0.91 (0.18)
COPD	78,760	18,525	0.29 (0.61)	0.28 (0.38)	0.87 (0.22)
Hip Replacement	24,603	10,083	0.12 (0.37)	0.14 (0.31)	0.94 (0.17)
Knee Replacement	53,647	12,189	0.11 (0.36)	0.12 (0.26)	0.94 (0.17)

Among the three medical MS-DRGs shown in table 3.11, CHF has the highest mean number of ED visits per 30 day episode, closely followed by COPD. But when aggregated to the TIN level, the difference between the two diseases increases, largely because CHF episodes with 30 day ED visits were likely to have sicker index hospitalizations where they were treated by more TINs. The TIN's average probability of no avoidable ED utilization per 30 day episode is again much lower for CHF compared to the other medical conditions. The TIN's average probability of no avoidable ED utilization per 30-day episode is the same at the mean, for both hip and knee replacement.

The average probability of no ED visits per episode for TINs varies by MS-DRG condition. Conditions with higher probability of ED visits per episode, like CHF and COPD have lower probability of no ED visits per episode at the TIN level. Figure 3.7 (a), below is a bar graph showing the distribution of the TIN's Average Probability of Avoidable ED utilization per 30-day CHF Episode for CHF TINs. Approximately 52 percent of the CHF TINs have no avoidable ED visits for all of their CHF episodes. The distribution of the average probability of avoidable ED utilization per TIN is positively skewed with a long tail. Because the average probability of avoidable ED utilization is the sum of probability scores for the three types of avoidable ED visits – it is greater than or equal 1 for <3 percent of CHF TINs [27]. We winsorized the TIN's average probability of 30-day Avoidable ED utilization to 1 and reframed it as the average probability of no avoidable ED utilization per episode. Its distribution for 30-day CHF episodes is presented in the bar graph in figure 3.7 (b).

[26] 5 percent of TINs had average probability of avoidable ED utilization per episode greater than 1 if they managed multiple episodes where the probability of avoidable ED utilization per episode was 1.

[27] Less than 3 percent of CHF TINs had average probability of avoidable ED utilization per episode greater than 1 if they managed multiple episodes where the probability of avoidable ED utilization per episode was 1.

Figure 3.7(a): Distribution of TIN’s Average Probability of Avoidable ED Utilization per 30-day CHF Episode for CHF TINs

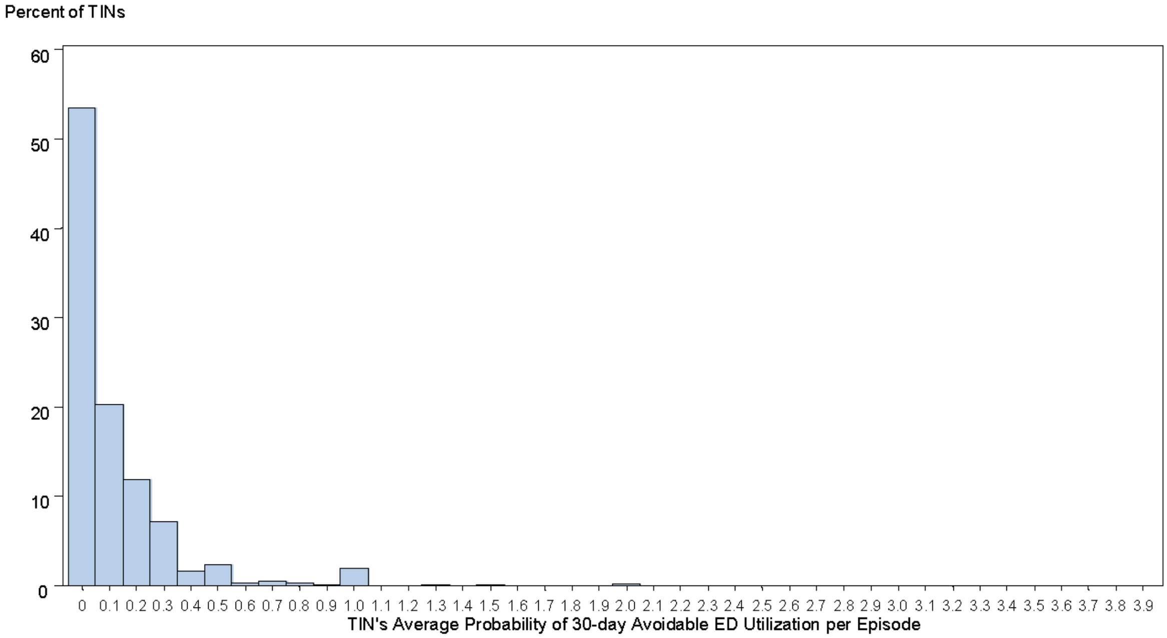
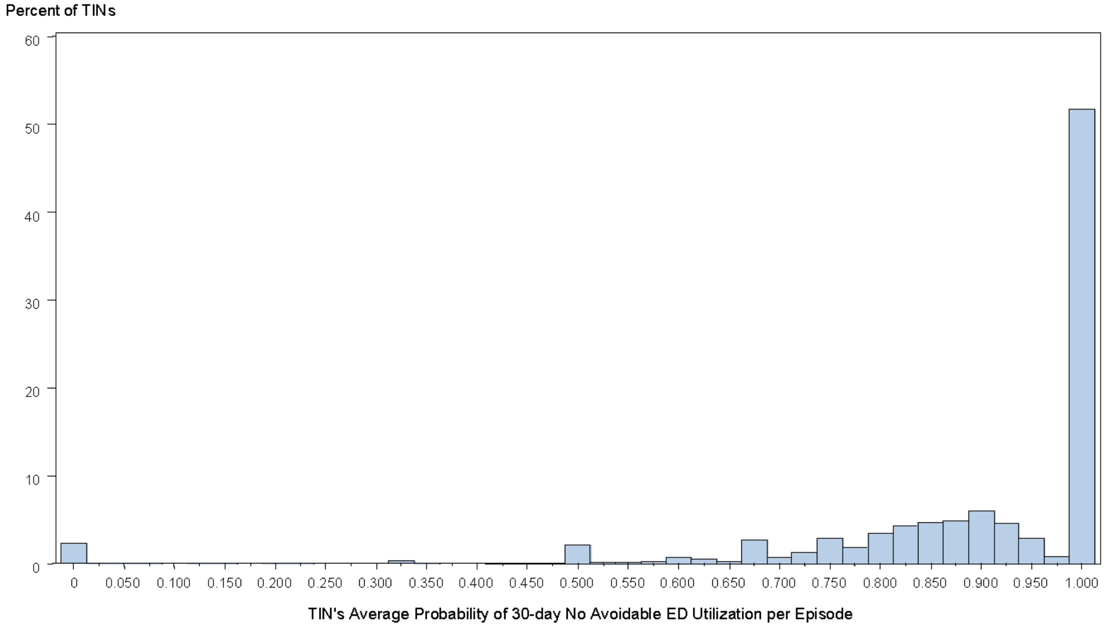


Figure 3.7 (b): Distribution of TIN's Average Probability of No Avoidable ED Utilization per 30-day CHF Episode for CHF TINs



The bar graph in Figure 3.7(b) above shows the distribution of the average probability of no avoidable ED utilization per 30-day episode for CHF TINs. TINs with average probability close to 1 have very low avoidable ED utilization for their 30 day CHF episodes while TINs far from1 have a lower probability of no avoidable ED utilization. This distribution in figure is Figure 3.7 (b) is very similar to the distribution observed for the average unadjusted survival probability and PPR-free likelihood for 30-day episodes of CHF, in Figures 3.6 (a) and 3.6 (b) respectively. This raises the question of whether risk adjustment models similar to those applied for mortality and readmission measures would work for avoidable ED utilization. However, we did not attempt to perform any risk adjustment to the avoidable ED utilization measures because of various shortcomings in the Billings algorithm in its current form, that were beyond the scope of this project to correct.

- The algorithm was based on the experience of an emergency department in New York City, and thus may not be entirely transferrable to other locations – e.g., less urban or rural settings.
- The algorithm does not consider the cumulative or interactive effects of multiple diagnoses.

The MS-DRG TEP identified avoidable ED visits as an important quality measure for measuring quality of care provided by physician entities in hospitals, giving it a weight of 53 out of 100. We positively framed this measure as the probability of no avoidable ED visits per 30- or 60-day episode for TINs. We use this measure for calculating a composite quality score for 30- and 60 day episodes for TINs, to be discussed later. However, we recommend that CMS consider addressing the shortcomings of the Billings algorithm, discussed above; if it is to be incorporated into a value based purchasing system. Our use of the algorithm in its current form thus should be considered illustrative.

V. Hospital Level Measures

Due to limited episode level quality measures for MS-DRG conditions, the MS-DRG TEP recommended that hospital level process and outcome measures be linked to MS-DRG conditions for those conditions to which these measures are applicable. We linked two publicly available hospital quality measure sets to the MS-DRG episodes: CMS' Hospital Compare process of care measure set, and the AHRQ Patient Safety Indicators (PSI) measure set. These measures are at the hospital level, and they do not vary by episode duration. As a result, they are the same for 30- and 60-day episodes across TINs.

A. Hospital Level Process Measures: Hospital Compare Process of Care Measure Set

The Hospital Compare Processes of Care Measure Set contains process measures collected over a 12 month period at the hospital level for AMI, pneumonia, CHF, and the Surgical Care Improvement Project. The September 2009 Hospital Compare Release contained process measures reported by hospitals for cases seen from January 2008 through December 2008. We did not include mortality and readmission measures from hospital compare for two reasons. First, these measures can be computed and risk adjusted at the episode level from claims, which we do separately. Second, the mortality and readmission measures are collected and reported over a period of 36 months in hospital compare and thus do not line up with annual cycles of performance measurement that are more appropriate for a payment system.

We linked hospital level measures from the Hospital Compare Process Measure Set to appropriate to MS-DRG episodes:

- AMI – Measures for AMI were linked to episodes for Medical AMI, AMI with CABG, and AMI with PTCA.
- Pneumonia and CHF – Measures for pneumonia and CHF were linked to episodes for pneumonia and CHF respectively.
- Surgical care measure set – Measures for the surgical care measure set were linked to all surgical MS-DRGs (as shown previously in Table 3.8(b)).

We used the hospital provider number on the index hospitalization to link an episode to the hospital level measure set. We then attributed the measures to all TINs involved in the care of the episode and aggregated the measures to the TIN level. For a TIN that treated episodes in multiple hospitals, the TIN's aggregate measure was a weighted average of the TIN's episode volume in the hospitals.

Table 3.12 (a): Descriptive statistics for CHF Hospital Compare Measure for CHF TINs

Process Measure	Average TIN Score: Mean (Std)
CHF-1: Patients Given Discharge Instructions	0.82 (0.15)
CHF-2: Patients Given an Evaluation of Left Ventricular Systolic Function	0.95 (0.08)
CHF-3: Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction	0.92 (0.08)
CHF-4: Patients Given Smoking Cessation Counseling	0.97 (0.09)

Table 3.12 (a) summarizes the descriptive statistics for the CHF hospital compare measures attributed to the CHF TINs. The means can be interpreted as follows: the average hospital quality score for CHF patients receiving ACE inhibitors or ARBs (measure CHF-3) across hospitals attributed to CHF TINs is 92 percent. Figure 3.8 is a bar graph that shows the distribution of average hospital scores for measure CHF-3[28] across all the CHF TINs. The distribution has a long left tail, with most CHF TINs providing care for CHF episodes in hospitals that have higher scores, closer to 1. Few TINs provide care for CHF episodes in hospitals with low CHF hospital compare scores. Nevertheless the variation in average hospital compare scores is reflective of the variation in process quality for CHF episodes in different hospitals, resulting from variation in how physicians provide recommended care to CHF patients.

Table 3.12 (b): Correlation Matrix for CHF Hospital Compare Measures (and level of significance)

CHF Measure	CHF-1	CHF-2	CHF-3	CHF-4
CHF-1: Patients Given Discharge Instructions	1.00000			
CHF-2: Patients Given an Evaluation of Left Ventricular Systolic Function	0.48687 (<.0001)	1.00000		
CHF-3: Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction	0.55171 (<.0001)	0.46318 (<.0001)	1.00000	
CHF-4: Patients Given Smoking Cessation Counseling	0.47919 (<.0001)	0.56041 (<.0001)	0.34562 (<.0001)	1.00000

Tables 3.12(c) and 3.12(d) summarize the descriptive statistics for the AMI process measures attributed to the medical AMI TINs, and surgical care process measures attributed to AMI- CABG TINs. With the exception of the measures for patients receiving fibrinolytic medication within 30 minutes of arrival and patients receiving percutaneous coronary intervention within 90 minutes of arrival, all process measures have high scores at the TIN level. Table 3.12 (e) summarizes the descriptive statistics for Pneumonia process measures attributed to Pneumonia TINs. It also shows high average scores for hospitals attributed to Pneumonia TINs for all the measures in the measure set. The correlation matrix for the AMI, Surgical Care and Pneumonia measure sets are presented in Tables 7-A, 7-B and 7-C of Appendix 7. Here again we see modest positive correlation among all measures in the measure set – suggesting that they are again reflective of an underlying quality construct.

[28] CHF-3 is patients given ACE inhibitor or ARB for left ventricular systolic dysfunction.

Table 3.12 (c): Descriptive statistics for AMI Hospital Compare Measure for Medical AMI TINs

Process Measure for AMI	**Average TIN Score: Mean (Std)**
AMI-1:Patients Given Aspirin at Arrival	0.973 0.0459)
AMI-2: Patients Given Aspirin at Discharge	0.954 (0.07043)
AMI-3: Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction	0.928 (0.10404)
AMI-4: Patients Given Smoking Cessation Advice/Counseling	0.979 (0.07505)
AMI-5: Patients Given Beta Blocker at Discharge	0.959 (0.06865)
AMI-7A: Patients Given Fibrinolytic Medication Within 30 Minutes Of Arrival	0.448 (0.21794)
AMI-8A:Patients Given PCI Within 90 Minutes Of Arrival	0.767 (0.13057)

Table 3.12 (d): Descriptive statistics for Surgical Care Hospital Compare Measure for AMI - CABG TINs

Process Measures for Surgical Care	**Average TIN Score: Mean (Std)**
Surg-Inf-1: Patients given Antibiotic at right time	0.922 (0.0692)
Surg-Inf-2: Patients given right kind of Antibiotic	0.956 (0.04323)
Surg-Inf-3: Patients' Antibiotic stopped at right time	0.87 (0.10007)
Surg-Inf-4: Heart Surgery Patients with Blood Sugar under Control	0.861 (0.16536)
Surg-Inf-6: Patient receiving safe hair removal	0.968 (0.06159)
Surg-VTE-1: Patients who received treatment to prevent blood clots after surgery	0.89 (0.08647)
Surg-VTE-2: Patients who got treatment at the right time to prevent blood clots after surgery	0.859 (0.09414)

Table 3.12 (e): Descriptive Statistics for Pneumonia Hospital Compare Measures for Pneumonia TINs

Process Measures for Pneumonia	**Average TIN Score: Mean (Std)**
PNEU-1:Patients Given Oxygenation Assessment	0.996 (0.01187)
PNEU-2: Patients Assessed and Given Pneumococcal Vaccination	0.864 (0.13249)
PNEU-3: Patients Receiving ED Blood Culture Prior to First Hospital Antibiotic	0.928 (0.06066)
PNEU-4: Smoking Cessation Counseling	0.944 (0.10823)
PNEU-5: Patients Given Initial Antibiotic(s) within 6 Hours After Arrival	0.936 (0.04859)
PNEU-6: Patients Given the Most Appropriate Initial Antibiotic	0.891 (0.07123)
PNEU-7: Patients Assessed and Given Influenza Vaccination	0.817 (0.15152)

B. AHRQ Patient Safety Indicators

AHRQ PSIs are a set of outcome measures for potentially avoidable complications that patients may experience while hospitalized. The AHRQ PSIs are computable from claims if their hospitalization claims have present on admission (POA) indicators for diagnosis codes. Unfortunately, the Medicare inpatient hospitalization claims that we had did not have POA indicators. We thus used a substitute method: we took hospital level PSI measures and linked them to MS-DRG conditions, for those conditions for which these measures were appropriate (summarized earlier in Tables 3.8 (a) and (b)). PSI measures at the hospital level are collected and reported over a 20-month period at the hospital level for seven of AHRQ's PSIs:

1. PSI 5- Foreign body left during procedure;
2. PSI 6- Iatrogenic Pneumothorax;
3. PSI 7- Catheter related blood infections;
4. PSI 11- Postoperative respiratory failure;
5. PSI 12- Postoperative sepsis;
6. PSI 14-Postoperative wound dehiscence; and
7. PSI 15- Accidental puncture or laceration.

PSIs 5 and 7 are also reported as hospital acquired conditions in Hospital Compare. Hospital Compare's 2011 release reported hospital level PSI measures from October 2008-July 2010. Earlier releases of Hospital Compare did not report PSIs. So we used the 2011 release for illustrative purposes. Using the hospital provider number on the index hospitalization, we linked condition episodes to the hospital level PSI measure set. We then attributed the measures to all TINs involved in the care of the episode and aggregated the measures to the TIN level. For a TIN that treated episodes in multiple hospitals, the TIN's aggregate measure was a weighted average of the TIN's episode volume in the hospitals. PSI rates aggregated at the TIN level were rates of adverse outcomes per 1000 hospital discharges. We reframed the PSI measures positively at the TIN level as PSI-free rate per hospital discharge.

Table 3.13: Descriptive Statistics for Hospital Level PSI Measures for Hip Replacement TINs

Patient Safety Indicator	TIN's PSI Rate per 1000 Discharges: Mean (Std)	TIN's PSI-Free Rate per Discharge: Mean (Std)
PSI-5: Foreign body left during procedure	0.028 (0.064)	1.0000 (0.0001)
PSI-6: Iatrogenic Pneumothorax	0.381 (0.154)	0.9996 (0.0002)
PSI-7: Catheter related blood infections	0.441 (0.474)	0.9996 (0.0005)
PSI-11: Postoperative respiratory failure	9.075 (3.248)	0.9909 (0.0033)
PSI_12: Postoperative sepsis	5.765 (3.021)	0.9942 (0.0030)
PSI_14: Postoperative wound dehiscence	2.108 (0.404)	0.9979 (0.0004)
PSI_15: Accidental puncture or laceration	1.898 (0.793)	0.9981 (0.0008)

Table 3.13 above, summarizes the descriptive statistics for the PSI measures attributed to hip replacement TINs. The TINs' average PSI rate per 1000 discharges is low for most of the PSIs and the reframed PSI-free rate per discharge is almost equal to 1. Figures 3.9 (a) and (b) below are bar graphs showing the distribution of accidental laceration/puncture rate per 1000 discharges and the positively framed accidental laceration/puncture-free rate per discharge. The advantage of reframing PSIs as a positive outcome is that it allows us to identify TINs associated with hospitals that perform better with respect to patient safety outcomes. In the bar graph in Figure 3.9 (b) hip fracture TINs treating episodes in hospitals with lower accidental laceration/puncture rates are closer to 1.

Figure 3.9 (a): Distribution of TIN's Average Hospital Accidental Laceration/Puncture Rate per 1000 discharges for Hip fracture TINs

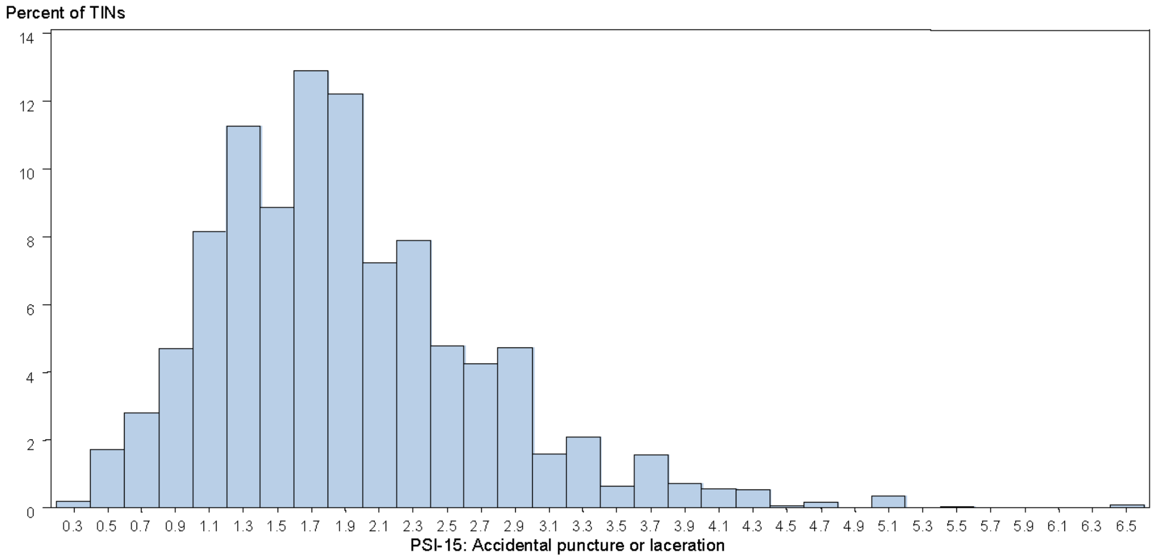
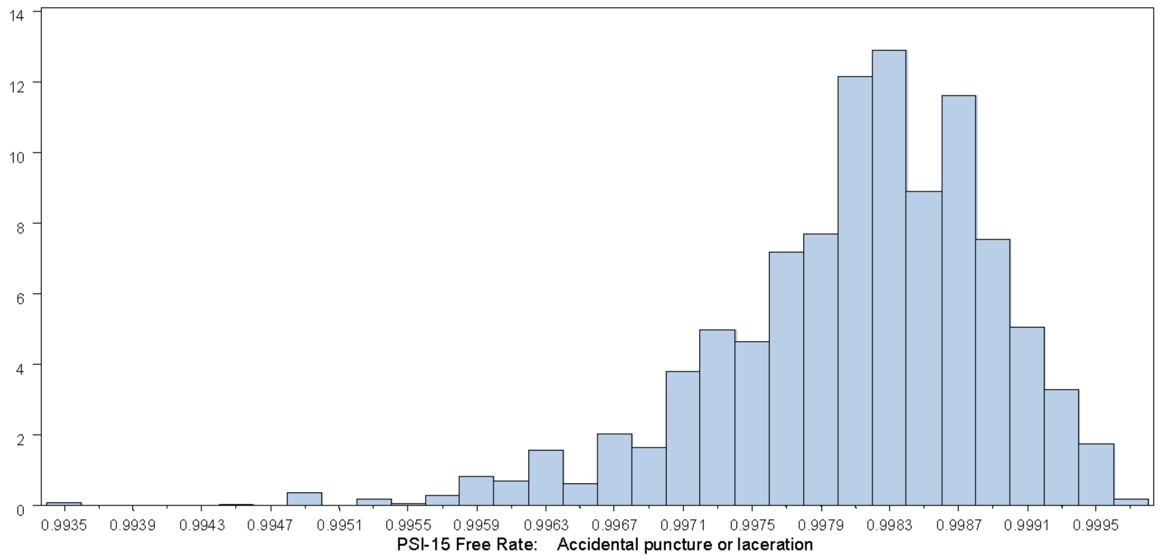


Figure 3.9 (b): Distribution of TIN's Average Hospital Accidental Laceration/Puncture-Free Rate discharge for Hip fracture TINs



The correlation matrix for patient safety indicators, in Appendix 8- Table 8 A, shows that *unlike hospital compare measures, there is very little correlation between the PSIs*. Catheter related blood infections and post-operative sepsis are moderately correlated (0.33) – the former often preceeds the latter in the causal pathway. But other PSI measures are very poorly correlated with one another (with correlation coefficients<±0.10). This shows that patient safety in hospitals is a much broader construct of quality, and in this case is defined by different, unrelated patient safety measures.

C. Issues in Constructing a Composite Quality Measure

For all the selected MS-DRG conditions, we identified, obtained, and – where appropriate – risk adjusted, five different sets of quality measures. As discussed in the previous section, three of these were episode level outcome measures, and two were hospital level measure sets, one that identified processes of care and the other outcomes.

To combine such measures requires that we establish weights for each. We did so by using expert judgment to rate the measures, as part of the work of the MS-DRG TEP panel (see discussion in section IV above). The TEP weights for MS-DRG quality measure sets are shown in Table 13.4 (a) and rescaled to 1. The rescaled TEP weights were rounded off and then used to construct a single composite quality measure for each MS-DRG condition.

Table 3.14 (a): Unscaled, Rescaled and Rounded off Technical Expert Panel Weights for MS-DRG Quality Measures

Quality Measure Set	Unscaled TEP Weight	Rescaled Weight	Rounded off Weight
Probability of No ED Utilization	53	0.22	0.2
Risk Adjusted Potentially Preventable Rehospitalization Free Likehood	62	0.26	0.3
Risk Adjusted Survival Likelihood	33	0.14	0.1
Medicare Hospital Compare Process Measure Rate	42	0.18	0.2
AHRQ PSI Free Rate	50	0.21	0.2
Total	240	1.00	1.00

For each of the hospital level quality measure sets, we examined the mean and variation in TEP-assigned weights. Table 13.4 (b) below shows the unscaled and rescaled TEP weights for the AHRQ hospital patient safety indicators. When rescaled, the TEP weights for the individual PSIs was almost the same as weighting all the PSIs equally. The TEP weighting for hospital compare measures was also equivalent to equal weighting. *We hence decided to weight all individual hospital compare measures and PSIs for a given condition equally*.

Table 3.14 (b): Unscaled and Rescaled Technical Expert Panel Weights for Hospital Patient Safety Indicators

PSI Measure	Unscaled TEP Weight	Rescaled Weight
PSI-5: Foreign body left during procedure	78	0.15
PSI-6: Iatrogenic Pneumothorax	68	0.14
PSI-7: Catheter related blood infections	73	0.14
PSI-11: Postoperative respiratory failure	65	0.13
PSI-12: Postoperative sepsis	74	0.15
PSI-14: Postoperative wound dehiscence	71	0.14
PSI-15: Accidental puncture or laceration	75	0.15
Total	503	1

Creating a final risk-adjusted composite quality measure raises a series of issues. Among the most important that we considered were the following:

i. Making Meaningful Imputations for Missing Values

Missing values for quality make it hard to create a composite quality measure. We made meaningful imputations if there were missing values for a quality measure. For instance, hospitals with low volumes had missing values for certain PSIs or hospital compare process measures. We therefore imputed the average values for hospitals in the sample for a condition of interest, in place of the missing values. In another instance, we had excluded episodes where the beneficiary died within the episode window, in estimating risk adjusted 30- and 60-day PPR-free rate and probability of ED utilization. We set the missing values for PPR-free rate and probability of ED utilization for such excluded episodes to zero. By making such an imputation, we imply that in the event of death during a beneficiary episode- the other avoidable episode-level measures for that beneficiary episode are set to their lowest possible value, thereby denying the TIN any credit for the beneficiary episode.

ii. Transforming Raw Scores to z-scores

Raw scores can be easily weighted when they are transformed to z-scores. This solves an important problem: many of the quality scores are measured on different scales, making nonsense of untransformed combinations. Transforming raw scores to z-scores before weighting them allows for weighting of scores on a common scale. All our quality measures have interval data – making them amenable to z-score transformation. Transforming raw scores to z-scores is a better option than transforming raw scores to ranks. Ranks have problems at both the means and the extremes. At the mean, a small change in raw score results in a large change in rank, while at the extreme a large change in raw scores results only in small change in rank. However, Z-scores however have a potential problem: the raw scores lose their meaning in terms of their standard deviations. Z-scores may hence magnify small differences in quality measures. But given that the quality measures are reweighted to weights shown in table 13.4, transforming to z-scores and then weighting is the ideal compromise.

D. Results

The distribution of risk adjusted composite quality scores for 30-day CHF episodes across CHF TINs is shown in the bar graph in Figure 3.10(a). The risk adjusted composite quality score is simply the percentile rank of the composite quality z-score. We choose to represent the composite quality score as a percentile rank because it is more easily interpretable than z-scores. Representing the composite quality score as a percentile rank allows it to range from zero to 100 (if the composite were to be represented as a z-score, it would have a mean of zero and a standard deviation of 1, making interpretation difficult at first glance). The mean of risk adjusted composite quality score across the CHF TINS is 49.5, and TINs seem to be evenly spread at the high and low ends of the CHF quality composite. The ‘normal’ distribution of CHF TIN quality scores might be explained by the large number of CHF TINs (21,120 TINs). In contrast, the distribution of risk adjusted composite quality scores for 30-day episodes of hip replacement with fewer TINs (10,083) – shown in the bar graph in Figure 3.10 (b) – is more skewed to the left, meaning that more TINs have better composite quality scores for hip replacement. The distribution of risk adjusted composite quality scores for 60-day episodes of CHF and hip replacement across TINs for those conditions, shown in Appendix 11, Figures 11.A and 11.B respectively, parallels the distribution seen for 30-day episodes.

Figure 3.10(a) Distribution of Risk Adjusted Composite Quality Score for 30-day CHF Episodes Across CHF TINs

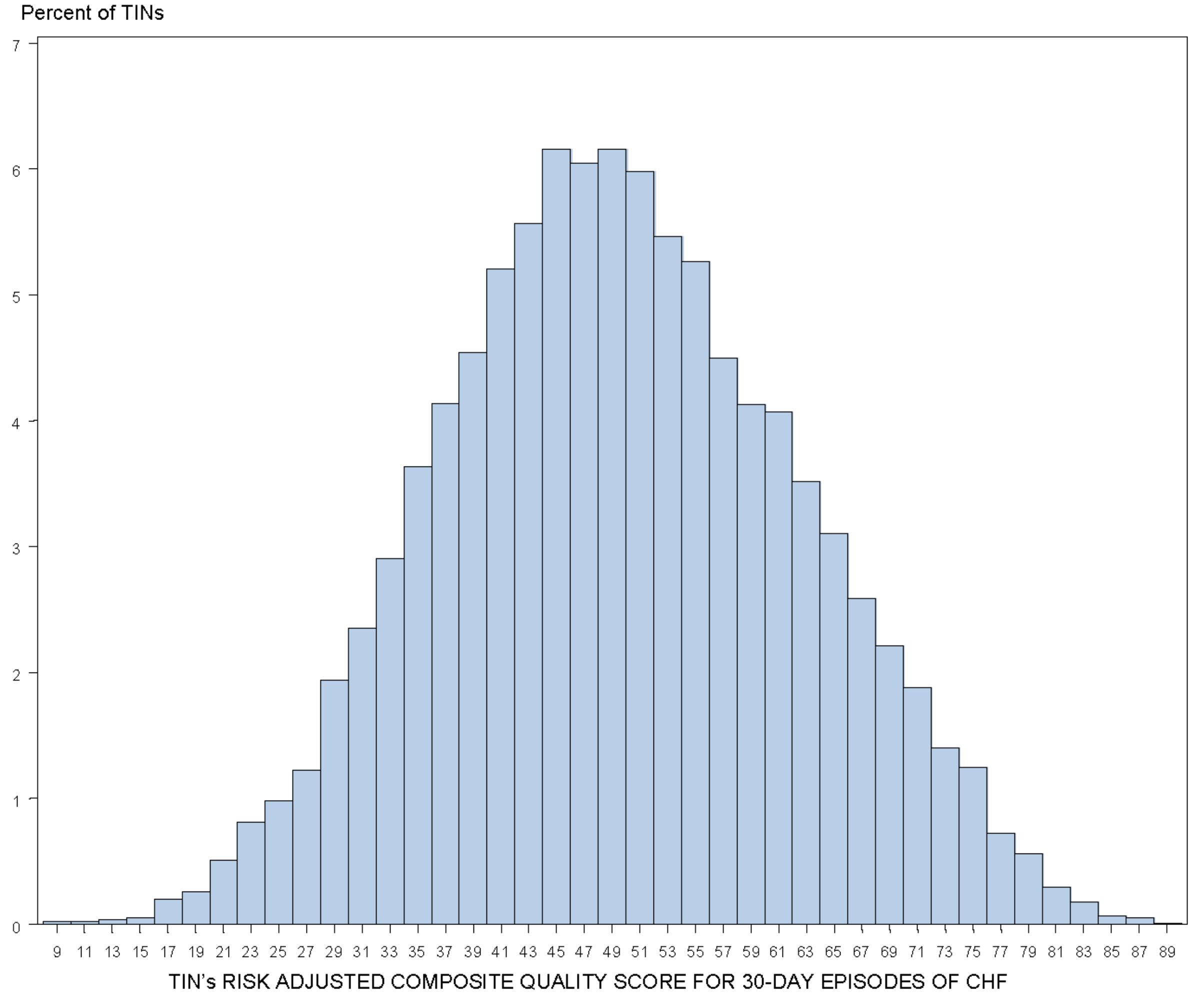
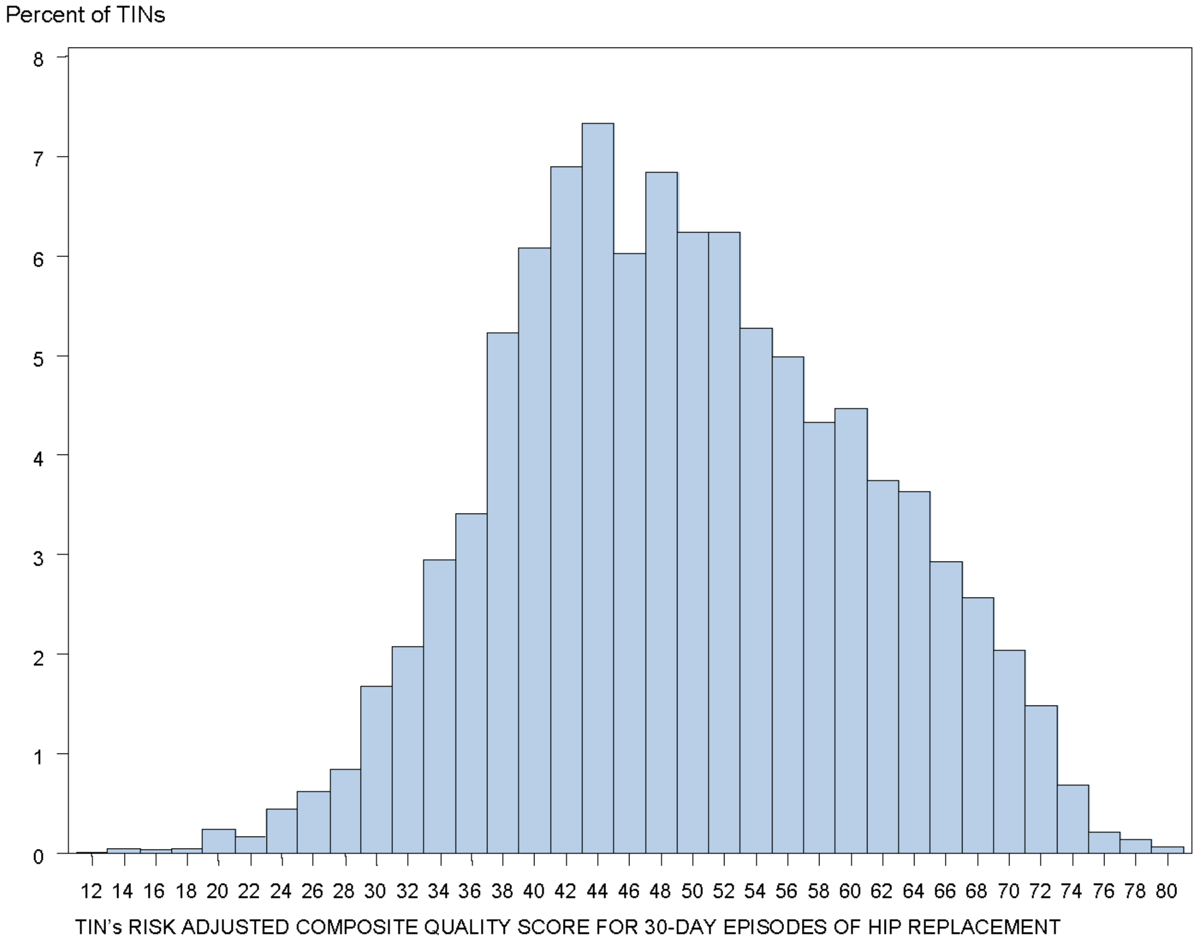


Figure 3.10(b) Distribution of Risk Adjusted Composite Quality Score for 30 day Hip Replacement Episodes Across Hip Replacement TINs



The fact that the variation in the composite quality scores shows this kind of distribution across TINs has a very important quality: *it allows us to tier TINs by quality* – e.g., by considering TINs within each of the mean episode proportion quintiles. TINs can then be rank ordered by their risk adjusted costs within each quality tier, thus allowing us to identify TINs that provide more 'value'. We will have more to say on this issue in Chapters 5 and 6, when we consider implementation issues and how to translate these measurement methods into a functioning, value based payment system.

Chapter 4

The Beneficiary Level Approach

In order to tailor the value based payment system to the physician's practice setting, we designed two approaches to measuring physician performance. Chapter 3 laid out the approach for measuring cost and quality for physicians who practice in the hospital. In this chapter, we detail the second approach: for physicians practicing primarily in outpatient settings. We term this the "beneficiary level" approach, because it holds a physician or physician group accountable for a beneficiary's entire year of care. Since all care during the year is included, the physician is held responsible for the cost of hospitalizations as well as the appropriateness of hospitalizations and emergency department utilization as part of the quality measures.

The essential features of the beneficiary level approach are:

I. The attribution rule – To hold outpatient physicians accountable, we first need to establish how care in the uncoordinated world of FFS Medicare will be attributed to a particular physician or physicians.

II. The units of analysis – Given our choice on attribution, how wide is the "window" over which care will be attributed to the physician? And will care be attributed to physicians one by one, or to groups of physicians (by Tax Identification Number, or TIN)?

III. Analytic issues in estimating cost and quality – This section reviews certain analytic issues that have to be resolved before we can perform the cost and quality measurement activities that we are proposing. The main issues to be discussed are the data we used for the Beneficiary Analysis and how we develop estimates from it; and general issues in risk adjusting cost and quality performance measures.

IV. Estimating attributed costs – Given an attribution rule and unit of analysis, computation of the actual cost of care is relatively straightforward. But considerations of fairness and efficiency require that measures of cost should be risk adjusted, and this section addresses special issues in risk adjusting costs. In addition, since certain characteristics of health care data require the analyst to pay special attention to the specification of the cost equations, this section describes and justifies the specification we have chosen.

V. Quality measures – We need to develop quality measures to appraise the physician's performance that are fair, actionable and meaningful to physicians. A key requirement in selecting such measures is that they all must be computable from claims data, so that they readily can be implemented as part of Medicare payment. All of the measures we have selected, as well as our approach to combining them, were vetted by a panel of national technical experts. The measures fall in three broad areas:

1. Potentially avoidable events, including
   a. Ambulatory care sensitive (ACS) hospitalizations
   b. Potentially preventable rehospitalizations
   c. Potentially preventable emergency department visits

2. Monitoring measures
3. Screening measures

VI. *Combining cost and quality into a performance measure* – Individual cost and quality measures must be aggregated into a composite index of performance, to provide a single metric of performance for the value based payment system. We take up these issues in Chapter 5.

VII. *Translating performance measures into payment policy* – How can the methodologies described above best be translated into an actual payment system? Chapter 6 addresses this key issue and includes considerations of peer grouping and stability of the performance measures over time.

I. Attribution

The purpose of a value based payment system for physician payment is to reward physicians for good performance – high quality and low cost care. Both cost and quality of care are measured at the point that the physician interacts with the beneficiary, and thus there must be some way to attribute measures of cost and quality to the physicians who are to be held accountable for the care those beneficiaries receive. This is a non-trivial issue in the uncoordinated Medicare FFS system, because Medicare beneficiaries typically have many different physicians caring for them, and there is little in the system itself to help allocate responsibility for the care among these physicians.

There are two ways to classify all attribution systems:

- *Ex ante versus ex post systems* – In *ex ante* systems, physicians know in advance which beneficiaries will be attributed to their practices, whereas in *ex post* systems, beneficiaries are assigned to physicians at the end of a reporting period.

- *Active versus passive systems* – In active attribution systems, the physician and patient agree that the beneficiary's care will be attributed to the specific physician, whereas in passive attribution systems, beneficiaries are assigned to physicians without either the beneficiary's or physician's prior knowledge or consent.

These systems can be combined in different ways. For example:

- We might use a passive, *ex post* attribution system – e.g., use analyses of claims after the fact to assign beneficiaries to the physician who provided the plurality of the beneficiary's non-hospital Evaluation & Monitoring (E&M) visits. Other examples of passive *ex post* attribution include the CMS Medicare Physician Group Practice (PGP) Demonstration (CMS, 2009) and the CMS Resource Use Report initiative (CMS, 2008b).

- An example of an active, *ex ante* system is the United Kingdom's National Health Service. Patients must sign up with primary care doctors who are their entry to the system (including access to specialist care). In turn, these primary care doctors are responsible for the quality and much of the cost of their care. The system makes explicit efforts to link quality of care to payments (Roland, 2004). By contrast, in the U.S. in the traditional FFS Medicare program, active, *ex ante* attribution is difficult because (unlike the NHS) there is no "gateway" physician, and beneficiaries are not restricted in their choice of physicians. There is no way formally to assign responsibility for a beneficiary's care "before the fact." The structure of the program invites uncoordinated care.

- The CMS Pioneer Accountable Care Organization (ACO) model uses a passive, but *ex ante* attribution system (http://innovations.cms.gov/Files/fact-sheet/Pioneer-ACO-General-Fact-Sheet.pdf)l. Physicians are told in advance which patients will be attributed to them, but physicians agree only to abide by the attribution *rule*. There is no *ex ante* contractual agreement between a specific patient and the physician, and the patient is not bound by any obligation to use the physician as her primary source of care.

In a normative sense, some kind of active, *ex ante* attribution would be preferable, because it would require the physician who is to be held accountable to accept responsibility before the fact – there is a critical element of awareness and consent, which could be a firm institutional and behavioral basis for performance measurement and accountability. Unfortunately, no model of that sort is possible in the current FFS Medicare program because beneficiaries are free to see any provider of their choice without any type of preauthorization.

The inability to implement an *ex ante* attribution in FFS Medicare also results in an important policy problem. *Ex post* attribution systems are incapable of attributing beneficiaries who used no health care services during the observation period. Thus, a physician who has taken steps to ensure that her patient requires minimal professional intervention will receive no credit if the beneficiary uses no Medicare-covered services during the ensuing observation period. The physician cannot be rewarded for her good efforts, which gives the efficient physician insufficient credit and – given the reduced incentive to the physician that results – harms the Medicare program.

We did explore the possibility of an attribution that involved elements of an active, *ex ante* system: the Medicare Physician Quality Reporting System (PQRS).[29] The 2006 Tax Relief and Health Care Act (TRHCA) (P.L. 109-432) authorized CMS to establish the PQRS, which enables physicians and other eligible "professionals"[30] to report data on health care quality and health outcomes beyond the measures available in traditional administrative (e.g., claims) data. The reporting system is voluntary. The first reporting period was the second half of 2007 (CMS, 2008a).

[29] The material that follows in this section is taken from the CMS website http:/www.cms.hhs.gov/pqri/. Note that, when first introduced, PQRS was termed the Physician Quality Reporting Initiative, a name that was changed in 2011 to Physician Quality Reporting *System*. For simplicity, we will use PQRS to label both versions throughout this report, regardless of the year.

[30] "Eligible professionals" include Doctor of Medicine, Doctor of Osteopathy, Doctor of Podiatric Medicine, Doctor of Optometry, Doctor of Oral Surgery, Doctor of Dental Medicine, Doctor of Chiropractic, Physician Assistant, Nurse Practitioner, Clinical Nurse Specialist, Certified Registered Nurse Anesthetist (and Anesthesiologist Assistant), Certified Nurse Midwife, Clinical Social Worker, Clinical Psychologist, Registered Dietician, Nutrition Professional, Audiologists (as of 1/1/2009), Physical Therapist, Occupational Therapist, Qualified Speech-Language Therapist (as of 7/1/2009). (http://www.cms.hhs.gov/PQRI/Downloads/EligibleProfessionals.pdf)

By submitting a PQRS report on a patient, the physician might be seen to take some responsibility for the care the patient receives. Thus, PQRS reporting might form the basis of an active and somewhat *ex ante* attribution system. We knew in advance that the PQRS was a very new program and reporting rates were low. Nonetheless, we compared PQRS-based attribution to the widely used "plurality of E&M visits" rule. Our preliminary analyses were based on 2008 data from Colorado. We showed as expected that physicians who provided PQRS 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 PQRS report on at least one beneficiary. We later extended that analysis to data from five states for the years 2008 and 2009. That analysis is described in detail in Appendix 12. We found that PQRS reporting was increasing rapidly from 2008 to 2009, but participation still was so low that PQRS reporting was not a feasible basis for attribution in this project. With the encouragement of the Beneficiary TEP panel, we turned our attention to more common *ex post,* passive attribution systems.

There are many possible *ex post*, passive attribution algorithms. For example, through retrospective analysis of Medicare claims, the care of Medicare beneficiaries could be attributed to the physician from whom the beneficiary obtained most of her office visits or largest dollar volume of her Medicare claims (a plurality rule). Alternatively, every physician who submitted a bill for a beneficiary could be assigned a part of the responsibility for the beneficiary's cost and quality of care based on the proportion of dollars or visits attributable to each physician (a proportionate rule). Mehrotra, et al. (2010) compared eleven different attribution algorithms and found that assignment of physicians to cost categories is sensitive to the choice of algorithm.

Because the purpose of the beneficiary level analysis is to develop performance measures for physicians practicing primarily in outpatient settings, we wanted an attribution system that reflected outpatient care of beneficiaries. Based on advice from the April 2010 Technical Expert Panel, we selected the "plurality of the beneficiary's non-hospital evaluation and management (NH-E&M) visits" as the attribution system for the beneficiary level analysis. We refer to this rule as the "plurality rule." Appendix 12 contains a detailed analysis comparing the results from this attribution to system to one based on PQRS reporting. Given the low rates of PQRS reporting for the time being, this plurality rule is far more defensible than using PQRS as the basis for a payment system in the near term.

II. The Units of Analysis – "Beneficiary Year" and Tax Identification Number (TIN)

We address the two main issues regarding the unit of analysis:

- Over how long a time period should the beneficiary's care be observed; and
- Should accountability be assigned to individual physicians or to groups of physicians.

A. The Window of Accountability – the "Beneficiary Year"

Since the Beneficiary Analysis holds a physician or physician group accountable for a fixed period of care for a beneficiary, we obviously need to decide how long that period of time should be. Admittedly, the length of time over which a beneficiary's cost and quality of care are assessed is somewhat arbitrary. However, health care costs exhibit wide variance, and so shorter observation periods introduce more random variation into a beneficiary's observed level of cost than would longer periods. Shorter observation windows also increase the probability that tests or screenings that should be performed on a regular basis might be performed as recommended, but fall outside the observation window. However, longer observation windows can drive a wedge between the physician's actual performance and the accompanying incentive payments by lengthening the time between the behavior and associated reward.

We chose a year as our observation period for the beneficiary level analysis, as a reasonable compromise between observation periods that are too short versus too long and as a natural fit to administrative cycles – for example, fee "updates" or raises in FFS Medicare currently are granted on an annual basis, so cost measurement is somewhat simplified. (We note, however, that extending the length of the observation to two years or constructing a rolling average cost measure – perhaps giving more weight to more recent years – would smooth out even more of the random variation in beneficiary-level cost data.)

B. Individual or Group Unit of Analysis

Individual physicians and other health professionals in the FFS Medicare program are identified by their National Provider Identifier or NPI. However, we chose the tax identification number or TIN as the unit of our analyses and the recommended organizational unit for incentive payment purposes. For several reasons.

First, our earliest analyses of performance measures made it clear that cost and quality analysis at the individual physician (NPI) level only would exacerbate the pervasive problem of small sample sizes. Individual physicians will have fewer attributed beneficiaries than groups of physicians (e.g., TINs). Second, we wanted to encourage coordination of care, and it made sense to start with the physicians who billed under the same TIN and thus had an organizational relationship of some kind with each other. Our system leaves the distribution of incentive payments to physicians within the TIN up to the TIN. Physicians who bill under more than one TIN could receive incentive payments under different TINs, but not for the same beneficiary under the plurality attribution rule, because the plurality rule results in unique pairings of beneficiaries and TINs.

III. Analytic Methods

In this section, we review analytic issues that have to be resolved before we can perform the cost and quality measurement activities that we are proposing. The main issues to be discussed are:

- Data for the Beneficiary Analysis.
- General issues in risk adjusting cost and quality performance measures.

A. Data for the Beneficiary Analysis

Early in the project, we were provided a sample of data on 2008 beneficiaries in Colorado, and some of our early analyses were based on that sample. However, the final beneficiary level analyses (development of our cost and quality measures and the combination of cost and quality measures) are based on 2008 and 2009 data from five states: California, Colorado, New Jersey, North Dakota, and Florida. These states were chosen by CMS to represent a mix of geographic locations and population, demographic, and sociodemographic characteristics.

The sampling frame was a 100 percent sample of beneficiaries who lived in these states. We then collected all the claims for each beneficiary incurred during the year. Some of the claims for these beneficiaries come from providers based outside of the five states, since some of their care will have been obtained elsewhere.

The smallest unit of observation in the beneficiary level analysis is a year's worth of claims experience for a beneficiary assigned to a particular TIN. *Data on the cost and quality of care for individual beneficiaries assigned to the TIN are aggregated to produce the TIN's annual performance measures*. Those annual performance measures become the basis for determining whether or not the TIN receives additional payments as rewards for better quality and lower cost.

We identified 41,369 TINs identified by the plurality rule in the 2008 five state data and 42,123 in the 2009 data. Because we are interested in the stability of our performance measures over time, we limited our analytic sample to the 36,405 TINs that appeared in both the 2008 and 2009 data.[31] The samples for the cost and quality analyses are presented in subsequent tables in this chapter.

Table 4.1 shows the distribution of the number of attributed beneficiaries per TIN. Some of our subsequent analyses in this chapter were limited to the 65 percent of TINs with 20 or more attributed beneficiaries.

Table 4.1: Distribution of attributed beneficiaries per TIN

Number of attributed beneficiaries per TIN	Number of TINs	Percent of TINs
0 to 19	12,793	35
20 to 99	12,653	35
100 to 999	10,494	29
1,000 to 9,999	449	1
10,000 or more	16	0
Total	36,405	100

B. General Issues in Risk Adjusting Cost and Quality Performance Measures

Any effort to measure physician performance must confront a series of issues concerning risk adjustment. The reasons are straightforward. The overall purpose of this project is to develop performance measures that will be an effective basis for value based modifiers to physician payments. In order to be effective, the performance measures must be fair, actionable and meaningful to physicians. Fairness generally implies that the measures should be adjusted for factors beyond the health care provider's control – i.e., factors "exogenous to the provider."

[31] Dropping the TINs that don't appear in both years doesn't present a problem for using the methods we propose in a payment system. All it would mean is that such a TIN would get measured for bonus payments one year, but not the other.

Cost and quality data often are risk adjusted in an attempt to improve the fairness of a payment system. Data are considered actionable if the provider on whom data are reported is capable of changing the value of the performance variable. "Actionable" can be a contentious issue because there may be quality or cost domains which a provider cannot influence easily in the short run, but (through third parties, such as other providers) might be able to influence in the long run. For example, changing the incentives faced by primary care physicians or hospital discharge planners might result in better post-discharge follow-up of the patient by providers other than the primary care physician. There is a tension between the variables one might adjust for in the pursuit of fairness *given the current FFS Medicare delivery system,* and the variables one might adjust for in an attempt to incent providers to take a more comprehensive interest in their patients' welfare.

There are two broad approaches to presenting risk adjusted data. Both involve regressing the variable of interest (e.g., cost, quality, or a combination of the two) on a set of explanatory variables representing factors that affect cost or quality, but are beyond the physician's control (i.e., exogenous factors).

In the first approach, the residuals from that regression represent the differences between actual value of the measure and the expected value of the measure, given the values of the exogenous variables for that unit of observation (e.g., a TIN). That approach is shown in the equation below:

$$Y_{ij} = X_{ij}\beta + u_{ij} \tag{4.1}$$

where:

Y_{ij} is the variable of interest (cost, quality or performance) for the i$^{\text{th}}$ beneficiary attributed to the j$^{\text{th}}$ TIN

X is a vector of exogenous variables, for which the physician is *not* held responsible

β is a vector of estimated coefficients, and

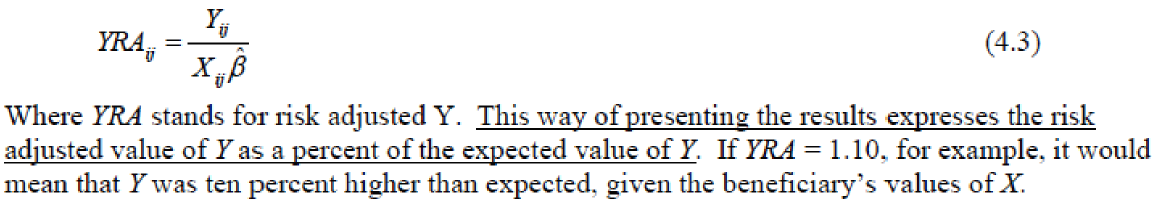
u is the error term, representing the costs for which the physician *is* held responsible.

Once the equation and the coefficients $\hat{\beta}$ are estimated, the residual can be computed as:

$$\hat{u}_{ij} = Y_{ij} - X_{ij}\hat{\beta} \quad (4.2)$$

$\hat{u}_{ij}$ can be positive or negative. In the case of cost, if $\hat{u}_{ij}$ is positive and equal to $500, it means that the TIN's actual cost for that beneficiary in that year was $500 above the cost that would be expected, given beneficiary's values of the X variables. Similarly, if $\hat{u}_{ij}$ is negative and equal to *negative* $500, it means that the TIN's actual cost for that beneficiary in that year was $500 *below* the cost that would be expected, given beneficiary's values of the X variables.

The second approach takes the results from the regression and forms a ratio:



$$YRA_{ij} = \frac{Y_{ij}}{X_{ij}\hat{\beta}} \qquad (4.3)$$

Where *YRA* stands for risk adjusted Y. This way of presenting the results expresses the risk adjusted value of *Y* as a percent of the expected value of *Y*. If *YRA* = 1.10, for example, it would mean that *Y* was ten percent higher than expected, given the beneficiary's values of *X*.

In the Beneficiary Analyses, we chose the first approach for two reasons. First, given Medicare's fiscal difficulties, we believe it is important to pay close attention to actual dollar amounts, rather than percentage deviations from expected values. If one type of patient or procedure is more expensive than another, that patient or procedure might deserve much closer attention for that reason alone, even if the percentage variations from expected values were less.

Second, the numerator of the right-hand side of equation 4.3 (Y_{ij}) when aggregated to the provider level (i.e., Y_j) often is subjected to further adjustment using methods referred to as random effects or shrinkage estimators, which are used to compensate for the fact that samples on which Y_j is based are small, leading to a large standard deviation of Y_j . Note that this approach is not novel: it is used to adjust the numerator of equation 4.3 in the estimation of hospital mortality and readmission data reported in Medicare's Hospital Compare system. When the samples on which Y_j is based are small, leading to a large standard deviation of Y_j , the shrinkage estimator produces a weighted average estimate of Y_j that combines the actual observed Y_j with the mean of *Y* across all providers in the sample.

The use of shrinkage estimators is fraught with conceptual and technical problems and has come under sharp criticism in the health services research literature (Silber, et al., 2010).[32]

We agree with Silber, et al.'s critique. Given that critique, the additional problems of shrinkage estimators, and the different problems presented by percentage cost differentials that mask the dollar amount of true cost differentials, we chose to use the residuals in equation 4.2 as our measures of risk adjusted *Y*.

[32] Silber, et al., are concerned that the jth hospital's characteristics, particularly the condition-specific volume of patients treated in the hospital, are omitted when computing the mean to which the jth hospital's mean should be shrunk. Silber, et al., maintain that low volume hospitals are known to have higher risk adjusted mortality rates (lower quality), yet the CMS shrinkage model shrinks the low volume hospital's mortality rate towards the mean for all hospitals, not the mean for low volume hospitals, thereby making low volume hospitals look artificially safer. Silber proposes shrinking hospital mortality rates to the volume-adjusted mean, rather than the grand mean for all hospitals. In fact, Silber's critique applies to *any* hospital characteristic that creates a distinct and somewhat stable group to which the jth hospital's mean more properly should be shrunk. But there is a second problem with shrinkage estimators. All shrinkage estimators involve an important element of uncertainty. The advantage of shrinkage estimators lies in the powerful result that they can be proven, theoretically, to reduce the mean squared error of predicted hospital mortality rates across three or more hospitals, *regardless of the hospitals' true means*. Uncertainty arises in practice because it is impossible to know the true mean of any hospital. (That's why the mean of the jth hospital has to be estimated in the first place.) So to evaluate the success of a shrinkage estimator or to compare one shrinkage estimator specification to another, the analyst must make a somewhat arbitrary assumption regarding the jth hospital's true mean. For example, one might evaluate the mean squared prediction error using the subsequent year's value of mortality for each hospital as an estimate of its "true" mean.

IV. Estimating Risk Adjusted Costs

In order to estimate the “cost” of attributed services, we need to consider two issues: certain special issues in risk adjusting costs, and questions about the specification of the cost equations.

A. Special Issues in Risk Adjusting Costs

The cost data available for this project consisted of Medicare paid claims for beneficiaries in our five state sample incurred during the calendar year 2008 or 2009, with the exception of Part D (outpatient drug) claims.[33] All 36,405 TINs had valid average cost data (per beneficiary per year). In the incentive payment system we propose, the cost data for 2008, for example, would be used to calculate the incentive payments that would be paid in 2009 – i.e., the incentive payments for a given year’s performance are actually paid out the following year. Our beneficiary level risk adjustment data consist of the standard demographic variables and the condition categories or CCs available on Medicare claims data (as are used in Medicare Advantage to risk adjust monthly payments to health plans). These are the X_{ij} variables in equation 4.1 above.

A remaining question, however, is what calendar year’s data should be used to calculate the risk adjusted 2008 cost data? Should the 2008 values of X_{ij} be used to adjust the 2008 cost data? That approach is referred to as *concurrent* adjustment, and was used in the CMS PGP Demonstration Project. Or should 2007 values of X_{ij} be used to risk adjust the 2008 data? That approach is referred to as *prospective* adjustment, and a variant of that approach is used to calculate payments to Medicare Advantage plans.[34] While it would seem obvious to use the 2008 X_{ij} to adjust the 2008 cost data, the concern is that the CCs are triggered by concurrent utilization, and thus physicians would be rewarded for higher utilization by giving them higher values of the risk adjustment variables. The prior year’s values of X_{ij} are somewhat more exogenous to the physician.

When we put this question to our Beneficiary TEP, they recommended that we use the concurrent (in the above example, 2008) hierarchical condition categories (HCCs) to risk adjust the beneficiary level 2008 data, *but* that we use the HCCs judiciously, paying particular attention to the incentives for provision of good primary care and efficient use of the hospital. That recommendation resulted in the following approach to risk adjustment in the beneficiary level approach.

1. We divided the claims data by grouping payer types into inpatient and outpatient claims.

[33] We did not have access to Part D (outpatient prescription drug) data for this project. While spending on outpatient drugs likely is correlated with spending on other Medicare-covered services, the addition of Part D data could move some relatively healthy beneficiaries from zero spending to some positive level of spending, but would be unlikely to affect attribution under the plurality rule.

[34] For MA payments, the beneficiary’s values of X_{ij} from the prior year are multiplied by dollar weights to arrive at the MA plan’s payment for the ith beneficiary in the current year.

a. The inpatient claims category consisted of institutional claims for hospitalizations.

b. The outpatient claims consisted of all other institutional and non-institutional claims (minus Part D claims), including physician/carrier, outpatient, durable medical equipment, home health, hospice, and skilled nursing facility. The Medicare payment amount was used as the cost variable, and all calendar year Medicare payments as grouped by the payer types were summed at the beneficiary-level for the cost equation calculations.

2. We risk adjusted the claims, using the regression approach described above, for all the demographic factors and recalculated HCCs.

3. For inpatient claims, the ambulatory care sensitive conditions (ACSs), such as asthma were excluded from the HCC recalculation for inpatient claims because we did not want to adjust inpatient claims for hospitalizations which could have been prevented with better ambulatory care.

Thus, TINs are not penalized for seeing asthmatic patients in the office. That is, asthma is included among the risk adjustors (the X_{ij} *)* in the outpatient version of equation 4.1. However, TINs are penalized for asthma patients who are hospitalized because asthma is excluded from the X_{ij} in the inpatient version of equation 4.2.[35]

B. What Specification Should be Used in Cost Equations?

Health expenditure data have some particular characteristics that require the analyst to pay special attention to the specification of the cost equations. Some background will help to clarify this.

As Buntin, et al. (2004) note, there typically are three problems with health expenditures data:

1. No negative observations;
2. A mass of observations at zero; and
3. A skewed distribution of positive observations.

To those, we add:

4. Misspecification of the functional form of the relationship between the explanatory and dependent variables.

There are two additional aspects of our analysis that warrant attention. First, as noted above, our risk adjustment system uses different variables to risk adjust the inpatient and outpatient equations, so we must have different equations for those types of costs. Second, we

[35] This approach is by no means perfect and even could have the opposite of the intended effect – making TINs with ACS admissions appear cheaper. Ideally, one would simply add the cost of ACS admissions into the TIN's risk-adjusted cost. However, the identification of a hospitalization as an ACS admission is not a perfect science when using administrative data because the ACS diagnosis may be listed as a co-morbidity rather than principle diagnosis or vice versa. The mapping of ACS conditions into the HCCs, resulting in two different sets of risk adjusters was seen as a second best alternative. This would be a fruitful area for future research.

are interested only in prediction of average costs at the TIN level, not the individual beneficiary level, which has the fortunate effect of reducing, to a degree, the influence of high cost outlier beneficiaries.

Detailed investigation of the varying specifications of cost equations were not a major focus of this project. We built on prior work in this area. Our experience with cost equations is consistent with that of other researchers (Buntin and Zaslavsky, 2004; Manning and Mullahy, 2001) who find that no single best approach to modeling health care cost data ensures accurate predictions. In fact, Buntin and Zaslavsky (2004) found that several models produced similar and fairly reliable predictions of costs in Medicare data.

Early in the project we compared two different approaches to estimating risk adjusted cost. We used the 2008 Colorado data for this analysis because those were the cost data we had available at the time. The unit of analysis in both approaches was the individual beneficiary. The two approaches also used the same risk adjustment variables. However, they differed in the estimation method.

In both approaches there was only one equation for outpatient cost equation. The dependent variable in the outpatient cost equation was the total outpatient cost for the beneficiary during the calendar year. The outpatient cost equation took the following form:

$$\textit{Outpatient Cost}_i = X_i\hat{\beta}_i^{OUTPATIENT} + \hat{u}_i^{OUTPATIENT} \qquad (4.4)$$

where "i" indexes the beneficiary

X_i = a vector of risk adjustment variables (HCCs plus demographic factors)

$\hat{\beta}_i^{OUTPATIENT}$ = a vector of estimated coefficients

$\hat{u}_i^{OUTPATIENT}$ = the residual outpatient cost after risk adjustment.

$\hat{u}_i^{OUTPATIENT}$ is the outpatient cost for which the physician will be held responsible. Notice that $\hat{u}_i^{OUTPATIENT} = \textit{Actual Outpatient Cost}_i - X_i\hat{\beta}_i^{OUTPATIENT}$. In other words, the physician is held responsible for the arithmetic difference between the beneficiary's actual outpatient cost and the beneficiary's predicted outpatient cost.

The same approach holds for the inpatient cost equation, but due to the large proportion of beneficiaries who had no inpatient costs, the inpatient cost model was estimated in two parts. The first part of the inpatient cost model was a logit equation that yielded the risk adjusted probability that the beneficiary had some inpatient costs (versus no inpatient costs). That equation took the form:

$$\textit{Probability that inpatient costs}_i > 0) = \frac{1}{1+e^{-X_i\hat{\gamma}}} \qquad (4.5)$$

The second part of the inpatient cost model was an equation that predicted the risk adjusted *level* of inpatient costs, based on the sample of beneficiaries who had some inpatient costs. That equation took the form:

$$Inpatient\ Cost_i\,|_{\text{Given some inpatient cost}} = X_i\hat{\beta}_i^{INPATIENT} + \hat{u}_i^{INPATIENT} \qquad (4.6)$$

To obtain the risk-adjusted *residual* expected cost for the ith beneficiary (the inpatient equivalent to $\hat{u}_i^{OUTPATIENT}$), we used equation 4.6 to predict the level of inpatient cost (given some inpatient cost) for all beneficiaries. Then we multiplied the result of equation 4.5 times the result of equation 4.6 to get the beneficiary's risk-adjusted *expected* inpatient cost as follows:

$$Expected\ Inpatient\ Cost_i = \left(\frac{1}{1+e^{-X_i\hat{\gamma}}}\right) \times \left(X_i\hat{\beta}_i^{INPATIENT}\right) \qquad (4.7)$$

Next, we subtracted that expected level of inpatient cost from the beneficiary's *actual* level of inpatient cost (which was be zero for many beneficiaries) to obtain the same type of risk adjusted residual for inpatient costs that we obtained for outpatient costs in equation 4.4:

$$\hat{u}_i^{INPATIENT} = Actual\ Inpatient\ Cost_i - Expected\ Inpatient\ Cost_i \qquad (4.8)$$

Next, we added $\hat{u}_i^{OUTPATIENT}$ and $\hat{u}_i^{INPATIENT}$ together to obtain the total risk adjusted residual cost for the ith beneficiary, for which the physician was to be held responsible. Finally, we added up the total risk adjusted residual costs for all beneficiaries attributed to a TIN by the plurality rule, and divided by the number of attributed beneficiaries to obtain the TIN's average risk adjusted (residual) cost. Positive numbers meant that the TIN's costs were above the risk adjusted expected cost. Negative numbers meant that the TIN's costs were below the risk adjusted expected cost.

Our two estimation approaches differ in the way equations 4.4 and 4.6 are estimated. In the first approach, equations 4.4 and 4.6 were estimated using ordinary least squares (OLS). In the second approach, they were estimated with a generalized linear model (glm) using the gamma family with a log-link function (Stata Version 11). Otherwise, the two approaches combined the results from equation 4.4 through 4.8 in an identical manner.

The results of from the two models are summarized in Table 4.2 and shown in Figures 4.1 and 4.2.[36] The sample was limited to TINs to which 20 or more beneficiaries were assigned. Table 4.3 shows that the OLS model reproduces both the mean and the minimum and maximum values of TIN-level average cost more accurately than the glm model.

[36] Full regression results can be found in Appendix 18, Table 5. Rsquareds in the cost regressions ranged from 0.37 to 0.41. The pseudo-Rsquared in the logit analysis of the probability of positive inpatient spending was 0.36. In the 5 state data used for the full analsyis, Rsquareds in the OLS cost regressions ranged from 0.19 to 0.28.

Table 4.2: Actual Average Cost at the TIN Level Versus Predictions from the OLS and GLM Models (2008 Colorado Data)

Variable	Mean	Standard deviation	Minimum	Maximum
Actual average cost	$7,592	$5,744	$546	$47,925
OLS predicted average cost	$7,828	$4,662	$1,636	$47,780
GLM predicted average	$9,179	$7,487	$2,491	$92,963

N=7,391 TINs

The results in Table 4.2 are illustrated graphically in the scatterplots in Figures 4.1 and 4.2. In both figures, actual average costs are shown on the vertical axis and predicted average costs are shown on the horizontal axis. The diagonal line shows the points (on the 45 degree line) where actual costs and predicted costs are equal. Risk adjusted costs are the difference between actual and predicted (expected) costs and thus the distribution of points around the diagonal line shows the distribution of risk adjusted costs (graphically, the vertical distance from each point to the 45-degree line). TINs represented by points above the diagonal line have higher than expected costs while points below the line have lower than expected costs.

Figures 4.1 and 4.2 show scatterplots of actual cost versus predicted cost from the OLS and glm models. If predicted costs were equal to actual costs, the scatter of points would like along the solid lines shown in the diagram. The results in Figures 4.1 and 4.2 suggest that for the majority of observations, the simple OLS model provides a better fit to the actual TIN-level data, because the actual points are more tightly clustered around the 45-degree line.

For very high cost practices, the OLS model under-predicts actual cost (the points are farther from the 45-degree line), but the glm model tends to over-predict costs for the majority of TINs (a much larger proportion of the data points are to the right of the 45-degree line). We chose to proceed with the OLS model, but again, we emphasize that there is nothing in our overall approach that precludes substitution of different specifications of the cost model.

Figure 4.1: Actual Average Cost at the TIN Level versus Predictions From the OLS Model

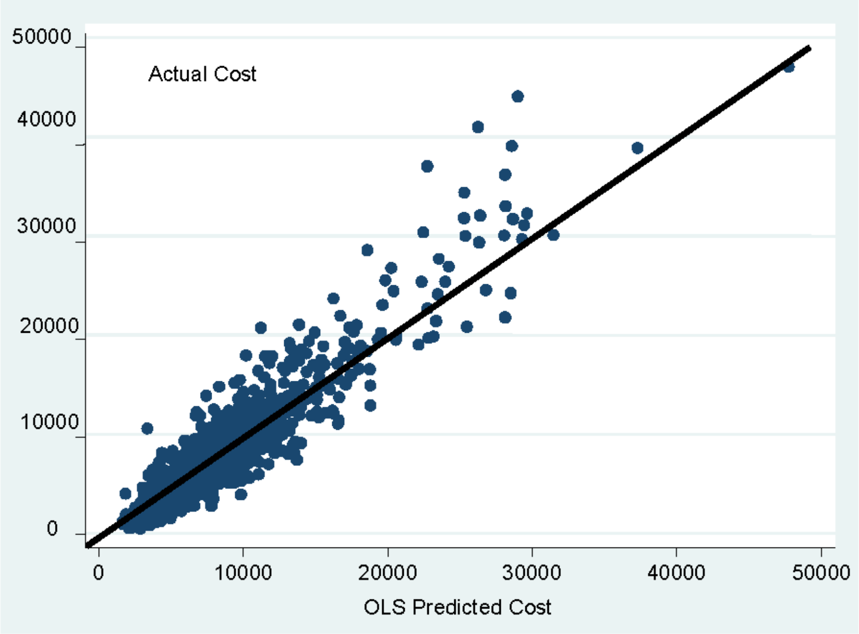
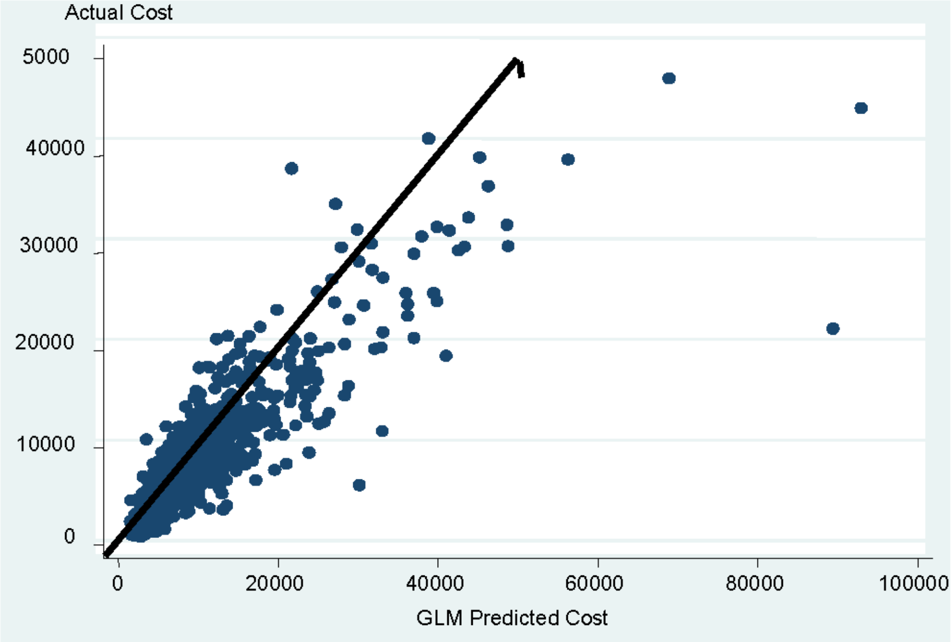


Figure 4.2: Actual Average Cost at the TIN level versus Predictions From the GLM Model



The discussion above summarizes the most important issues in estimating risk adjusted costs, and the choices we have made on them. We now turn to the other side of performance measurement: quality.

V. Quality measures

The ideal measures of health care quality in the FFS Medicare program would be measures of the beneficiary's health and functional status. Unfortunately, such measures currently are not available on the entire population, and certainly not in administrative data that would produce samples sizes required to compute reliable measures at the individual physician or TIN level. PQRS may improve that situation in the future. For now, it does not. So in order to develop a practical set of measures for a current payment system, *we rely on claims-computable quality measures*. Fortunately, the list of such measures has expanded considerably in recent years, to provide a useful and well rounded portfolio of measures of physician performance.

Our discussion of quality measures is detailed. It begins with three introductory sections:

A. The process for choosing and vetting quality measures – the beneficiary level technical expert panel

B. Overview of the quality measures selected out of the TEP process.

C. The special problem of choosing denominators for quality measures that apply to different subpopulations.

We then move to a conceptual discussion of three of the more interesting quality measures we have chosen: measures of “avoidable” utilization among our quality measures.

D. Avoidable emergency department visits.

E. Ambulatory care sensitive admissions.

F. Potentially preventable readmissions.

In the last section of our quality discussion, we briefly discuss our results for a set of screening and monitoring measures we selected.

A. The Beneficiary Level Technical Expert Panel (TEP)

As in the MS-DRG approach described in Chapter 3, we used an organized panel of experts in April 2010 to evaluate our methodology and to rate the quality measures we tentatively had chosen for the Beneficiary Analysis.

Before convening the TEP, we did extensive work to select attributable quality measures, through a process similar to what was done for the MS-DRG approach described earlier. 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 then developed sets of candidate measures from these sources to be considered for the Beneficiary Approach. We did analyses of the frequency of events after we chose the candidate measures. Then, we used the organized judgments of the TEP panel to refine and formally rate the relative usefulness of the measures.

The Beneficiary Approach measures were organized into the following broad categories for the TEP:

- Potentially avoidable utilization
- Screening measures
- Monitoring measures

Some of these measures represent measures that have been applied at more aggregate levels such as entire communities, but to our knowledge have not been used for physician performance measurement in public programs. The specific quality measures (the fifth category) were taken from those developed or endorsed by the National Committee for Quality Assurance (NCQA) or the National Quality Forum (NQF).

We used the two-day Beneficiary panel meeting in April 2010 to introduce TEP members to the details of this approach, to discuss the quality measures the project was considering, and to reach some understanding of the rating process that was to follow the meeting. After the TEP meetings, with suitable preparation, TEP members then were asked to rate the usefulness of potential quality measures for the Beneficiary approach. As in the case of the MS-DRG approach, the task of the Beneficiary panel rated 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 the Beneficiary results the average ratings on the usefulness of measures for the general constructs are much higher than the MS-DRG ratings, and the variance is considerably 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, PQRS 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 PQRS measures, 47 received a score of 70 or better. There are two ways to treat measures with low ratings. One option would have been 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 could have chosen 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 also because they 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 distribution of the ratings. That procedure still left 47 measures for analysis.

The TEP process was as useful for the Beneficiary analysis as it had been for the MS-DRG analysis. The TEP did a good job of winnowing our candidate cost and quality measures and approving our approach to combining cost and quality measures – the primary purpose for which it was used.

B. List of Quality Variables

Based on our review of literature and results from our TEP discussion, we selected *representatives* of three types of quality measures for development in our proposed system: potentially avoidable utilization, screening measures, and monitoring measures. Nothing in our system precludes the choice of a different or more expansive set of quality measures.

All of our quality measures are computable from Medicare claims data. With the exception of potentially preventable rehospitalizations, all can be computed using free, publicly available software (the algorithm for potentially preventable rehospitalizations is available commercially from 3M). The representative measures we have selected· are:

A. Potentially avoidable utilization measures:

d. Emergency department (ED) visits per beneficiary

e. Avoidable ED visits (Billings, Parikh and Mijanovich, 2000)

- Non-emergent ED visits per beneficiary
- Primary care treatable ED visits per beneficiary
- Preventable/avoidable ED visits per beneficiary

f. Ambulatory care sensitive (ACS) admissions per beneficiary (AHRQ, 2011)

g. Potentially preventable rehospitalizations per admission at 15, 16-30 and 31-60 days post-discharge (Goldfield, et al., 2007)

B. Screening measures:

   a. Colon cancer per eligible beneficiary (CMS, 2011)

   b. Breast cancer screening per eligible beneficiary (CMS, 2011)

C. Monitoring measures (CMS, 2011):

   a. Diabetes[37]
      - HbA1c screening per eligible beneficiary (CMS, 2011)
      - LDL-C screening per eligible beneficiary (CMS, 2011)
      - Medical attention for nephropathy per eligible beneficiary (CMS, 2011)

   b. Kidney disease
      - Lipid testing per eligible beneficiary (NQF, 2012)

   c. Cardiovascular disease
      - CVD-LDL testing per eligible beneficiary (NQF, 2012)

We defined the eligible beneficiaries for the avoidable utilization measures on the theory that all beneficiaries are at risk for avoidable ED visits and ambulatory care sensitive admissions, but only hospitalized beneficiaries are at risk for potentially preventable rehospitalizations. The number of eligible beneficiaries for each of the screening and monitoring measures were defined by the developers of each respective measure.

C. Choosing the Right Denominators for the Quality Measures

When we combine risk adjusted costs with quality measures to produce a measure of value, we compare costs and quality only for those beneficiaries who are eligible for the particular quality measures. We refer to the eligible beneficiaries as the "denominator population." The beneficiaries who actually receive the recommended care or who experience avoidable utilization comprise the "numerator" of the quality measure.

What is the proper denominator when comparing *different* quality measures? The same patient can be part of multiple denominators. For example, some female diabetic beneficiaries also will be found in the denominator population for breast cancer screening. All diabetic beneficiaries will be in the denominator population for avoidable emergency department visits and ambulatory care sensitive hospitalizations. Should the latter two quality measures be computed and compared only for diabetics or for the larger denominator populations who are at risk for those measures?

If the performance measures are to be the basis for an incentive payment system, then we recommend that each quality measure should be computed for *all* beneficiaries eligible for that quality measure. Our reasoning is based on the high rate of co-morbidities among Medicare beneficiaries. Note that there are desirable and undesirable forms of double-counting, and we need to establish clear standards for the two possibilities. For example, in our view, it is appropriate to reward a TIN for providing recommended monitoring of a diabetic beneficiary and then to reward the TIN again if the same diabetic beneficiary avoids unnecessary ED visits.

[37] Eye exams for diabetics and amputations were excluded due to very high and low frequencies, respectively.

However, it is not appropriate to reward the TIN for avoiding unnecessary ED visits for a diabetic beneficiary and then to reward the TIN again for avoiding the *same unnecessary ED visits* for the *same beneficiary* because that beneficiary also has hypertension, kidney disease or some other comorbid condition that places them in another denominator population. *Thus, we recommend that incentive payments should be structured so that each attributed beneficiary can contribute to multiple incentive payments for the different quality measures, but only one incentive payment per quality measure.*

D. Avoidable Emergency Department (ED) Visits

This section begins our discussion of three broad types of quality measures. We have two measures of emergency department (ED) utilization: the total number of ED visits per attributed beneficiaries and *avoidable* ED visits per attributed beneficiary. The purpose of the avoidable ED utilization measure is to assess the appropriateness of ED use among the beneficiaries assigned to a TIN. In Medicare claims data, each ED visit has at least one diagnosis code, but may have multiple diagnosis codes. We use an algorithm developed by Billings, Parikh and Mijanovich (2000) to assign a level of appropriateness to each ED visit. We refer to this algorithm as the Billings algorithm.

The Billings algorithm considers the 640 diagnosis codes that include most of the high-volume codes for ED visits. Based on clinical judgment, the Billings algorithm assigns each of the 640 codes a probability of four different categories of appropriateness.[38] The assigned probabilities sum to 1.0 across the four categories for any single diagnosis. For a more detailed discussion of the algorithm and the steps needed to implement it, see Appendix 13.

The four categories of appropriateness in the ED algorithm are:

1. Truly emergent – ED care needed, and not preventable/avoidable – Immediate care in an ED setting is needed, and the condition could not have been prevented/avoided with ambulatory care.

2. Emergent – primary care treatable – Care is needed within 12 hours, but care could be provided in a typical primary care setting. If care was needed within 12 hours but could have been treated in a primary care setting, it falls into a category of possibly inappropriate visits.

3. Emergent – ED care needed, but preventable /avoidable – Immediate care in an ED setting is needed, but the condition potentially could have been prevented or avoided with timely and effective ambulatory care. Here, the issue is not whether the patient should have gone to the ED, but rather whether better ambulatory care would have obviated the necessity for the ED visit.

4. Non-emergent – Cases where immediate care is not required within 12 hours. ED care may have been unnecessary if the patient could have waited more than 12 hours and sought care in the ambulatory care system.

[38] Diagnosis codes related to injuries, mental health, alcohol abuse, and substance abuse are not classified by the Billings algorithm and are excluded from our performance measure calculations.

We apply the Billings algorithm to each ED visit's diagnosis codes to determine the probabilities of each of the four categories of appropriateness for each diagnosis. The diagnosis with the highest score for "emergent-ED care needed: not preventable/avoidable (injuries included)" – i.e., the code that implies the likelihood of an *appropriate* ED visit – is identified as the diagnosis for the remainder of the analysis and is given that probability. This approach minimizes the likelihood that a TIN will be penalized for an ED visit that truly was appropriate. Using the single diagnosis derived in this manner, each visit is assigned probabilities for each level of appropriateness. This approach generates four probability "scores" for each ED visit. For any single visit the scores (probabilities) sum to one. (An example is included in Appendix 13.)

From these data, we develop two scores for each tax identification number (TIN):

1. The number of ED visits per beneficiary assigned to the TIN – The first score is simply the total number of ED visits of all kinds (appropriate and potentially avoidable) by beneficiaries assigned to the TIN divided by the total number of beneficiaries assigned to the TIN. Thus, the denominator population for all ED visits is the number of beneficiaries assigned to the TIN.

The average probability scores per beneficiary assigned to the TIN for the three types of *avoidable* ED utilization: non-emergent, emergent but primary care treatable, and emergent but preventable or avoidable – To arrive at this set of scores, we added the probability scores in each category of appropriateness across all the beneficiaries assigned to the TIN and divided by the total number of beneficiaries assigned to the TIN.[39]

The Billings algorithm is diagnosis specific and the assignment of appropriateness is ba*sed on a* single diagnosis, without considering the presence of additional diagnoses. In our application, the presence of an additional diagnosis that would have caused the visit to be rated more appropriate is not a problem because we choose the diagnosis that gives the visit the maximum probability of being rated appropriate.

The Billings algorithm represents an important step forward in measuring the appropriateness of ED utilization. Given the high cost and potentially negative quality implications of avoidable visits, the reasons to include such a measure in a value based payment system are compelling. But the algorithm has a number of shortcomings in its current form that were beyond the scope of this project to correct.

- The algorithm was based on the experience of an emergency department in New York City, and thus may not be entirely transferrable to other locations – e.g., less urban or rural settings.

- The algorithm does not consider the cumulative or interactive effects of multiple diagnoses.

[39] If we had divided by the number of ED visits rather than the number of attributed beneficiaries then the probabilities across all types of ED visits including truly emergent visits would sum to one. However, in that case, a zero could have been interpreted two ways: either then TIN had no beneficiaries with ED visits, or the TIN had no avoidable ED visits. Using our method avoids that problem and also acknowledges that every attributed beneficiary is at risk for both an ED visit and an avoidable ED visit.

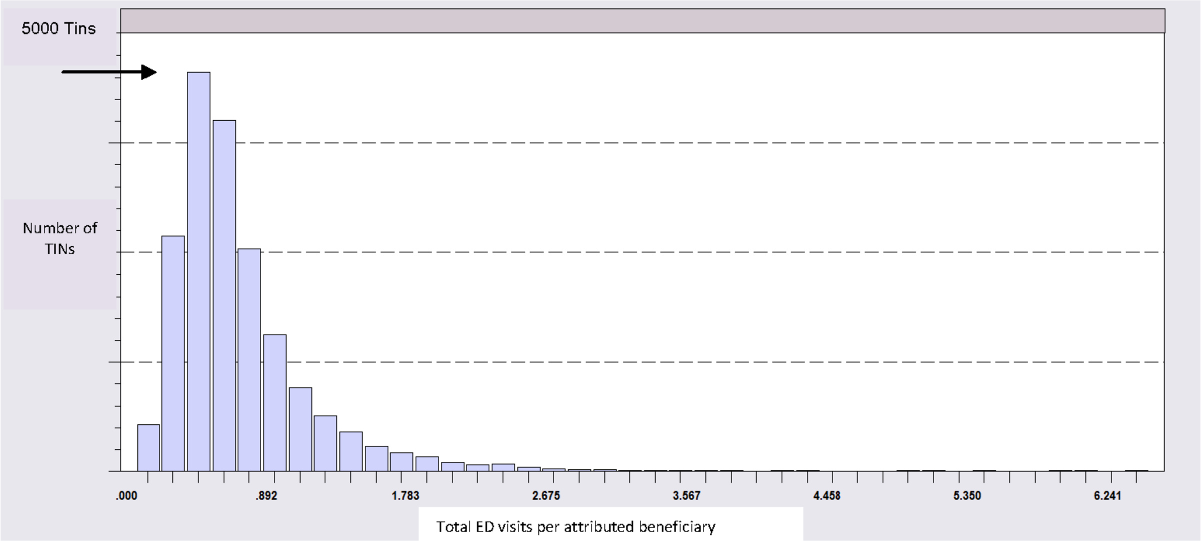
We recommend that CMS consider addressing these shortcomings if the Billings algorithm is to be incorporated into a value based purchasing system. Our use of the algorithm in its current form thus should be considered illustrative.

Our descriptive statistics for avoidable ED visit measures are shown in Table 4.3. The results in the table show that the TINs in our sample had an average of 0.62 ED visits per beneficiary attributed to the TIN. Said another way, on average, every group of 100 attributed beneficiaries had 62 ED visits. Among the avoidable ED visit measures, the highest average was for primary care treatable ED visits. On average, every group of 100 attributed beneficiaries had 11 primary care treatable ED visits, plus 8 non-emergent ED visits and 5 preventable ED visits.

Table 4.3: Descriptive statistics for avoidable ED visits (2008 five state data)

Measure	Mean	Standard Deviation	Number of TINs with 20 or more attributable beneficiaries
Total ED visits per beneficiary	0.62	0.82	23,612
Non-emergent ED visits per beneficiary	0.08	0.14	23,612
Primary care treatable ED visits per beneficiary	0.11	0.25	23,612
Preventable or avoidable ED visits per beneficiary	0.05	0.12	23,612

Figure 4.3: The Distribution of TINs by ED Visits per Attributed Beneficiary (2008 five state data – TINs with 20 or more attributed beneficiaries) N=23,612



E. Ambulatory Care Sensitive Admissions

Inpatient admissions are extremely costly. When inpatient admissions are possibly avoidable through better ambulatory care, such admissions may be indicative of poor quality care.

The Agency for Health Research and Quality (AHRQ) has developed an algorithm to identify admissions that are suspected of being avoidable because of the likelihood that they have resulted from poor ambulatory care. The ACS algorithm (AHRQ, 2011) is relatively easy to apply: the algorithm simply flags the ACS diagnosis codes of suspect admissions. The denominator for ACS admissions is all beneficiaries attributed to the TIN.

The ACS measure was reviewed by the Beneficiary TEP in 2010. The panel concluded that it would be both fair and advisable to hold the attributable physician responsible for the ambulatory care sensitive admissions – that is, if a physician's patient is hospitalized for an ACS condition, the physician's risk adjusted costs should rise.

Table 4.4: Ambulatory Care-Sensitive Medical Conditions

Immunization-preventable conditions	**Acute conditions**
Hemophilus meningitus (320.0)	Bacterial pneumonia (481, 482.2, 482.3, 482.9, 483, 485, 486)
Pertussis (033)	Cellulitis (681, 682, 683, 686)
Polio (045)	Dehydration (276.5)
Rheumatic fever (390, 391)	Gastroenteritis (558.8)
Tetanus (037)	Iron deficiency anemia (280.1, 280.8, 280.9)
	Kidney/urinary tract infection (590, 599.0, 599.9)
	Severe ear, nose, and throat infections (382, 462, 463, 465, 472.1, 20.01)
Chronic conditions	Untyped conditions
Asthma (493)	Angina (411.1, 411.8, 413)
Chronic obstructive pulmonary disease (491, 492, 494, 496, 466.0)	Congenital syphilis (090)
Congestive heart failure (428, 518.4)	Dental conditions (521-523, 525, 528)
Diabetes with ketoacidosis or hypersmolar coma (250.1-250.3)	Failure to thrive (783.4)
Diabetes with specified manifestations (250.8-250.9)	Nutritional deficiency (260-262, 268.0-268.1)
Diabetes without specified complications (250.0)	Pelvic inflammatory disease (614)
Grand mal seizure disorders (780.3, 345)	Tuberculosis (011-018)
Hypertension (401.0, 401.9, 402.0, 402.1, 402.9)	
Hypoglycemia (251.2)	

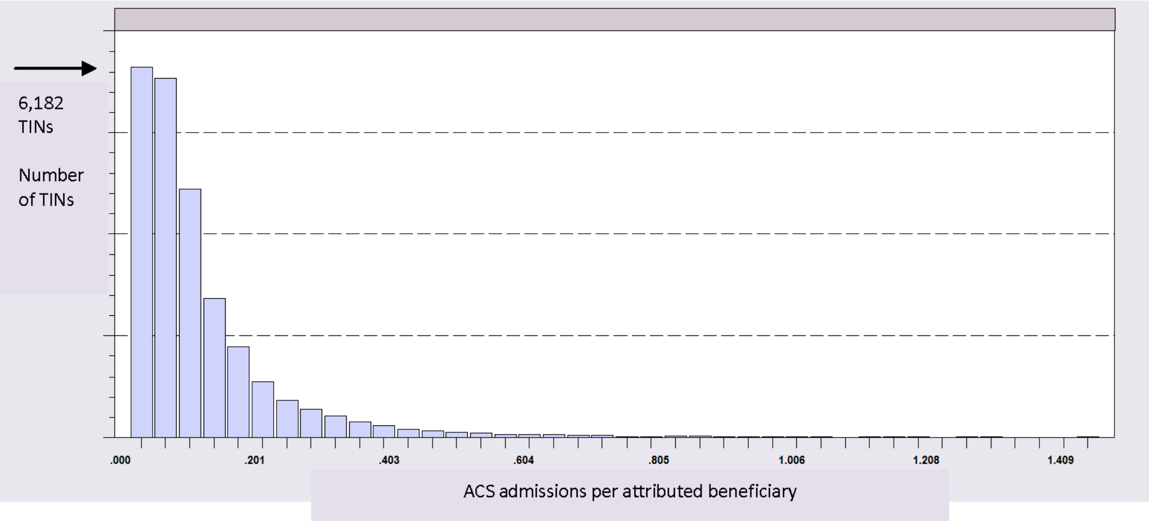
Table 4.5 shows the descriptive statistics for TIN-level ACS admissions. All of the TINs with 20 or more attributable beneficiaries had at least one beneficiary who had been hospitalized during the observation period. The mean of ACS admissions per beneficiary is interpreted as follows. For every group of 100 attributed beneficiaries, on average, there will be nine ACS admissions.

Table 4.5: Descriptive statistics for ACS admissions (2008 5 state data – TINs with more than 20 attributed beneficiaries)

Measure	Mean	Standard Deviation	Number of TINs with eligible beneficiaries and 20 or more total attributable beneficiaries
Ambulatory care sensitive admissions per beneficiary	0.09	0.16	23,612

Figure 4.4 shows the distribution of ACS admissions per attributed beneficiary across the TINs with 20 or more attributed beneficiaries. As in the case of ED visits, the distribution has a long right tail, indicating that some TINs have a large number of ACS admissions per beneficiary.

Figure 4.4: The Distribution TINS by ambulatory care sensitive (ACS) admissions per attributed beneficiary (2008 5 state data) N=23,612



F. Potentially Preventable Readmissions

One source of costly inpatient care is potentially preventable readmissions. Readmissions are especially interesting from a quality point of view because they may result from actions taken or omitted during/after the initial hospital stay. As noted by Goldfield, et al. (2008, p. 75),

> "A readmission may result from incomplete treatment or poor care of the underlying problem, or may reflect poor coordination of services at the time of discharge and afterwards, such as incomplete discharge planning and/ or inadequate access to care. Readmissions are important not only as quality screens, but also because they are expensive, consuming a disproportionate share of expenditures for inpatient hospital care. Readmissions can therefore focus attention on the critical time of an acute illness when the patient is in transition between inpatient and outpatient phases of treatment."

Goldfield and colleagues (2008) at 3M have developed an algorithm to evaluate readmissions to determine whether they are "potentially preventable." Each All Patient Refined DRG (APR

DRG) has four severity of illness levels that are used to assign to each discharge a probability that the discharge would result in a PPR. Those probabilities are used to calculate the expected rate of PPR readmissions to which an actual rate can be compared. 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.

The denominator for the potentially preventable rehospitalization (PPR) measure is the number of beneficiaries with at least one hospitalization attributed to the TIN. The descriptive statistics for the PPR measure are shown in Table 4.6. The means are interpreted as follows. In a sample of 100 attributed beneficiaries who have at least one hospitalization, on average there were 15 potentially preventable readmissions in the first 15 days following discharge, 9 in the next 15 day period and 11 in the third period ranging from 31 to 60 days post-discharge.

Table 4.6: Descriptive Statistics for Potentially Preventable Rehospitalizations (2008 five state data)

Measure	Mean	Standard Deviation	Number of TINs with eligible beneficiaries and 20 or more total attributable beneficiaries
Potentially preventable rehospitalizations (15 days post-discharge) per hospitalized beneficiary	0.15	0.28	23,464
Potentially preventable rehospitalizations (16-30 days post-discharge) per hospitalized beneficiary	0.09	0.17	23,464
Potentially preventable rehospitalizations (31-60 days post-discharge) per hospitalized beneficiary	0.10	0.17	23,464

Figure 4.5: The distribution of TINs by Potentially Preventable Rehospitalizations 0 to 15 days Post-discharge per Attributed Hospitalized Beneficiary (2008 five state data – TINs with more than 20 attributed and hospitalized beneficiaries) N=10,0026

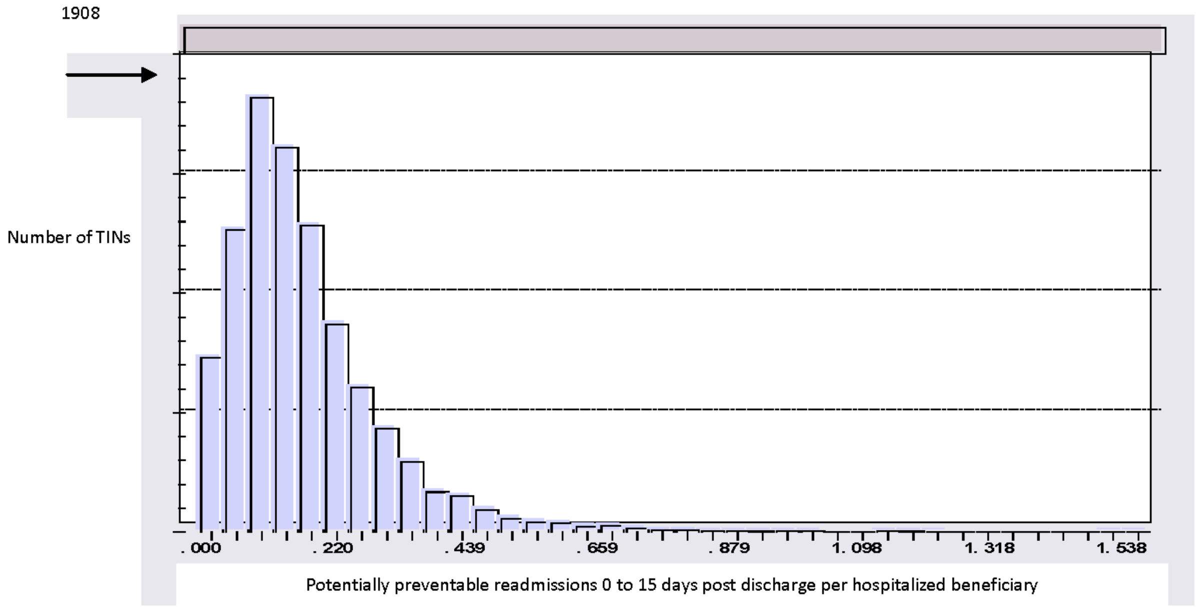


Table 4.7 shows the means and standard deviations of the screening and monitoring measures in the 2008 five state data. (Additional results on the screening and monitoring measures are presented in Appendix 15.) It is important to interpret the data carefully. If practice guidelines suggest that a monitoring measure should be performed every three years, for example, then the rates in Table 4.7 should be multiplied by three to obtain an estimate of compliance with the guidelines – on the assumption that the same rate would apply across the three years to the complete group of beneficiaries attributed to the TIN, to give a rate per beneficiary for the measure. *Our interest is less in whether physicians meet practice guidelines, (though that could well be a concern in a different context) but more in their performance relative to other physicians who are providing services to the same eligible beneficiaries*.

Table 4.7: Descriptive Statistics for the Screening and Monitoring Measures (2008 five state data)

Measure	Mean	Standard Deviation	Number of TINs with eligible beneficiaries and 20 or more attributable beneficiaries
All attributable beneficiaries			23,612
Colon cancer screening	0.15	0.12	23,508
Breast cancer screening	0.19	0.15	19,648
Diabetes measures			
HbA1c test	0.67	0.22	14,190
LDL-C test	0.68	0.24	14,190
Medical attention for nephropathy	0.59	0.21	14,190
Lipid test	0.55	0.25	5,344
CVD-LDL test	0.55	0.24	23,479

The smallest eligible population was beneficiaries with kidney disease, which affected the results for lipid testing. Sample sizes that small point to one of the main reasons why we have recommended separate incentive payments for different categories of quality measures (see discussion in Chapter 6). If the lipid measure had to be included *in a composite measure for all physicians*, the number of TINs for which the composite quality measure could be computed would be dramatically reduced, since a composite measure of necessity requires that each of the component measures can be computed for all of the TINs.

The pattern of correlations among the quality variables further supports that recommendation:

- Some of the measures are highly correlated with each other. Correlations among the three diabetic monitoring measures range from 0.50 to 0.81. Correlations among the avoidable ED visit and ACS admission measures range from 0.42 to 0.90.
- The potentially preventable readmission measures are moderately correlated with each other and with the ED and ACS measures (0.33 to 0.55).
- Otherwise, the correlations among groups of measures, or among the screening and monitoring variables, do not exceed 0.27.

These results have useful implications for the purposes of comparing cost and quality (i.e., for which measures we group together, before developing a combined, cost-quality measure). Based on these results, we grouped the ED and ACS admissions into one category; the potentially preventable readmission measures into a second category; and kept the remaining screening and monitoring measures as separate categories.

The very different cost implications associated with avoidable utilization versus screening and monitoring measures provides yet another argument for keeping the avoidable utilization measures together as a set and treating the screening and monitoring measures as separate categories. Again, we emphasize that there is nothing in our approach that precludes a different grouping of the quality measures.

Chapter 5

Combining Risk Adjusted Cost and Quality Measures into a Single "Efficiency" Score

The previous two chapters have described in detail how we propose to measure cost and quality for physicians practicing in inpatient and outpatient settings. In a payment system that sought to reward physician performance, it would be possible to keep cost and quality measures separate and somehow reward each separately. But the purpose of a value based payment system is to reward quality care that is relatively low cost – i.e., to point to more efficient care, rather than care that is only high quality or only low cost. To do that, we need some way to measure performance on cost and quality together. In this chapter, we lay out the ways we propose to do that for our MS-DRG and Beneficiary approaches in inpatient and outpatient settings, respectively.

We begin by discussing the relationship between cost and quality, and then discuss a series of popular approaches to combining measures of cost and quality. Our final two sections discuss how we propose to combine cost and quality measures in the MS-DRG approach ("tiering") and the Beneficiary approach (data envelopment analysis).

I. The Relationship of Cost and Quality

The combination of cost and quality addresses the simple question, "What are Medicare beneficiaries, the Medicare program, and taxpayers getting for the money spent on Medicare beneficiaries?"

Attempts to combine cost and quality into a single index of value or performance immediately confront the question, "What is the relationship between cost and quality?" In the standard economic literature, higher quality is assumed to imply higher cost. "You get what you pay for," is the operational rule of thumb in most economic activity. But that characterization of cost and quality does not easily accommodate the concepts of inefficiency or waste, or the production of services that actually might be harmful to consumers. Those concepts are not meaningful in the economists' perfectly competitive market. Wasteful firms would be driven from the market and there would be no demand for harmful services by perfectly informed consumers.

In health policy discussions, however, it is widely assumed that the health care sector exhibits considerable inefficiency and overuse of services which could expose consumers to dangerous environments (e.g., hospitals) and treatments that carry considerable risk. Certain types of preventive care (e.g., some immunizations and care in the first trimester of pregnancy) also have the potential to produce a positive return on investment – that is, to save more than a dollar for each dollar spent covering or delivering the service. Thus, there is some possibility that higher "quality" in health care will be associated with lower overall cost.

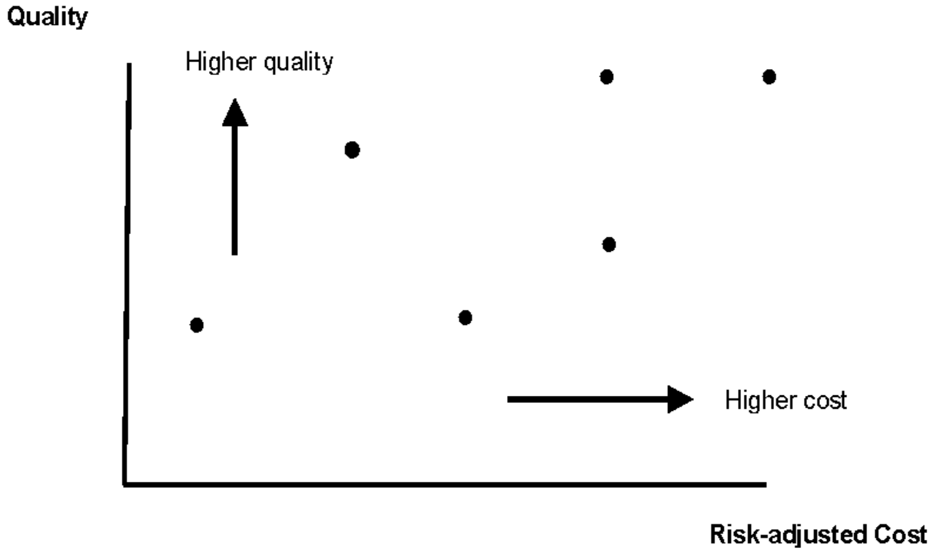
Among our quality measures, for example, reduction of avoidable utilization of the ED and hospital admissions certainly would be expected to reduce overall cost per episode or per beneficiary per year. In fact, reducing unnecessary care must *deterministically* result in lower cost. The same is not necessarily true of our other measures: e.g., screening and monitoring. It is possible that following practice guidelines in screening and monitoring could reduce expenditures per beneficiary per year, but that result is not inevitable.

No method of combing cost and quality solves a thorny problem CMS will face in those instances where higher quality care is more expensive. Suppose, for example, that more careful monitoring of diabetics produces better health outcomes, but at higher overall cost – i.e., better monitoring, while producing "value," does not have a positive return on investment. In that case, CMS will have to answer the policy question, "How much quality does the Medicare program want to pay for?" In the happy event that higher quality is less expensive, the question can be avoided.

A. Popular Approaches to Combining Cost and Quality

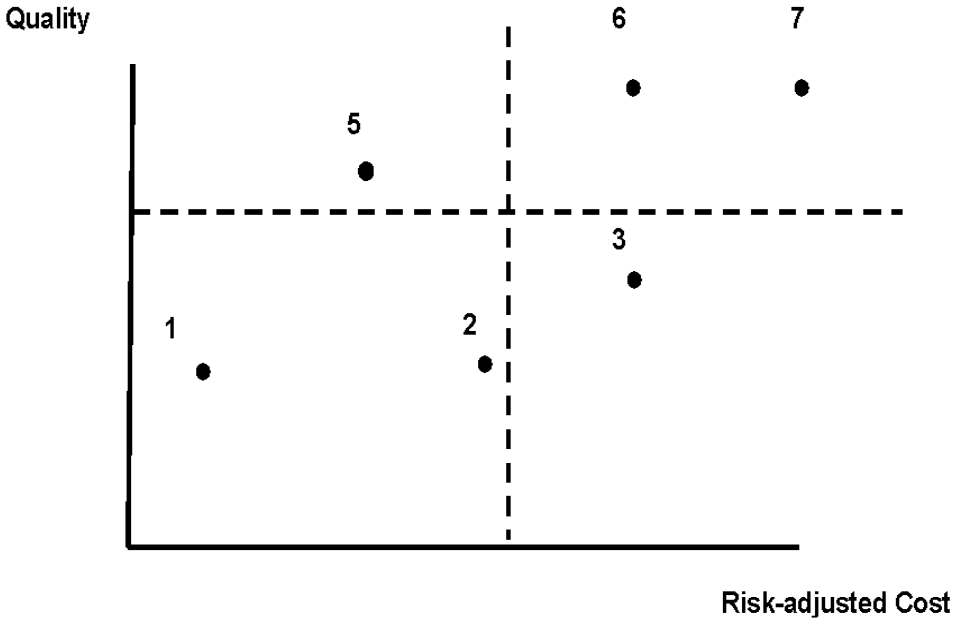
There are a number of ways to combine cost and quality scores. Consider the scatter plot of risk adjusted cost versus a single quality measure, shown in Figure 5.1. Each dot in the figure represents a physician or TIN. TINs that are further to the right on the horizontal axis have higher costs per beneficiary per year. TINs that are higher on the vertical axis have higher quality measures.

Figure 5.1: A Scatterplot of Quality versus Risk Adjusted Costs (each dot is a physician or TIN)



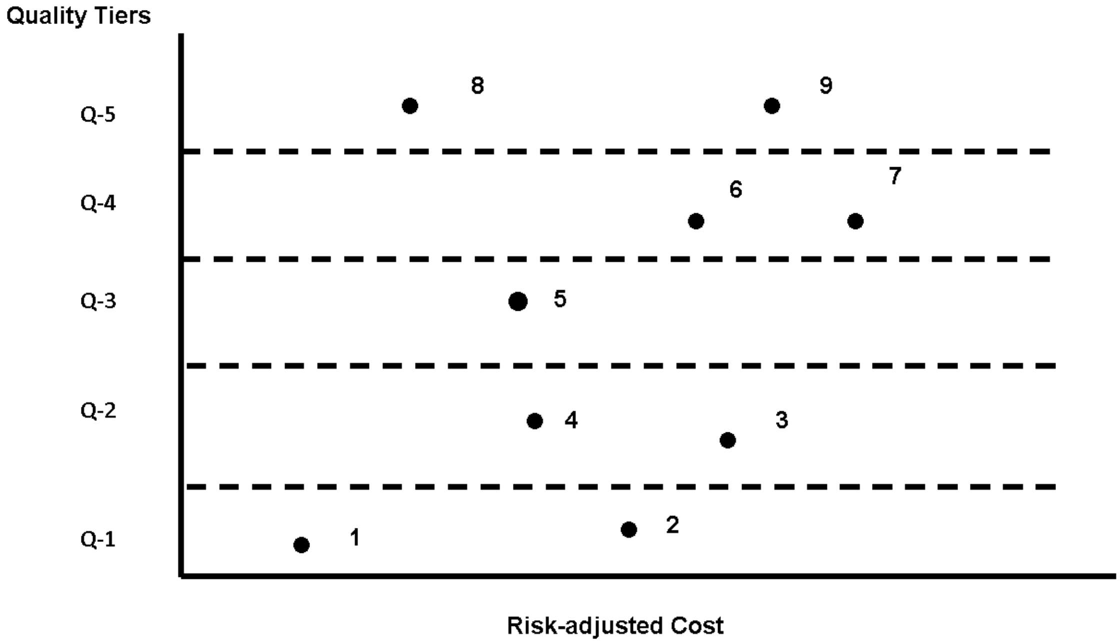
One approach to combining cost and quality scores is to choose points on the cost and quality axes that divide physicians into high and low quality and high and low cost, as shown in Figure 5.2. Physician 5 would represent a desirable combination of relatively low cost and relatively high quality.

Figure 5.2: A Quadrant Approach to combining Cost and Quality



A second approach is to create tiers of quality and compare cost within each tier, as shown in Figure 5.2(a). Within each quality tier, practices would be rewarded for being lower cost. Physicians 6 and 7 lie in the highest quality tier, but Physician 6 has lower cost than Physician 7. In some systems there might be only two tiers, one representing performance above a minimum quality benchmark, the other representing performance below the benchmark. For physicians above the benchmark, the reward is based entirely on risk adjusted cost. The MS-DRG approach uses this tiering method to combine cost and quality, as discussed in more detail in Section III.

Figure 5.2(a) A Tiering Approach to Combining Cost and Quality



Both the quadrant and the tiering approaches have advantages and disadvantages, and each can work well in particular situations. Both require the somewhat arbitrary choice of the number of tiers and the placement of the cut points, although "above or below the mean" certainly is an appealing starting point for the quadrant approach. The tiering approach does not attempt to distinguish between varying levels of quality within each tier and only compares risk adjusted cost.

B. A Third Alternative – Data Envelopment Analysis (DEA)

A third alternative for combining cost and quality data is data envelopment analysis, or DEA. Like the quadrant and tiering approaches, DEA compares the cost of physicians who are producing the same level of quality, but it achieves that result through a different method that avoids the need to establish arbitrary quadrants or tiers. DEA has been used in hundreds of studies of efficiency. The majority of the health-related studies focus on hospital efficiency (Mutter, et al. (2011)), but there have been applications to physician practices as well (Andes, et al., 2002).

The theory underlying this application of DEA is relatively straightforward. The physician who treats a Medicare beneficiary provides services in return for Medicare payments and beneficiary copayments. The physician strongly influences which services will be provided and thus the cost to Medicare for the care of each beneficiary.

In the DEA model, that cost is treated as the *input* to the production of outputs of interest. The outputs are quality measures, such as those discussed for the Beneficiary approach in Chapter 4. We emphasize that nothing about the DEA approach requires the use of those specific measures or any particular attribution rule, such as the plurality attribution rule we have chosen for that approach. CMS could choose any cost and quality measures and any attribution rule. The DEA approach even can accommodate multiple separate measures of cost, such as inpatient costs versus outpatient costs. *The only requirement is that there are some measures of cost and some measures of quality that are aggregated and attributed to some "accountable unit" or "provider."*

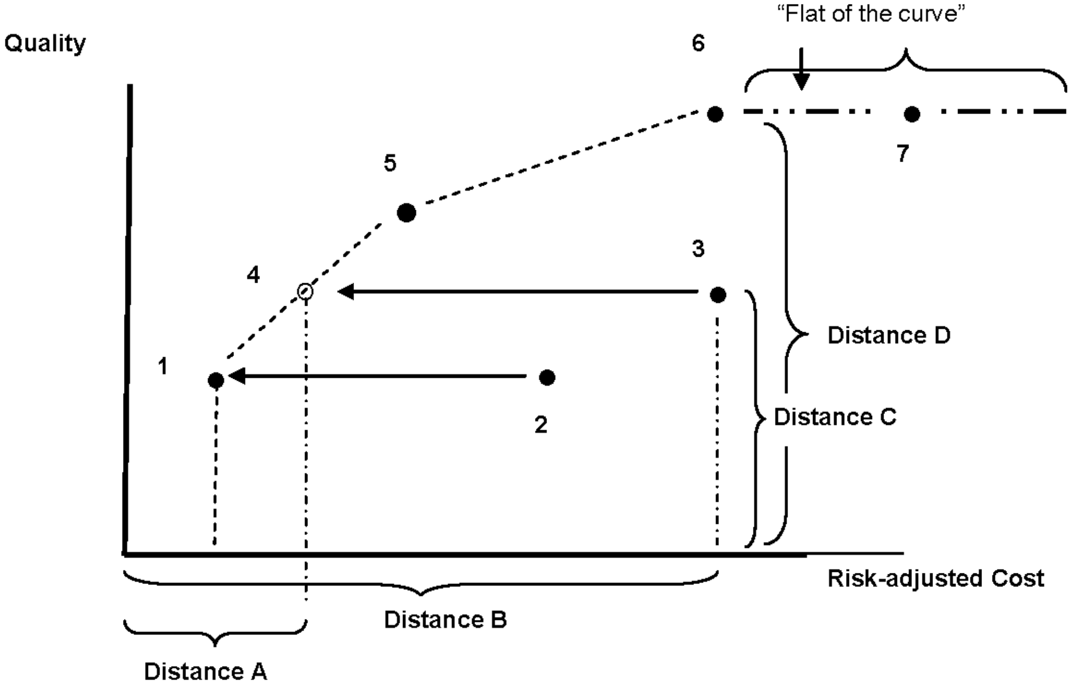
The question that DEA addresses is, "What 'outputs' of value are being produced from the 'input' of Medicare dollars?" In our application of DEA, the unit of analysis is the TIN. In the discussion that follows, we explain DEA in terms of the Beneficiary approach, though it is not in principle limited to that approach.

DEA can accommodate multiple inputs and multiple outputs. In our analysis, there is only one input – average risk adjusted Medicare expenditures per beneficiary per year – but there are multiple outputs – in the form of the quality measures discussed earlier.

A diagrammatic example of DEA analysis is shown in Figure 5.4. Suppose there is only one quality measure, shown on the vertical axis and one cost measure (risk-adjusted cost), shown on the horizontal axis. Thus, inputs are shown on the horizontal axis (less is better) and the single quality output on the vertical axis (more is better). A physician who is higher on the vertical axis has better quality for the same resources, and a physician who is farther to the left on the horizontal axis has lower cost for the same quality.

Assume there are seven physicians. If we take quality and cost data for all seven and plot them on the graph, the outermost points define a "frontier." Physicians 1, 5, and 6 are on that frontier – it is a cost-quality frontier in that there is no physician who has a better result on one variable (cost or quality), without having a worse result on the other. The frontier is defined not only by those physicians, but by the straight lines connecting them. DEA is a *linear* programming algorithm and thus the frontier is defined by the straight line segments connecting physicians on the frontier.

Figure 5.3 Diagrammatic Example of DEA Analysis



Physicians 1, 5 and 6 are producing the maximum amount of quality for their given levels of risk-adjusted cost. Physician 2 is not on the frontier because Physician 1 is producing the same level of quality at lower cost. Physician 3 is not on the frontier, either. Even though there is no *actual* physician producing the same level of quality at lower cost, an important assumption underlying DEA is that the frontier defined by Physicians 1, 4, 5, 6 and 7 is continuous,[40] and thus there is a hypothetical Physician 4 that could produce the same level of quality as Physician 3 but at the cost indicated by Physician 4.

DEA analysis produces two measures of efficiency: input efficiency and output efficiency.

- Input efficiency measures the degree to which a physician could reduce her risk-adjusted cost while maintaining the same level of quality. That is, for any given level of quality – a horizontal line on Figure 5.3 – what is the ratio of cost of a *frontier* physician to the cost of a specific *actual* physician? Physician 3, for example, has the same quality as (hypothetical) Physician 4. Physician 3's input efficiency is defined as the ratio of Physician 4's cost to Physician 3's cost, or Distance A divided by Distance B in the figure.

  Thus, physicians on the frontier have an input efficiency of 1.0, because, by the

[40] A more complete description of data envelopment analysis and empirical results from the 2008 Colorado data can be found in Appendix 14.

definition of the frontier, there is no physician farther to the left along a given horizontal line defining a particular level of quality.

- Output efficiency measures the degree to which a physician could increase quality, holding risk-adjusted cost constant. That is, for any given level of risk-adjusted cost – a vertical line on Figure 5.3 – what is the ratio of the quality of an *actual* physician to the level of quality of a *frontier* physician? Physician 3, for example, has the same cost as Physician 6 in Figure 1. Physician 3's output efficiency is defined as the ratio of Physician 3's quality to Physician 6's quality, or Distance C divided by Distance D in the figure.

Thus, input efficiency measures the degree to which costs, in principle, could be reduced without reducing the level of quality. Output efficiency measures the degree to which quality, in principle, could be increased without increasing costs (where the slope of the frontier is positive).

In the application we propose, input efficiency clearly is the preferred measure of a physician's performance. The reason is that output efficiency cannot distinguish between physicians 6 and 7 in Figure 5.3. For both physicians, output efficiency is equal to 1.0 because they both are on the upper part of the frontier (each of them is getting as much quality as is possible at their given level of cost – that's what it means for them to be on the frontier). However, input efficiency is greater for Physician 6 than Physician 7 because both are producing the same level of quality, but Physician 6 has lower risk-adjusted cost. Physician 7 is said to be on the "flat of the curve" with respect to quality – spending more money but producing no higher quality than Physician 6.

The graphical example discussed above involves only one input (cost) and one output (one quality measure). In the case of one input and one output, input efficiency for the s^{th} physician can be computed using high-school algebra, as it is simply the intersection of the horizontal line extending left from the physician's point in the cost/quality space to the frontier line. One simply writes down the equation for the line that connects Practices 1 and 5, substitutes Practice 3's value of quality, and solves for risk-adjusted cost, which is Distance A.

However, many health conditions such as diabetes might have multiple quality measures. One of the important advantages of the DEA model over many other models (e.g., stochastic frontier models) is that it is not limited to one input and one output. There can be multiple inputs and multiple outputs. In this application, that means that there can be multiple cost and quality measures – that is, we can make policy decisions about which cost and quality measures we think should be used to measure the performance of physicians, and the DEA model will permit us to combine the inputs or outputs we have chosen.[41] When a second quality measure is added, the frontier becomes a two-dimensional plane. More quality measures create a multi-dimensional output surface, but the principle is the same.

Even input efficiency poses an important challenge, however, because *input efficiency makes no distinction between high and low levels of quality – a physician can be rated highly efficient by producing either low or high quality care at low cost relative to other physicians at*

[41] The LIMDEP computer program that we are using for the analysis in this project currently limits the number of inputs and outputs to a total of 19 variables. That is a limitation of the program, not the method.

the same level of quality. Thus, DEA, alone, does not answer the question posed above, "How much quality is CMS willing to pay for?" In Figure 5.3 above, notice that Physician 1 is much cheaper than Physician 6. Is the additional, efficiently produced quality of Physician 6 worth the additional cost?

This is a difficult issue to resolve in principle. As it happens, however, the results we present later in this chapter suggest that the issue perhaps can be avoided: specifically, our results suggest that physicians with higher input efficiency scores tend to have both lower costs *and* higher quality than physicians with lower efficiency scores.

Dropping or adding a quality measure could change the rank of a physician relative to other physicians in the peer group. (Not surprisingly, if we dropped or added a measure on which a physician did particularly well or poorly, the physician's relative performance would change.) However, *linear* transformations of any quality measure for all physicians will not change the relative rank of the physician. Thus, for example, standardizing a quality measure by subtracting its mean and dividing by its standard deviation would not affect the results of DEA analysis. "Importance weights" derived, for example, by weighting each quality measure by the total number of beneficiaries affected by the measure (as opposed to the total number of attributed beneficiaries for each physician) also would have no effect on the results from DEA analysis.[42]

Nonetheless, there are ways that CMS could give more weight to individual quality measures. CMS could create composite measures prior to the DEA analysis that give greater weight to some measures. For example, if CMS wanted to give greater weight to diabetes care rather than colon cancer screening, CMS could create a composite measure that did that, then run DEA on the composite measure. Alternatively, CMS could establish minimum floor levels of quality for some measures, denying payment updates to providers who did not meet the minimum quality levels, notwithstanding their input efficiency scores. Conversely, CMS could give bonus payments to providers who scored well on individual quality measures, in addition to, or even regardless of, the provider's input efficiency score from the DEA analysis.

II. Combining Cost and Quality in the MS-DRG Approach

With the background provided above, we can now move forward to describe the particular methods we propose to use, beginning with the MS-DRG approach. The methodology for combining cost and quality in the MS-DRG approach is based on a "tiering" of the quality results among TINs that have comparable proportions of the episode cost (for reasons we will explain). We then present illustrative results showing how this approach works in application.

A. Using Tiering for Combining Cost and Quality for the MS-DRG Approach

In the MS-DRG approach we attribute risk adjusted costs of 30 and 60 day episodes to TINs based on a proportion rule. The TIN's responsibility for an episode depends on the proportion of its Part B billing during the initial hospitalization. However, all of the quality of care associated with the episode is attributed to any TIN who provided care during the initial

[42] However, transformations that differ from one physician to another such as shrinkage estimators (discussed in Chapter 4) will change the relative rankings.

hospitalization for the episode. We use this alternative rule for attributing quality because TINs providing care for an episode are jointly involved in producing quality for the episode.

The proportionate rule for attributing cost to TINs in the MS-DRG approach means that all comparisons of cost and quality must be done *within* peer groups based on the mean proportion of costs attributed to the TIN – put differently, TINs are only compared to other TINs with comparable proportions of attributed costs. Otherwise TINs with a lower mean proportion of episode cost would seem to be more efficient – as they would be producing 100 percent of an episode's quality (since all of the quality is attributed) with a fraction of the episode's cost attributed to them. Again, by using differing attribution rules for cost and quality, we do not view TINs producing output quality using input costs. Hence we do not use DEA for combining costs and quality in the MS-DRG approach.

The risk adjusted episode costs attributed to TINs can be made more comparable across TIN's by separating the TINs into peer groups – based on the mean episode cost proportion that is attributed to them. We group TINs first into five peer groups based on the mean episode proportion of cost attributed to them. TINs with 0-20%, 20-40%, 40-60%, 60-80% and 80-100% of episode responsibility respectively were grouped together into five peer groups. In this context, the peer groups represent workably "comparable proportions" of mean attributed episode proportion, for purposes of performance assessment, as they are narrow enough to permit us to make meaningful comparisons of the variation in quality among TINs within each peer group.

The tiering approach to combining costs and quality within a peer group of TINs is shown in Figure 5.3 above and proceeds in three steps:

- First, we establish cost peer groups as described above, based on the mean episode proportion of cost attributed to them.

- Second, *within each episode-proportion peer-group*, we order TINs into tiers based on their mean risk adjusted composite quality score. The number of tiers chosen is arbitrary. TINs could be ordered into quartiles, quintiles or even deciles based on their mean composite quality score. In Figure 5.1, TINs are grouped into quintiles based on their quality score. TINs belonging to quality tier Q-1 have the lowest quality composite scores while those in tier Q-5 have the highest quality score. We do not attempt to distinguish the varying levels of quality within each tier, because quality here is represented as a composite quality score, which could be achieved in a variety of ways.

- We then compare the TINs' average risk-adjusted episode costs within each quality tier. In the highest quality tier Q-5, TIN-8 is lower cost and hence more efficient than TIN-9. TINs 6 and 5 are lower cost compared to TIN-9, but also provide lower quality of care compared to TIN-9.

In summary, in the tiering approach to combining costs and quality, TINs would be proportionately rewarded for (i) being in a peer group that provides greater proportion of the care during an episode, (ii) belonging to a higher quality tier within a peer group, and (iii) being lower cost within a quality tier.

B. Tiering Results for 30DAY Episodes of CHF and Hip Replacement

To clarify how the tiering approach might work, we work through a set of results for combining costs and quality using the multistate data for 30-day episodes of two selected MS-DRGs: CHF and hip replacement.

Even before combining costs and quality using the tiering approach, it is useful to examine how the two vary across TIN peer groups, shown in Tables 5.1 (a) and (b) for CHF and hip replacement, respectively:

- Lowest quintile of mean episode proportion: Approximately 43 percent of CHF TINs and 74 percent of hip replacement TINs are in the lowest episode proportion peer group – i.e., they have 0-20% of the episode proportion attributed to them.

- Highest quintile: Only 3% of CHF TINs and 4% of hip replacement TINs are in the highest episode proportion peer group with 80-100% episode responsibility.

The variation in risk adjusted episode costs increases across peer groups with a greater proportion of responsibility for the episode – there is a higher variation in the higher peer groups. Thus, the standard deviation in risk adjusted episode costs for the lowest proportion peer group for CHF is approximately $1,200, but is$12,350 for the highest peer group. For hip replacement, the standard deviation in risk adjusted episode costs for the lowest proportion peer group is approximately $900, but is$4,100 for the highest peer group.   Peer grouping TINs allows us to separate TINs by their differing levels of responsibility for episodes, reflected in differing variation in risk adjusted costs. TINs in the higher peer groups with greater responsibility for the episode, are also responsible for most of the variation in risk adjusted episode costs. Hence incentive payments should be structured to offer greater rewards to such TINs compared to those in lower peer groups. This will be discussed in more detail in Chapter 6.

Table 5.1 (a): CHF: Descriptive statistics for TIN Episode Proportion Peer Groups

TIN Episode Proportion Peer Group	Number of TINs (Percent of TINs)	TINs Average Episode Proportion: MIN-MAX	TINs Average Episode Proportion: Mean (Std)	TIN's Average Risk Adjusted 30-day Episode Cost: Mean (Std)	TIN's Average Risk Adjusted Composite Quality Score for 30-day Episode: Mean (Std)
1	9,007 (43%)	0 - 0.2	0.11 (0.05)	$146 ($1,207)	51 (14)
2	7,266 (34%)	0.2 - 0.4	0.25 (0.03)	$128 ($2,053)	50 (12)
3	3,155 (15%)	0.4 - 0.6	0.48 (0.06)	$153 below expected ($3,146)	48 (12)
4	1,018 (5%)	0.6 - 0.8	0.68 (0.06)	$420 below expected ($4,453)	45 (12)
5	674 (3%)	0.8 – 1.0	0.91 (0.06)	$57 ($12,351)	40 (12)

Table 5.1 (b): CHF: Risk Adjusted Composite Quality Score within Quality Quintile Tiers across TIN Episode Proportion Peer Groups

TIN Episode Proportion Peer Group	**Risk Adjusted** Composite Quality Score in Quality Quintile 1: MIN-MAX	**Risk Adjusted** Composite Quality Score in Quality Quintile 2: MIN - MAX	**Risk Adjusted** Composite Quality Score in Quality Quintile 3: MIN - MAX	**Risk Adjusted** Composite Quality Score in Quality Quintile 4:MIN -MAX	**Risk Adjusted** Composite Quality Score in Quality Quintile 5: MIN - MAX
1	9 - 39	39 - 47	47 - 54	54 - 64	64 - 88
2	12 - 40	40 - 47	47 - 53	53 - 61	61 - 84
3	9 - 38	38 - 45	45 - 50	51 - 58	58 - 86
4	18 - 35	35 - 42	42 - 48	48 - 55	55 - 87
5	12 - 30	30 - 36	36 - 42	42 - 50	50 - 85

Table 5.2 (a): Hip Replacement: Descriptive statistics for TIN Episode Proportion Peer Groups

TIN Episode Proportion Peer Group	Number of TINs (Percent of TINs)	TIN's Average Episode Proportion: MIN-MAX	TIN's Average Episode Proportion: Mean (Std)	TIN's Average Risk Adjusted 30-day Episode Cost: Mean (Std)	TIN's Average Risk Adjusted Composite Quality Score for 30-day Episode: Mean (Std)
1	7,487 (74%)	0 - 0.2	0.09 (0.06)	$88 ($906)	50 (12)
2	1,097 (11%)	0.2 - 0.4	0.26 (0.05)	$170 ($2,887)	49 (11)
3	335 (3%)	0.4 - 0.6	0.51 (0.06)	$160 ($3,654)	49 (11)
4	832 (8%)	0.6 - 0.8	0.71 (0.05)	$162 ($2,876)	48 (9)
5	371 (4%)	0.8 - 1	0.89 (0.07)	$509 below expected ($4,161)	45 (9)

Table 5.2 (b): Hip Replacement: Risk Adjusted Composite Quality Score within Quality Quintile Tiers across TIN Episode Proportion Peer Groups

TIN Episode Proportion Peer Group	Risk Adjusted Composite Quality Score in Quality Quintile 1: MIN-MAX	Risk Adjusted Composite Quality Score in Quality Quintile 2: MIN -MAX	Risk Adjusted Composite Quality Score in Quality Quintile 3: MIN -MAX	Risk Adjusted Composite Quality Score in Quality Quintile 4: MIN -MAX	Risk Adjusted Composite Quality Score in Quality Quintile 5: MIN -MAX
1	14 - 40	40 - 46	46 - 53	53 - 61	61 - 80
2	17 - 40	40 - 45	45 - 52	52 - 60	60 - 79
3	21 - 39	39 - 45	45 - 51	51 - 58	58 - 78
4	17 - 40	40 - 45	45 - 50	50 - 55	55 - 75
5	20 - 38	38 - 42	42 - 46	46 - 52	52-75

While the variation in TIN's risk adjusted episode *costs* increases across peer groups, there is much less variation in quality across the episode proportion peer groups: the variation in risk adjusted composite quality scores for the TINs is more or less similar across the peer groups for both CHF and hip replacement.

The variations in risk adjusted episode costs and quality across the episode proportion peer groups are in line with what should be expected *based on the differing attribution rules for the two*. TINs are attributed costs according to the proportion of their billing. Hence TINs with lower proportion of billing (Peer Groups 1 and 2) have lower variation in cost compared to TINs with higher proportion of billing (Peer Groups 4 and 5). By contrast, the entire episode's quality is attributed to each TIN (irrespective of the proportion TIN's billing for the episode). As a result, TINs in all of the peer groups tend to have more similar composite quality scores.

Tables 5.1 (b) and 5.2 (b) above, show the risk adjusted composite quality score range within quality quintile tiers across the episode proportion peer groups for CHF and hip replacement, respectively. TINs in Quintile 1 have the lowest average risk adjusted composite quality scores, while those in Quintile 5 have the highest risk adjusted composite quality score. The upper and lower bounds of the risk adjusted composite quality scores for the five quality tiers are more or less similar, but not exactly uniform across the five episode proportion peer groups. But as we are more interested in discerning *relative differences in quality within a peer group*, rather than differences in quality across different peer groups, quality tiers can be used to compare the performance of TINs within a peer group. That is, while these differences in variation are descriptively interesting, they do not complicate the performance comparison that is the point of the peer grouping in the first place.

Figure 5.4 (a): CHF: Distribution of TIN's Average Risk Adjusted 30-day Episode Costs across Risk Adjusted Composite Quality Score Tiers for TINs in Highest Episode Proportion Peer Group (Number of TINs=674)

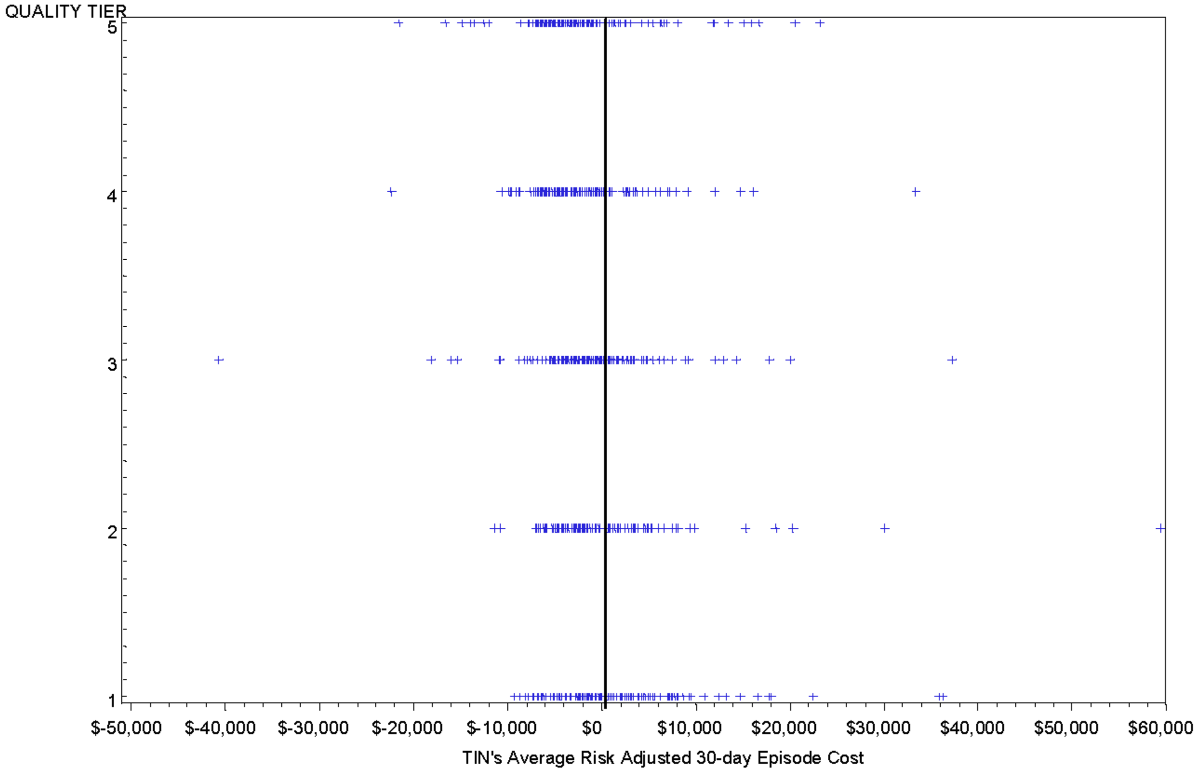
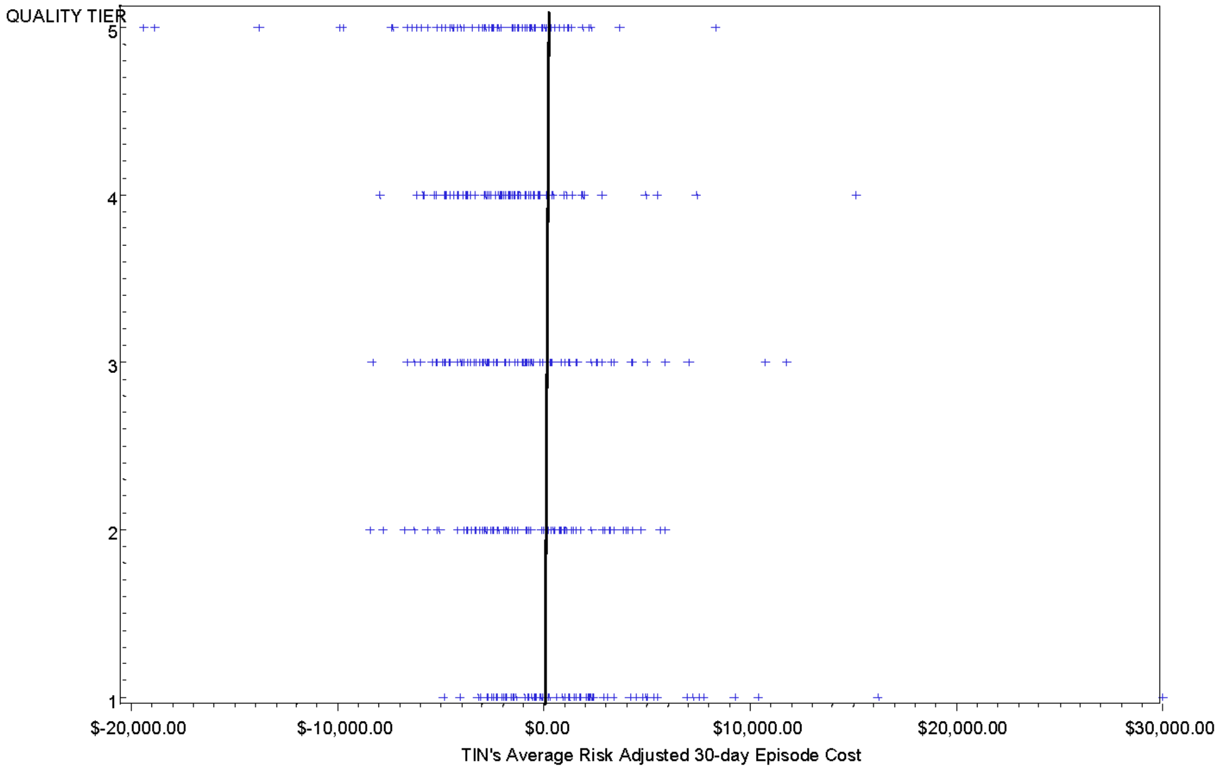


Figure 5.4 (b): Hip Replacement: Distribution of TIN's Average Risk Adjusted 30-day Episode Costs across Risk Adjusted Composite Quality Score Tiers for TINs in Highest Episode Proportion Peer Group (Number of TINs=371)



Figures 5.4 (a) and (b) above show the distribution of TINs' risk adjusted costs for 30-day episodes across quality score quintiles for TINs in the highest episode proportion peer group CHF and hip replacement respectively. TINs in this peer group have 80-100% of responsibility of the episode attributed to them. TINs in the higher quality tiers have better risk adjusted composite quality scores than those in lower tiers, as discussed earlier. For easier interpretation, a line is drawn at $0 expected cost. TINs with lower expected costs are more efficient than those with higher expected costs, within a quality tier (that is the point of the tiering of the TINs). TINs in quality tier 5 with the lowest cost are the best performers while those in tier 1 with the highest costs are the poorest performers. The key message of both figures is straightforward: *in both figures, there are more TINs below the expected cost in the highest quality tier compared to other quality tiers*. For example, in case of hip replacement there are more TINs below the expected cost in the highest quality tier compared to the other tiers.

This suggests that the relation between risk adjusted quality and risk adjusted episode costs might be linear. As most of the quality measures forming the composite are "avoidable" events – like mortality, readmissions, and avoidable ED visits, which are directly associated with higher episode costs – this linear association between quality and risk adjusted costs is not surprising. Poorer quality as a result of more "avoidables" invariably results in higher episode costs.

Another way we can make this relationship clear is to compare risk adjusted costs across TINs in each risk adjusted quality tier. To do that, we pooled TINs based on their cost z-score, as a way of standardizing differences in risk adjusted cost for comparison. The cost z-score for TINs in a peer group was calculated based on the mean and variance in risk adjusted cost for all the TINs in the peer group. TINs with cost z-scores below zero are below the mean cost for the peer group, and those with z-score above zero are above the mean cost of their peers.

Table 5.3 (a) shows the matrix of TINs across risk adjusted quality tiers for CHF. by their average risk adjusted 30-day episode cost z-score TINs belonging to the peer group with the highest episode responsibility (80-100% of episode). Table 5.3 (b) shows similar results for hip fracture.

Each quality quintile tier has 20% of the TINs in the peer group. TINs in quality tier 5 provide the highest quality, while TINs in quality tier 1 provide the lowest quality. There are fewer TINs in quality tier 5 that are above average cost of their peers (5% of 20% for CHF; 7% of 20% for hip replacement) than quality tier 1 (10% of 20% for CHF; 14% of 20% for hip replacement). Also, there are more TINs in quality tier 4 that below average costs (15% of 20% for CHF; 13% of 20% for hip replacement) than quality tier zero (10% of 20% for CHF; 6% of 20% for hip replacement). TINs in quality tier 5 with the lowest risk adjusted cost z-score are the most efficient, while TINs in quality tier 1 with highest risk adjusted cost z-score are the most inefficient within the episode proportion peer group.

These percentage distributions show again what Figures 5.2(a) and 5.2(b) show graphically: higher quality tiers have more TINs below the expected cost.

Table 5.3 (a): CHF: Percentage distribution of TINs across Risk Adjusted Quality Tiers by Average Risk Adjusted 30-day episode cost z-score for Episode Proportion Peer Group 5 (Number of TINs=674)

Risk Adjusted Quality Tier	**Risk Adjusted Cost z-Score < -2**	**Risk Adjusted Cost z-Score -2 - 1**	**Risk Adjusted Cost z-Score -1 - 0**	**Risk Adjusted Cost z-Score 0 - 1**	**Risk Adjusted Cost z-Score 1-2**	**Risk Adjusted Cost z-Score >2**	**Total**
5	0.0%	0.7%	14.5%	3.9%	1.0%	0.0%	20.1%
4	0.0%	0.1%	14.3%	5.1%	0.4%	0.1%	19.9%
3	0.1%	0.4%	12.1%	6.6%	0.7%	0.1%	19.9%
2	0.0%	0.0%	11.1%	7.9%	0.4%	0.3%	19.4%
1	0.0%	0.0%	9.6%	9.0%	1.0%	0.1%	19.6%
Total	0.1%	1.2%	61.6%	32.5%	3.5%	0.1%	100.0%

Table 5.3 (b): Hip Replacement: Percentage distribution of TINs across Risk Adjusted Quality Tiers by Average Risk Adjusted 30-day episode cost z-score for Episode Proportion Peer Group 5 (Number of TINs=371)

Risk Adjusted Quality Tier	Risk Adjusted Cost z-Score < -2	Risk Adjusted Cost z-Score -2 - -1	Risk Adjusted Cost z-Score -1 - 0	Risk Adjusted Cost z-Score 0 - 1	Risk Adjusted Cost z-Score 1-2	Risk Adjusted Cost z-Score >2	Total
5	1.4%	2.2%	9.8%	6.3%	0.3%	0.3%	20.0%
4	0.0%	1.6%	11.7%	5.7%	0.8%	0.3%	19.8%
3	0.0%	1.9%	9.5%	6.8%	1.6%	0.5%	19.8%
2	0.0%	1.9%	7.1%	8.4%	2.2%	0.0%	19.6%
1	0.0%	0.0%	6.0%	9.8%	2.7%	1.4%	18.5%
Total	1.4%	7.6%	44.1%	37.0%	7.6%	1.4%	100.0%

One question remains: do these relationships hold for *different episode proportion peer groups* (the data above focus exclusively on episode proportion peer group 5). Data in the Appendix V provide an answer to this question, in a format similar to the tables above. Tables 1 (a), 2(a) and 3(a) in that appendix show – for CHF, for episode proportion peer groups 5, 3 and 1 –the distribution of TINs across risk adjusted quality tiers by their cost z-scores. Tables 1 (b), 2(b) and 3(b) in Appendix V show the distribution of average risk adjusted 30-day episode costs for TINs across risk adjusted quality tiers by their episode cost z-score. The similar results for hip replacement episode proportion peer groups 5, 3 and 1 are presented in Tables 4(a), 5(a) and 6(a); and 4(b), 5(b) and 6(b) respectively.

The percentage distribution of TINs in the cost-quality matrix elements is similar across the peer groups (for a particular condition). As in our discussion above, for these other episode proportion peer groups. *There are more TINs below average cost in the higher quality tiers than the lower quality tiers*. As the as the proportion of the TINs responsibility decreases, from peer group 5 to 1, the variation in costs across TINs decreases and the risk adjusted costs come closer to zero.

A unit cost z-score hence represents more absolute dollar amounts in higher proportion peer groups than in lower proportion peer groups. This has to be kept in mind while weighting the performance of TINs across different TIN peer groups.

The varying levels of TINs' performance within and across peer groups should be weighted to appropriately to make them comparable in a payment system that rewards efficient TINs. The different methods for weighting TIN performance will be discussed in Chapter 6.

III. Combining Cost and Quality in the Beneficiary Approach

After considering different approaches to combining cost and quality for the beneficiary level analysis, we chose the third alternative – data envelopment analysis or DEA. In the beneficiary approach we use the plurality of non-hospital E&M visits as the attribution rule and attribute all the risk-adjusted costs and quality measures to the attributed physician (TIN). Thus, it is easier to think of the costs as inputs and the quality measures as output.

An alternative approach would have a cost function using a stochastic frontier approach.

Minimization the cost of producing a given level of output is the “dual” equivalent of maximizing output for a given level of cost. In the stochastic frontier approach, the dependent variable typically is the firm’s (TIN’s) total cost (not the cost per unit) and the explanatory variables are the level of output and input prices. The concept of a “frontier” is handled either by assuming that observations representing costs lower than that estimated minimum possible cost are due to measurement error, or by specifying the error terms to have a one-sided distribution such as half-normal.

The stochastic cost function approach faced a number of hurdles in this application. How should cost be defined? It the usual application it would be the firm’s internal cost of production, but that cost is not available in our data, and is of only secondary policy interest. The cost of primary interest is the cost incurred by Medicare beneficiaries and taxpayers.

Second, how should “total quantity” be defined? Should it be the total number of attributed beneficiaries, the total number of office visits or hospitalizations, the beneficiary’s health status (unavailable in the data) or some other measure? Third, how should inputs be defined? One possibility would be the number of hours spent by the physician and non-physician personnel, the square feet of office space, x-ray machines and other capital equipment. Again, those data not only are unavailable, they are not particularly relevant to the task at hand.

We already have controlled for geographic variation in input prices by standardizing our cost data for the Geographic Practice Cost Index (GPCI). Quality, one of the main foci of our analysis, is not as easily incorporated into a cost function of that form.

Rather than try to adapt the research questions underlying our analysis to the stochastic cost function approach designed to analyze firms’ internal costs, we adopted the DEA approach to address the relevant policy question, “What are taxpayers and beneficiaries getting for the Medicare dollars that are spent caring for beneficiaries?” In that approach, risk-adjusted Medicare dollars are the inputs and various measures of quality are the outputs.

A. DEA Results from the Five State Data

This section presents DEA results from the 2008 and 2009 five state dataset. We selected two sets of DEA results to present in detail:

- Diabetes monitoring measures; and
- Emergency department (ED) visits and ambulatory care sensitive (ACS) admissions.

We present separate DEA analyses for these two sets of measures. Although DEA analysis is capable of analyzing multiple outputs across many different types of quality measures, we noted earlier that requiring a physician to have adequate sample size across many different types of quality measures would reduce the number of physicians who could qualify for an incentive payment.

We chose these particular measures for several reasons:

- They represent different types of quality requiring different types of quality-improving initiatives on the part of the physician.

- Diabetes monitoring needs to increase in the majority of practices, while avoidable utilization needs to decrease.
- They both incorporate multiple measures of quality, thus illustrating that advantage of DEA analysis.

Because both diabetes monitoring and the combined ED/ACS measures involve multiple quality measures, it is not easy to present the results from those analyses in simple two-dimensional scatterplots. That is possible with measures such as colon and breast cancer screening and we have included those results in Appendix 15. Examination of the results in that appendix may help to clarify some of the concepts underlying DEA analysis.

The LIMDEP computer program that we used for the DEA analysis was limited to 11,750 observations, so when necessary, we drew random samples from the five state datasets that approached that sample size limit.[43] We also used only the TINs that appeared in both the 2008 and 2009 data so that we could compare the stability of the DEA performance measures over time. Finally, we restricted the analysis to TINs that had 20 or more beneficiaries in the denominator for the measure being evaluated. Twenty is the sample size at which the t-distribution is well approximated by the normal distribution. Again, we emphasize that our recommendation regarding reward payments in a value based system is that *those payments be made only to physicians who provide care to enough Medicare beneficiaries to allow a reasonably reliable estimate of performance, so these restrictions have a substantial policy basis and aren't merely matters of methodological convenience*.

B. Diabetes Monitoring

We begin the presentation of results with the diabetes monitoring measures for several reasons. First, there are three measures:

- HbA1c screening per eligible beneficiary (CMS, 2011)
- LDL-C screening per eligible beneficiary (CMS, 2011)
- Medical attention for nephropathy per eligible beneficiary (CMS, 2011).

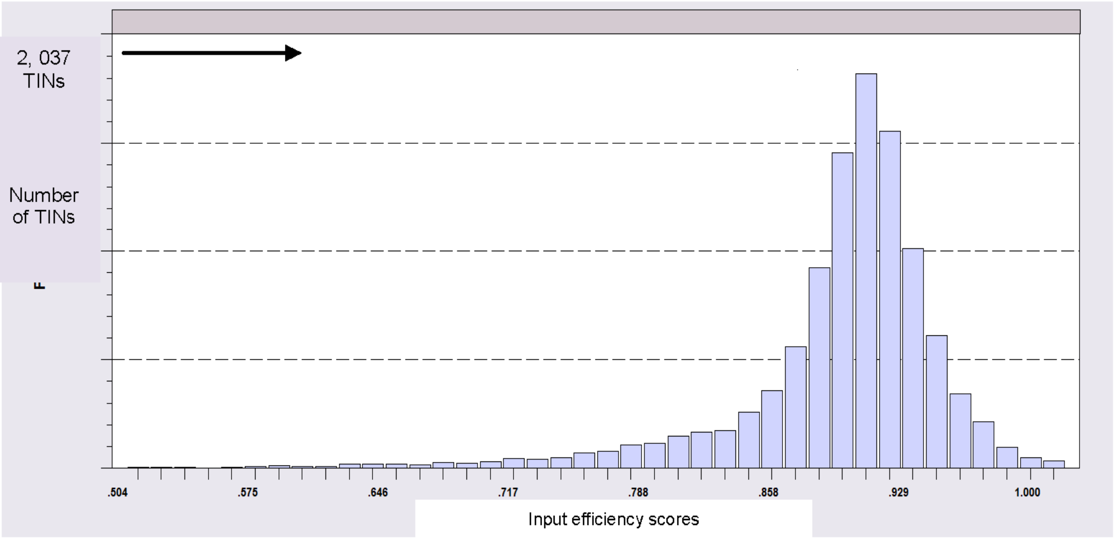
With three measures, we can illustrate DEA's capability to accommodate more than one measure of output. The three measures do not have to be combined into a single measure prior to comparing them to cost (although nothing precludes that approach). Second, the distribution of input efficiency scores for diabetes is similar to the distribution for other quality measures, so the results are not misleading.

Selecting cases as described above produced a sample of 11,637 TINs, each with 20 or more diabetic beneficiaries. None of the diabetes monitoring measures were strongly correlated with risk adjusted cost. But annual health care costs have a large random component, and it would be unusual if any variable, even a single measure of contemporaneous utilization, explained a large portion of the variance in costs.

[43] There are many software packages that include DEA analysis, including the freeware package "R." We have not explored the sample size limitations of all those packages.

DEA analysis using risk adjusted cost as the input[44] and the three diabetes monitoring measures as output produced TIN level input efficiency scores ranging from 0.5 to 1.0 (the 1.0 scores are on the frontier). *That means that a frontier practice could produce the same quality for approximately half the risk adjusted cost as the least input-efficient practice.* The mean efficiency score was 0.88 with a standard deviation of 0.05. The distribution of input efficiency scores is shown in Figure 5.5.

Figure 5.5: Distribution of TINs by Efficiency Scores for Diabetes Monitoring (2008 five state data) N=11,637



C. ED Visits and ACS Admissions

In our next example of DEA analysis, the input was risk-adjusted cost, as before, and the outputs were:

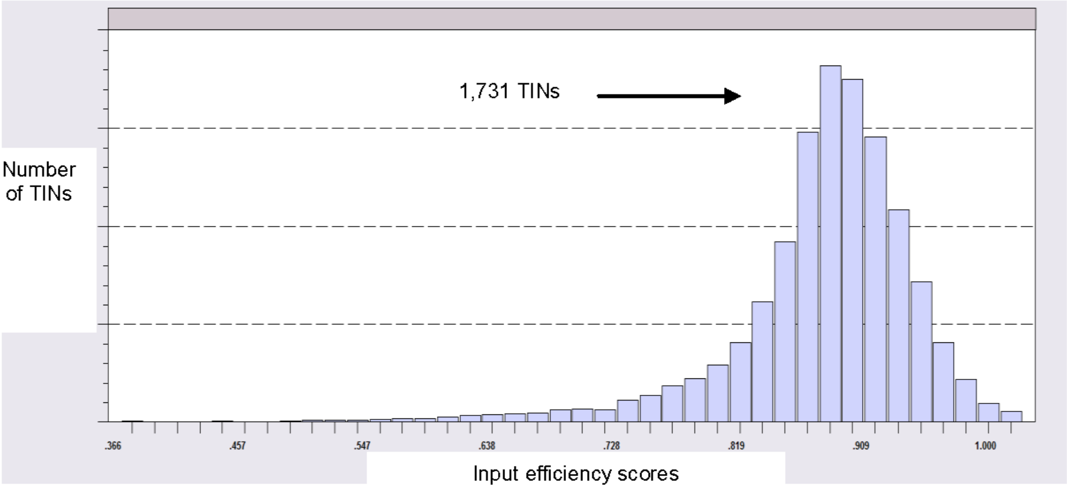
- ED visits per beneficiary
- Non-emergent ED visits per beneficiary
- Emergent, but primary care treatable ED visits
- Emergent, but preventable or avoidable ED visits
- Ambulatory care sensitive (ACS) admission per attributed beneficiary.

We drew a random sample of 11,741 TINs from the five state data for this analysis. The DEA analysis produced TIN level input efficiency scores ranging from 0.366 to 1.0 (the 1.0

[44] Our risk-adjusted cost measure is computed across all the beneficiaries attributed to the physician (TIN). Thus, we do not compute a separate risk-adjusted cost measure for diabetic beneficiaries versus the cost for hospitalized beneficiaries who are eligible for the potentially preventable readmission measure. Our rationale is as follows: consider a cost-conscious TIN that monitors Medicare spending per capita across *all* its attributed beneficiaries. That TIN might ask whether it is better to allocate Medicare expenditures to better monitoring of diabetic patients or to better follow-up care for beneficiaries being discharged from the hospital. In that case, looking only at the cost for the diabetic beneficiaries would be an inaccurate measure of the TIN's cost allocation efforts.

scores are on the frontier). *That means that a frontier practice could produce the same quality for approximately one-third the risk adjusted cost as the least input-efficient practice*. The mean efficiency score was 0.86 and the standard deviation was 0.06. The distribution of input efficiency scores is shown in Figure 5.6.

Figure 5.6: Distribution of Input Efficiency Scores for Avoidable Utilization (2008 five state data) N=11,741



Generally, we found that the DEA efficiency scores exhibited sufficient dispersion to provide a useful way of discriminating between high efficiency and low efficiency TINs, but that is all that we have established at this point in the analysis. There are a number of questions regarding the DEA scores that become relevant if they are to be used as a basis for value based payments. We take up those questions in the next chapter.

Similar results for the remaining quality measures are included in Appendix 15. In the next chapter we discuss how the results from the MS-DRG and beneficiary analyses are applied to a value based payment system.

Chapter 6

Translating Performance Measures into Payment Policy

Chapters 3, 4, and 5 above set forth methods for computing and combining TIN-level performance measures. But that is only one step in implementing a value based purchasing system for FFS Medicare. The reliability and stability of the performance measures must be assessed. Then the performance measures must be linked to an algorithm for producing incentive payments. As noted earlier, we recommend that decisions regarding value based incentive payments be separated from any across-the-board changes in physician fee levels. Value based incentive payments are *essentially* payments designed to reward good performance, not annual cost-of-doing-business increases. Each year's value based payments should be conditioned on good performance during the previous year. We begin by discussing the MS-DRG approach, then discuss the Beneficiary approach.

I. MS-DRG APPROACH

In this section we first discuss how performance measures for TINs obtained from the MS-DRG tiering approach can be applied in a functioning payment system. We then examine the stability of TIN performance measures for 30- and 60-day episodes for two selected MS-DRGs – CHF and hip replacement. The point of this discussion is primarily because the stability of measures may vary by the length of the episode window, and thus may influence which window we choose (30- or 60-day) We finally discuss the steps to implement the MS-DRG approach and issues around implementing it in the context of Medicare's payment policy for rewarding performance of hospital-based TINs.

As noted in Chapter 5, all comparisons of cost and quality must be done *within* peer groups, as the peer groups allow us to compare TINs with similar mean proportions of episode costs. If we did not do this, TINs with a lower mean proportion of episode cost would seem to be more efficient – as they would be producing 100 percent of an episode's quality with a fraction of the episode's cost attributed to them.

Tiered TIN measures of performance across different episode proportion peer groups for MS-DRG episodes can be used to reward TINs. Three steps are required to establish incentive payments for inpatient physicians. Payments must be distributed:

1. Across MS-DRGs;

2. Across "proportion of total episode cost" peer groups; and

3. Within "proportion of total episode cost" peer groups, based on cost and quality of care.

These three steps result in a set of weights that can be applied to each cost and quality measure within and across MS-DRGs and peer group.

A. Distributing Incentive Payments across MS-DRG Conditions

In a value based payment system, the payor seeks to provide varying rewards for physicians managing conditions that vary in difficulty – in particular, that vary in the challenge they present for hospital and post-acute care. For instance, an episode of hip replacement consumes more hospital and post-acute care than an episode of knee replacement, and an episode of non-laparoscopic cholecystectomy is more challenging to manage for hospital physicians than an episode of laparoscopic cholecystectomy. An inpatient payment system should be able to give greater rewards to hospital-based TINs that provide care to MS-DRG conditions that are more intensive and complex – not only in terms of cost, but also in terms of possible adverse episode outcomes.

Weighting MS-DRG conditions by cost within the 30- or 60-day episode window is relatively straightforward and that is the approach we illustrate in this section. *Relative weights based on cost can be obtained by dividing the mean standardized episode cost for that condition by the mean standardized episode cost across all MS-DRG conditions*. The episode costs we use for obtaining weights are not risk adjusted because we wish weight episodes on the basis of how expensive the conditions actually are for Medicare.

Table 6.1 below, shows the relative weights for 30- and 60-day episodes of the eleven MS-DRG conditions selected for this analysis (see Chapter 3). Among both the 30- and 60-day episodes, AMI with CABG has the highest relative weight (1.85 for 30-day episode and 1.66 for 60-day episode). Although a 60-day episode of AMI with CABG has higher episode costs than a 30-day episode of AMI with CABG, its relative weight is lower than that of the 30-day episode because 60-day episodes for other MS-DRG conditions are relatively high cost, too. Acute ischemic stroke closely follows AMI with CABG in its relative cost.

These relative weights for the condition episode can be used by CMS to reward TINs managing more intensive conditions.

B. Distributing Incentive Payments across TIN Peer Groups

Peer groups that account for a greater proportion of total mean episode costs should receive proportionately larger shares of the total incentive payments, on the premise that they have been responsible for more of the care (measured this approximate way). Tables 6.2(a) and 6.2 (b) below, show the percent of total incentive payments that would be distributed to each TIN peer group based on the average proportion of standardized 30-day episode cost for which TINs in the peer group are responsible. The tables shows data for CHF and hip replacement, respectively. Again the weights are based on amounts that are not risk adjusted to reflect how much Medicare actually pays TINs for differing proportions of episode responsibility. TINs in the highest episode proportion peer group for CHF (Peer Group 5) would be allocated 36 percent of the total payment incentive for CHF, because that is their proportion of the total standardized unadjusted 30-day episode cost. Those in the lowest episode proportion peer group would get 6 percent of the total payment incentive for CHF. The proportion of incentive payment allocated to TINs in the highest and the lowest proportion peer groups for hip replacement are 30 percent and 5 percent, respectively (Table 6.2(b)).

Table 6.1: Relative Weights for 30- and 60-day Episodes of Selected MS-DRG Conditions based on Relative Mean Standardized Episode Costs (in order of relative 30-day weight)

Condition	# of Episodes	Episode Length	Mean Standardized Episode Cost ($)	Relative Weight for Condition
Back Pain w Other Back Procedure	9,025	30-day	2,976	0.34
	9,025	60-day	5,963	0.48
Bronchitis	13,780	30-day	4,866	0.56
	13,780	60-day	7,582	0.61
Laparoscopic Cholecystectomy	15,851	30-day	5,459	0.63
	15,851	60-day	7,591	0.61
COPD	78,760	30-day	6,594	0.76
	78,760	60-day	10,654	0.86
AMI w PTCA	13,679	30-day	6,910	0.80
	13,679	60-day	9,927	0.80
Pneumonia	86,869	30-day	7,380	0.85
	86,869	60-day	11,119	0.90
Non-Laparoscopic Cholecystectomy	3,701	30-day	8,497	0.98
	3,701	60-day	11,528	0.93
CHF	107,185	30-day	8,974	1.04
	107,185	60-day	14,208	1.15
Knee Replacement	53,647	30-day	9,243	1.07
	53,647	60-day	10,738	0.87
Medical Back Pain	9,904	30-day	10,032	1.16
	9,904	60-day	14,713	1.19
Hip Replacement	24,603	30-day	10,513	1.22
	24,603	60-day	12,395	1.00
Back Pain w Spinal Fusion	6,332	30-day	12,463	1.44
	6,332	60-day	14,452	1.17
Stroke w Cerebral Infarct	43,246	30-day	12,887	1.49
	43,246	60-day	17,404	1.41
Medical AMI	34,194	30-day	13,281	1.54
	34,194	60-day	18,032	1.46
Hip Fracture	3,098	30-day	13,468	1.56
	3,098	60-day	18,078	1.46
Acute Ischemic Stroke	1,458	30-day	15,204	1.76
	1,458	60-day	19,920	1.61
AMI w CABG	2,559	30-day	16,000	1.85
	2,559	60-day	20,594	1.66

Table 6.2 (a): TIN Peer Group Weights for 30-day Episodes of CHF From TIN's Average Standardized 30-day Episode Cost

TIN Episode Proportion Peer Group	TIN's Proportion of Episode Cost: Min-Max	TINs' Average Standardized 30-day Episode Cost: MEAN	TINs' Average Standardized 30-day Episode Cost: WEIGHT
1	0 - 0.20	$1,221	0.06
2	0.20 - 0.40	$2,878	0.14
3	0.40 - 0.60	$4,113	0.19
4	0.60 - 0.80	$5,355	0.25
5	0.80 - 1	$7,623	0.36

Table 6.2 (b): TIN Peer Group Weights for 30-day episodes of Hip Replacement from TIN's Average Standardized d 30-day Episode Cost

TIN Episode Proportion Peer Group	TIN's Proportion of Episode Cost: Min-Max	TINs' Average Standardized 30-day Episode Cost: Mean	TINs' Average Standardized 30-day Episode Cost: Weight
1	0 - 0.20	$1,365	0.05
2	0.20 - 0.40	$3,663	0.13
3	0.40 - 0.60	$7,229	0.25
4	0.60 - 0.80	$7,890	0.27
5	0.80 - 1	$8,744	0.30

CMS also could set minimum thresholds of episode cost responsibility, if it chose. Given a fixed amount of money to distribute through incentive payments, the tradeoff would be between providing small incentives to the many physician entities that have small involvements in the care of MS-DRG episodes, versus providing larger incentives to fewer physician entities that play a much larger role. If the goal is coordination of care within a TIN, then CMS could choose to reward incentive payments to only those TINs that account for, say, at least 25 percent of an episode's cost on an average.

B. Distributing Incentive Payments Based on Quality and Cost within a Peer Group

Within a "proportion of total episode cost" peer group, CMS must determine whether and how much to reward TINs for being (a) high quality, (b) low cost, or (c) efficient (high quality and low cost). Table 6.3 shows one way in which the variation in TIN performance across quality tiers and cost z-scores might be approached. Quality tiers are regrouped into three levels and risk-adjusted cost z-scores into two, creating a six-cell matrix.

If quality is of paramount importance to CMS, then it can choose to reward TINs that have average or above average composite quality scores – allocating the highest weight to TINs in the highest quality tier, and progressively lowering the weight for TINs in lower quality tiers. If cost is of greater concern, it can chose to reward low cost TINs, using weights proportional to how low the TINs' mean risk adjusted costs or cost z-scores are. Finally, TINs could be rewarded only if they are both low cost and high quality, with weights determined by how much CMS is willing to pay for either. An example of the weighting approach is shown in Table 6.4, This model collapses several risk adjusted quality tiers and risk adjusted cost groups for simplicity. But the method is flexible, and ultimately, the numbers of cells and the weights can be determined by policy makers. The weights used in this example are for demonstration purposes only. Ultimately policy makers would have to determine what weights to use.

Table 6.3: Variation in TIN Performance across Risk Adjusted Quality Tiers and Cost z-Scores

Risk Adjusted Quality Tier	Risk Adjusted Cost z-Score < -2	Risk Adjusted Cost z-Score -2 - -1	Risk Adjusted Cost z-Score -1 - 0	Risk Adjusted Cost z-Score 0 - 1	Risk Adjusted Cost z-Score 1-2	Risk Adjusted Cost z-Score >2
5	<- High Quality - Low Cost			High Quality - High Cost ->		
4	<- High Quality - Low Cost			High Quality - High Cost ->		
3	<- Average Quality – Low Cost			Average Quality – High Cost ->		
2	<- Low Quality - Low Cost			Low Quality - High Cost ->		
1	<- Low Quality - Low Cost			Low Quality - High Cost ->		

Table 6.4 Illustrative Weights to Reward Quality and Cost

Adjusted Quality	Risk Adjusted Cost z-Score < -2	Risk Adjusted Cost z-Score -2 - -1	Risk Adjusted Cost z-Score -1 - 0	Risk Adjusted Cost z-Score 0 - 1	Risk Adjusted Cost z-Score 1-2	Risk Adjusted Cost z-Score >2
5	10			7		
4	10					
3	5			3		
2	1			0		
1						

D. Stability of TIN Performance Measures between 30- and 60- day episodes

We examined the stability of TIN's cost, quality and performance measures across 30- and 60-day episodes. As noted earlier, our interest in the stability of the measures is that, if cost, quality, and performance are stable for TINs across 30- and 60- day episodes, then it makes no difference if we use 30- or 60-day episodes to profile hospital based TINs. However if cost, quality and performance are very different for these two periods, then the decision to choose one of the episode lengths, or both, to profile hospital based TINs needs to be made.

In addition, the stability of cost, quality and performance measures for TINs could vary by type of MS-DRG. We therefore examined the stability of these measures for one medical MS-DRG (congestive heart failure) and one surgical MS-DRG- (hip replacement).

Tables 6.5 (a) and (b) below show the stability of risk-adjusted costs between 30- and 60-day episodes. The tables also show the 30- and 60- day risk adjusted composite quality measures for CHF and Hip replacement TINs respectively, across the episode proportion peer groups. The correlation between risk adjusted costs is approximately 0.80 for CHF TINs and 0.85 for hip replacement TINs across all episode proportion peer groups. What this means is that 64 percent of the variation in CHF costs and 72 percent of the variation in hip replacement costs is common between 30- and 60-day episodes. 60-day episodes, which have higher variation in costs, capture 36 percent and 28 percent of unique variation in risk adjusted costs for CHF and hip replacement respectively.

The correlation between risk adjusted composite quality measures is much lower for CHF (approximately 0.59) compared to hip replacement (0.85), for TINs across all episode proportion peer groups. Sixty-five percent of the variation in TINs' quality is unique across 30- and 60-day CHF episodes, while only 28 percent of the variation is unique for hip replacement.[45] The variation in the risk adjusted composite quality score comes wholly from episode level quality measures like mortality, avoidable readmissions and avoidable ED visits, as hospital-level quality measures are unchanged across 30 and 60 day episodes. *Hence it is to be expected that the stability in composite quality would be much poorer for between 30- and 60- day episodes of CHF (where mortality, avoidable readmissions and avoidable ED visits would be higher for 60-day episodes) than hip replacement*.

For each of the two conditions, we examined the stability of TIN risk adjusted quality tiers and TIN performance score (with respect to both quality and cost) over 30- and 60-day episodes. The analysis was performed separately for each episode proportion peer group and is presented in detail in Appendix 17.

The analysis shows that TINs *providing higher quality of care compared to their peers in 30-day episodes may not be providing higher quality for 60-day episodes. Similarly, lower quality TINs in 30-day episodes may not be lower quality in 60-day episodes. This is more likely for CHF TINs than hip replacement TINs. Hip replacement TINs are more stable with respect to their quality tiers between 30- to 60- day episodes compared to CHF TINs (whose quality scores are more likely to change if mortality, avoidable readmissions and avoidable ED visits are higher for 60-day episodes.* Changes in TIN's performance score ranks, going from 30- to 60-day episodes were more likely to come from changes in TIN's quality tiers than changes in TIN's risk adjusted cost z-scores. *Costs are therefore more stable than quality between 30- and 60-day episodes*

For surgical MS-DRGs like hip replacement, knee replacement, cholecystectomy, back pain etc., there is relatively better stability between 30- and 60-day quality measures. But for medical MS-DRGs like CHF, COPD, AMI, Pneumonia, etc., the stability between quality measures or tiers is much poorer between 30- and 60-day episodes. Moreover, 60-day episodes for MS-DRG capture more variation in both cost and quality at the TIN level and would hence be more useful for measuring TIN performance.

We therefore recommend that CMS use 60-day episodes, as they capture more variation in cost, quality and performance across TINs.

[45] The square of the correlation coefficient (ρ^2) gives the percentage variation that is common between the two variables of interest; while 1-(ρ^2) is the percentage variation between the two variables that is unique.

Table 6.5 (a). Stability between TIN's 30 and 60 day Episode Risk Adjusted Cost and Risk Adjusted Composite Quality Measure for CHF across TIN Episode Proportion Peer Groups

TIN's Episode Proportion Peer Group	TIN's Risk Adjusted 30-Day Episode Cost Mean (Std)	TIN's Risk Adjusted 60-Day Episode Cost Mean (Std)	Correlation Coefficient (Rho) between TIN's Risk Adjusted 30- and 60-day Episode Costs	TIN's Risk Adjusted Composite Quality Measure for 30-day Episode Mean (Std)	TIN's Risk Adjusted Composite Quality Measure for 60-day Episode Mean (Std)	Correlation Coefficient (Rho) between Risk Adjusted Composite Quality Measure for 30 and 60 day Episode
5	$57 ($12,351)	$185 below expected ($14,064)	0.74	40 (12)	48 (15)	0.63
4	$420 below expected ($4,453)	$657 below expected ($6,172)	0.86	45 (12)	51 (14)	0.63
3	$153 below expected ($3,146)	$203 below expected ($4,414)	0.81	48 (12)	51 (13)	0.59
2	$128 ($2,053)	$163 ($2,942)	0.79	50 (12)	49 (13)	0.58
1	$146 ($1,207)	$202 ($1,603)	0.83	51 (9)	49 (15)	0.57

Table 6.5 (b). Stability between TIN's 30 and 60 day Episode Risk Adjusted Cost and Risk Adjusted Composite Quality Measure for Hip Replacement across TIN Episode Proportion Peer Groups

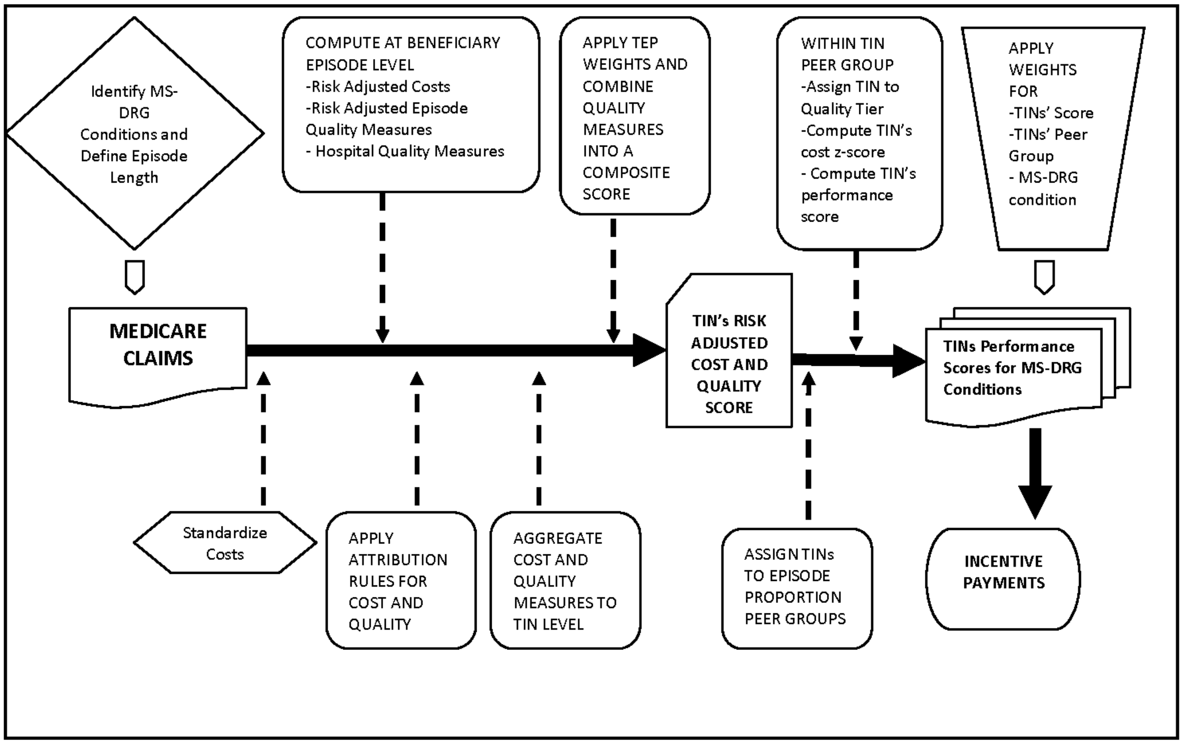
TIN's Episode Proportion Peer Group	TIN's Risk Adjusted 30-Day Episode Cost Mean (Std)	TIN's Risk Adjusted 60-Day Episode Cost Mean (Std)	Correlation Coefficient (Rho) between TIN's Risk Adjusted 30- and 60-day Episode Costs	TIN's Risk Adjusted Composite Quality Measure for 30-day Episode Mean (Std)	TIN's Risk Adjusted Composite Quality Measure for 60-day Episode Mean (Std)	Correlation Coefficient (Rho) between Risk Adjusted Composite Quality Measure for 30- and 60-day Episode
5	$509 below expected ($4,161)	$433 below expected ($5,634)	0.86	45 (9)	46 (9)	0.90
4	$162 ($2,876)	$81 ($3,430)	0.87	48 (9)	48 (9)	0.90
3	$160 ($3,654)	$81 ($5,724)	0.83	49 (11)	9 (11)	0.84
2	$170 ($2,887)	$326 ($3,596)	0.85	49 (11)	49 (11)	0.81
1	$88 ($906)	$97 ($1,278)	0.83	50 (12)	50 (12)	0.85

E. Steps to Implement MS-DRG Approach

The MS-DRG approach can be implemented for Medicare payment using the following steps outlined in the flow chart below:

- The MS-DRG conditions for which incentive payments would be provided and the episode length for which physician entities would be profiled are first defined.
- Medicare claims are input to the process. And costs are standardized.
- For each selected MS-DRG condition, costs are risk adjusted at the beneficiary-episode level. Risk adjusted costs are apportioned out to TINs providing care for the episode using the proportional attribution rule (proportion of part B billing for index hospitalization).
- Risk adjusted episode costs, and episode proportion are aggregated at the TIN level to get the mean risk adjusted episode cost and mean episode proportion for the TIN. TINs are assigned peer group based on their mean episode proportion.
- Episode level quality measures computed from claims are risk adjusted at the beneficiary episode level and then aggregated to the TIN level.
- Hospital level quality measures are linked to beneficiary episodes and aggregated to the TIN level.
- At the TIN level- episode level quality measures and hospital level quality measures are combined to a single risk adjusted quality score composite for the TIN, using weights from Technical Expert Panel
- In each episode proportion peer group, TINs assigned a risk adjusted quality tier based on their risk adjusted composite quality score, and a cost z-score based on their mean risk adjusted episode cost. TINs are then assigned a performance score based on their risk adjusted composite quality score and cost z-score.
- Incentive payments are assigned to TINs based on the weights for their MS-DRG conditions, peer-group and performance score with the peer group.

Figure 6.1. Steps to Obtain Physician Incentive Payments Using MS-DRG Approach


Identify MS-DRG Conditions and Define Episode Length
COMPUTE AT BENEFICIARY EPISODE LEVEL
-Risk Adjusted Costs
-Risk Adjusted Episode Quality Measures
- Hospital Quality Measures
APPLY TEP WEIGHTS AND COMBINE QUALITY MEASURES INTO A COMPOSITE SCORE
WITHIN TIN PEER GROUP
-Assign TIN to Quality Tier
-Compute TIN's cost z-score
- Compute TIN's performance score
APPLY WEIGHTS FOR
-TINs' Score
-TINs' Peer Group
- MS-DRG condition
MEDICARE CLAIMS
TIN's RISK ADJUSTED COST AND QUALITY SCORE
TINs Performance Scores for MS-DRG Conditions
Standardize Costs
APPLY ATTRIBUTION RULES FOR COST AND QUALITY
AGGREGATE COST AND QUALITY MEASURES TO TIN LEVEL
ASSIGN TINs TO EPISODE PROPORTION PEER GROUPS
INCENTIVE PAYMENTS

F. Implementing the MS-DRG Approach for Medicare Payment Policy

In this section we discuss some of the issues around implementing MS-DRG approach and suggest ways to strengthen it while applying it for future Medicare payment policy.

- We chose to attribute an MS-DRG episode to multiple physician entities based a proportionate Part B billing rule, after our attempts at using a single attribution approach were unsuccessful. The multiple attribution approach allowed episodes to be exhaustively attributed to all TINs providing care during the hospitalization, but approximately 40-45% of TINs on average had less than 10% of episode costs attributed to them.

- Even after moving from NPIs to TINs, the number of episodes attributed to physician entities remained low. The median number of episodes attributed to TINs was 3. And 25-35% of TINs had only one episode attributed to them.

- The problem of low episode proportions and low episode volume is a limitation of the MS-DRG approach. CMS could set thresholds for episode proportions and volumes that TINs should meet for MS-DRGs to be eligible for being profiled for performance. The level of the thresholds is in part a statistical issue, but is at some point likely to be arbitrary, though CMS may have policy considerations that assist in weighing where to establish the thresholds (e.g., concerns as to how inclusive CMS wants the measurement system to be, versus the defensibility of larger episode proportions and volumes for performance measurement).

- The MS-DRG approach uses a risk adjusted composite quality score to measure quality of care provided by TINs, and a performance score rank to measure TIN's performance. We suggest that the *composite* quality measure and performance score be used to facilitate the implementation of a payment system rewarding lower cost and higher quality. *Individual* quality measures should be reported to the TINs, showing benchmarks against peers, as these are more actionable for quality improvement than are the composite measures.

- Our choice of quality measures for MS-DRG episodes was constrained by what we could directly compute from claims. In addition, the TEP suggested that we use hospital level quality measures for episodes where appropriate. Hospital level measures, which are usually uniformly high or low across TINs, are not as useful in distinguishing performance among TINs, as compared to episode level measures. To successfully implement the MS-DRG approach or any episode-based approach of measuring performance, CMS will have to ensure that more quality measures are available at the episode level. With the availability of Present on Admission (POA) field on hospitalization claims, AHRQ's patient safety indicators can be obtained at the episode level. Hospital Compare process measures for AMI, CHF, Pneumonia and Surgical care improvement could be included as measures reportable as a part of the Physician Quality Reporting System. As of this date, a few process measures from the Hospital Compare measure set are being reported on the PQRS system.

- The tiered approach, used to combine cost and quality, is designed to provide maximal flexibility. It allows for value weights at several levels and can use a variety of subdivisions that are a matter of policy choice. This flexibility can also present a challenge. There is no obvious empirical basis for the weights. They will rely on some judgment. They can be used to pursue predetermined policy goals. The ultimate decision about how many cost-quality cells to create, what weights to apply to each, and how to treat the relative levels of participation must be made in a policy context.

II. Beneficiary Level Approach

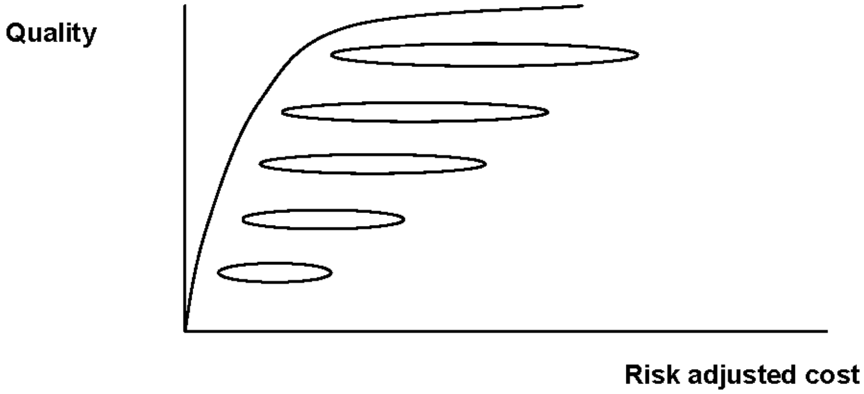
In this section, we first analyze the characteristics of DEA input efficiency measures for two types of measures: diabetic monitoring and the combination of emergency department utilization measures and ambulatory care sensitive hospitalization measures. Then we provide a detailed description of the way in which the DEA input efficiency scores can be incorporated into Medicare payment policy.

A. The Relationship of DEA Input Efficiency Scores to Cost and Quality Measures

There is nothing in the DEA analysis that requires the DEA input efficiency scores to be positively related to quality or negatively related to cost. The DEA frontier is defined by individual TINs and because the data for one TIN is independent of the data for other TINs the distribution of TINs has no necessary relationship to the position of the TINs on the frontier.

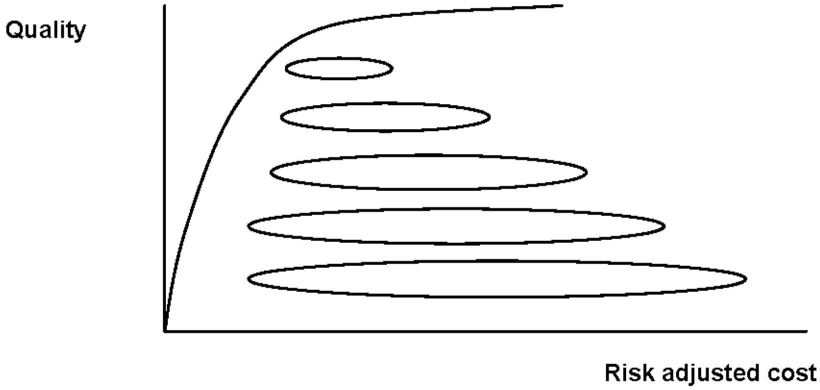
This point is illustrated in the following figures. In Figure 6.2, the dots represent the individual TINs that define the frontier, and the ovals represent the mass of observations at each level of quality. In Figure 6.3+, average efficiency (the ratio of distance A to distance B in Figure 5.4), is unrelated to cost and quality.

Figure 6.2: DEA Input Efficiency, Cost, and Quality



In Figure 6b.2, however, more efficient TINs exhibit higher quality and lower risk adjusted cost.

Figure 6.3: DEA Input Efficiency, Cost, and Quality



When DEA input efficiency measures are positively associated with higher quality and lower cost, we refer to the DEA measures as "internally consistent." In other words, the DEA input efficiency measure provides a useful summary of cost and quality. *This is an important point, because if DEA input efficiency is associated with higher quality and lower cost, then CMS can avoid the difficult question, "How much quality does the Medicare program want to buy for its beneficiaries?"* If DEA input efficiency scores were internally consistent, then a value based payment system based on internally consistent DEA input efficiency scores would reward both lower cost and higher quality.

In order to test for internal consistency, we compared the values of cost and quality measures for TINs with low and high values of DEA input efficiency on the diabetic monitoring measures and the combined ED and ACS admissions measures.

Table 6.6 compares the means and standard deviations of cost and measures of diabetes monitoring for TINs in the top quartile (25 percent) of efficiency scores to TINs in bottom quartile of efficiency. The results show that there is approximately a $12,400 difference in average risk adjusted cost per beneficiary between the high and low performing TINs on the diabetes monitoring measures. At the same time, all of the diabetes monitoring measures are about *20 to 30 percent higher in the high performing group*. Thus, the diabetes monitoring performance measures exhibit internal consistency.

Table 6.6: Comparison of cost and diabetes monitoring measures for TINs in the highest and lowest quartiles of DEA input efficiency TINs (2008 5 state data) N=11,637

	Mean for TINs in the highest quartile of DEA input efficiency	Standard deviation for highest efficiency TINs	Mean for TINs in the lowest quartile of DEA input efficiency	Standard deviation for lowest efficiency TINs
Risk-adjusted cost	$4,718.83 below overall average[46]	1,923.67	$7,701.72 above overall average	9,631.93
Hba1c test	.7846	1930	.6087	1910
LDL-C test	.8099	.1953	.6110	.2219
Medical attention for nephropathy	.6994	.2114	.5653	1748

Note: All differences in means are statistically significant at less than the 0.001 level

Table 6.7 shows the results of the same type of comparison for the combined ED and ACS measures. Here the results are even more pronounced. There was approximately a $12,700 difference in average risk adjusted cost per beneficiary between the high and low performing TINs on the ED/ACS measures. All of the ED/ACS measures are *two to seven times higher in the poor performing group* (higher in this case is bad as it means more inappropriate ED visits or ACS admissions). Thus, the ED/ACS measures also exhibit internal consistency. We obtained similar results for the other performance measures.

Table 6.7: Comparison of cost and ED/ACS measures for high and low performing TINs (2008 5 state data) N=11,741

	Mean for TINs in the highest quartile of DEA input efficiency	Standard deviation for highest efficiency TINs	Mean for TINs in the lowest quartile of DEA input efficiency	Standard deviation for lowest efficiency TINs
Risk-adjusted cost	$4,903.51 *below* overall average	1,879.43	7,688.26 *above* overall average	10,389.70
ED visits per beneficiary	0.3698	0.2325	1.0271	0.5110
Non-emergent ED visits per beneficiary	0.0549	0.0571	0.1132	0.0888
Primary care treatable ED visits per beneficiary	0.0702	0.0534	0.1665	0.0868
Preventable or avoidable ED visits per beneficiary	0.0269	0.0289	0.1044	0.0723
Ambulatory care sensitive admissions per beneficiary	0.0346	0.0351	0.2139	0.1576

Note: All differences in means are statistically significant at less than the 0.001 level

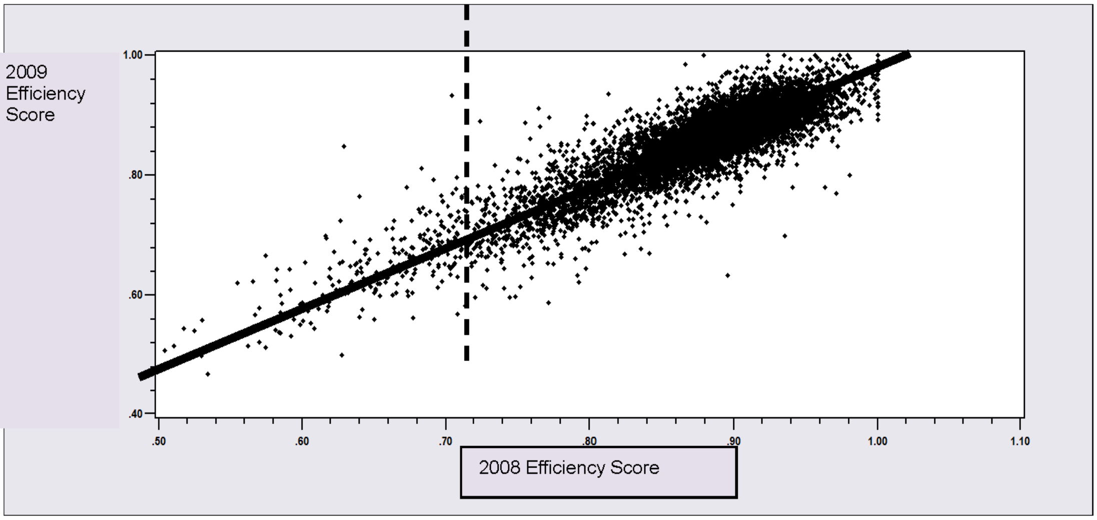
[46] The mean refers to the mean for diabetic patients.

B. Stability of TIN Performance Over Time

Any comparison of performance for the same TIN over multiple years combines three effects. First, one would expect a certain amount of stability in adjacent years, because in most cases the same physicians are likely to be associated with the TIN and the internal structure of incentives and rewards in the TIN are unlikely to change dramatically from one year to the next. A second, countervailing effect is that the TIN may be instituting changes in the practice that could affect the cost or quality of care. Third, even though we restrict our analysis to TINs with 20 or more attributed beneficiaries, there will be year-to-year sampling variation in unobserved variables affecting the performance scores. We cannot control for that random variation in unobserved variables.

Figure 6.4 shows a scatter plot comparing the 2008 and 2009 input efficiency scores for the same set of TINs. The diagonal line in Figure 6.8 traces the points at which the 2008 and 2009 scores are unchanged. Points above the diagonal line represent TINs for which the 2009 score was higher than the 2008 score, whereas points below the line represent TINs for which the 2009 score was lower than the 2008 score. It appears from the diagram that the input efficiency scores for most of the TINs declined from 2008 to 2009. In fact, the mean difference in input efficiency scores (2009 minus 2008) is -0.022.

Figure 6.4: Comparison of 2008 and 2009 input efficiency for diabetic monitoring (five state data)



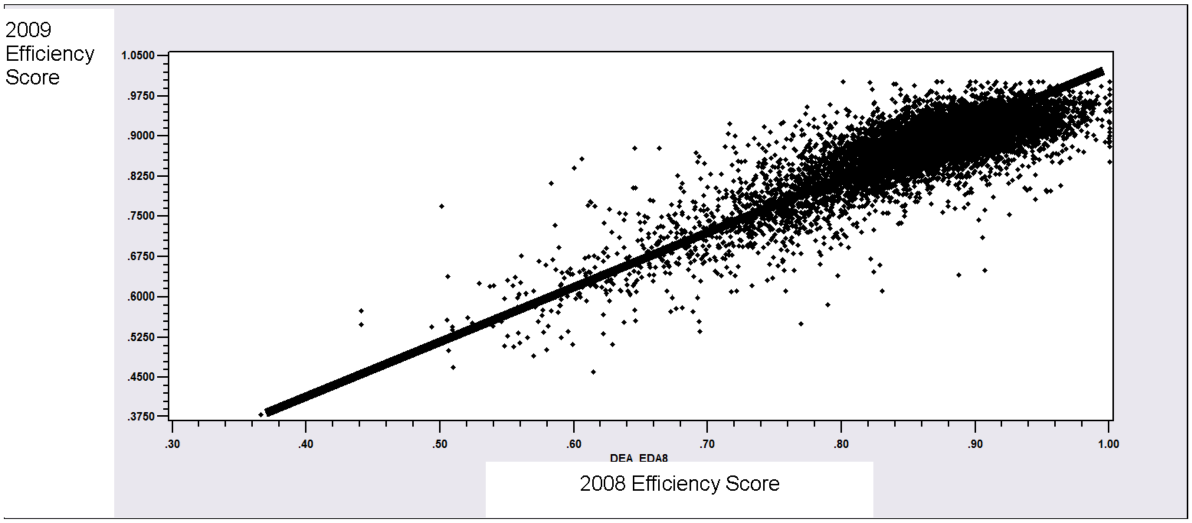
The results in Figure 6.4 could be due to:

- A decrease in diabetic monitoring holding cost constant.
- An increase in risk adjusted cost, with diabetic monitoring held constant.
- Favorable changes in both cost and quality, but with a smaller change in quality.
- Unfavorable changes in both cost and quality.

In fact all the diabetic monitoring measures increased slightly from 2008 to 2009, but so did risk adjusted cost, a result along the lines of the second possibility. The *ratios* of HbA1c monitoring and medical attention for nephropathy to risk adjusted cost increased from 2008 to 2009, but the ratio of LDL-C to risk adjusted cost fell over the same period, which could account for the overall decrease in performance.[47]

Figure 6.5 presents changes in efficiency scores for the ED/ACS measures from 2008 to 2009. The results are quite different from those for the diabetes measures. Across the full range of efficiency scores, the average scores for the ED/ACS measures improved slightly from 0.861 in 2008 to 0.877 in 2009. The improvement in efficiency can be traced to uniform improvement in all the ED and ACS variables from 2008 to 2009. Risk adjusted cost *increased* slightly from 2008 to 2009.

Figure 6.5: Comparison of 2008 and 2009 input efficiency for ED utilization and ACS admissions (five state data)



We examined the stability of TINs in the lowest and highest quartiles of efficiency for the diabetic monitoring and the ED/ACS measures from 2008 to 2009. The results are shown in Tables 6.8 and 6.9.

[47] A portion of the decline in input efficiency scores from 2008 to 2009 also could be due to regression to the mean. If some of the factors affecting input efficiency are somewhat random, then TINs with relatively higher values of those random factors in 2008 – possibly including TINs with efficiency scores above the 2008 mean – would be expected to have lower efficiency scores in 2009. The dotted line in Figure 6b.3 marks the mean of the input efficiency scores in 2008. Points to the right of the line represent TINs with higher than average efficiency scores in 2008 and indeed it appears that those TINs are disproportionately more likely to have lower efficiency scores in 2009 than in 2008. However, regression to the mean should have the opposite effect for TINs below the mean efficiency score for 2008. Their 2009 efficiency scores should be disproportionately *more* likely to increase from 2008 to 2009, and that does not appear to be the case.

- For diabetic monitoring, membership in the lowest quartile of efficiency was more stable than membership in the highest quartile. Among the 2,907 TINs in the highest quartile of diabetic monitoring performance in 2008, 1,923 or 66 percent remained in the top quartile in 2009. However, among 2,909 lowest quartile TINs in 2008, 2,245 or 77 percent remained in the lowest quartile in 2009.

Table 6.8: The Stability of the Diabetes Monitoring DEA Input Efficiency Measure From 2008 to 2009 (Five state data) Quartile 1 = Lowest Performance Quartile

		2009 performance quartile (1=lowest performance quartile)				
		Total	Quartile 1	Quartile 2	Quartile 3	Quartile 4
2008 performance quartile (1=lowest performance quartile)	Quartile 1	2909	2245	487	123	54
	Quartile 2	2908	494	1422	759	233
	Quartile 3	2909	128	784	1300	697
	Quartile 4	2911	43	214	731	1923
	Total	11637	2910	2907	2913	2907

- The results were similar for the combined ED and ACS measure. Among the 2,936 TINs in the highest quartile in 2008, 1,726 or 59 percent remained in the highest quartile in 2009. Of the 2,935 TINs in the lowest quarter in 2008, 2,162 or 74 percent remained in the lowest quartile in 2009.

Table 6.9: The Stability of the Combined ED and ACS DEA Input Efficiency Measure From 2008 to 2009 (five state data) Quartile 1 = Lowest Performance Quartile

		2009 performance quartile (1=lowest performance quartile)				
		Total	Quartile 1	Quartile 2	Quartile 3	Quartile 4
2008 performance quartile (1=lowest performance quartile)	Quartile 1	2935	2162	536	150	87
	Quartile 2	2936	504	1299	801	332
	Quartile 3	2934	173	773	1198	790
	Quartile 4	2936	97	325	788	1726
	Total	11741	2936	2933	2937	2935

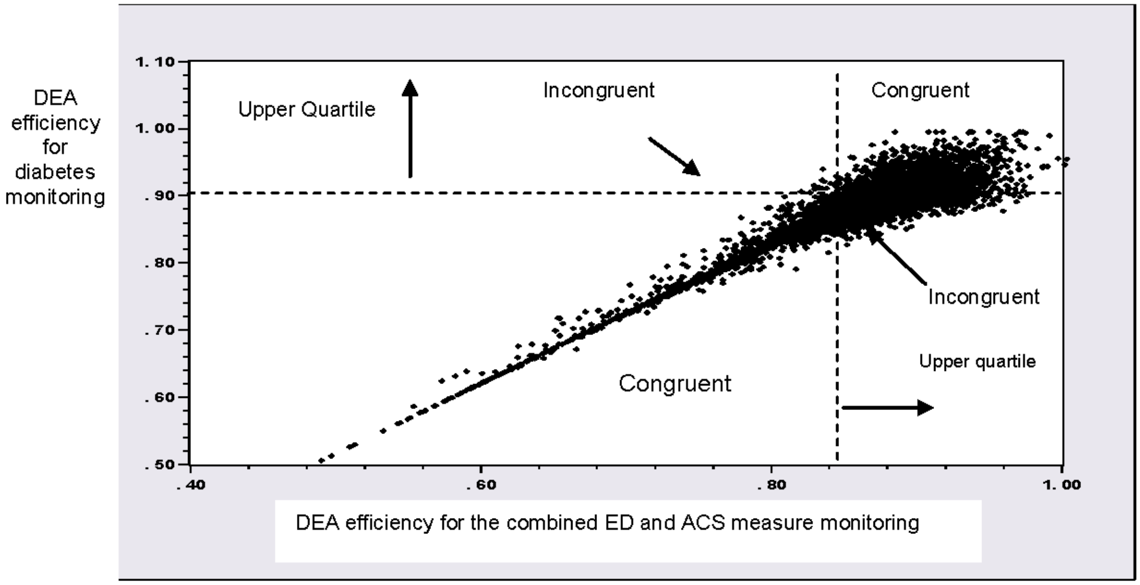
The results suggest that while there is movement among quartiles from year to year, there appears to be more stability in membership in the highest and lowest quartiles.

C. Relationship Between Different Efficiency Measures

Earlier we noted that the quality measures were highly correlated within categories such as diabetes monitoring, but not across categories. *As it turns out, however, the DEA input efficiency scores across all the quality measures are much more highly correlated than the quality measures themselves*. For example, the correlation of the DEA input efficiency scores for diabetes monitoring and the combination of ED visits and ACS admissions in the 2008 5 state data is 0.94 for the relatively small sample of 6,782 beneficiaries who appeared in both our subsamples (the beneficiaries eligible for the diabetes measures are a subset of the beneficiaries eligible for the ED and ACS measure). One obvious explanation is that the same annual risk adjusted cost per beneficiary is the input in each case, so one part of the DEA calculation tends to make the measures move together.

This high correlation across DEA input efficiency measures may prove useful for payment policy as discussed below. However, the story is a bit more complicated, as shown in Figure 6.6. The measures for diabetes monitoring and the combination of ED visits and ACS admissions are very highly correlated among low efficiency TINs, but less so for the high efficiency TINs. *In other words, a TIN that performs poorly on one measure performs poorly on the other, but the same relationship does not necessarily hold for membership in the upper quartiles.* The upper quartiles are shown for both measures. We use the term "incongruent" to describe TINs that are in the upper quartile of efficiency for one measure but not the other. Figure 6b.5 shows two incongruent areas. TINs in the upper area are in the upper quartile of DEA input efficiency for diabetes monitoring, but not for the combined ED and ACS measure. TINs in the lower incongruent area are in the upper quartile for the combined ED and ACS measure, but not for the diabetes monitoring measure.

Figure 6.6: The Relationship between Diabetes Monitoring and ED-ACS Quality Measures (2008 five state data) N=6,782



The consistency of the DEA input efficiency scores for poor performing TINs has interesting implications for payment policy. A perennial concern in pay-for-performance systems is that providers might focus only on the rewarded dimensions of quality – the equivalent of "teaching to the test" in educational reform. If TINs that perform poorly on one measure tend to perform poorly on others, then basing incentive payments on DEA efficiency scores could minimize the chance of missing a TIN that happened to excel at an *unmeasured* category of quality. We explored whether the efficiency scores for one measures might be predictive of the actual cost and quality measures for another measure. This point is illustrated in the following table which takes TINs that were in the highest and lowest efficiency quartiles *for diabetes monitoring* and compares the performance of those TINs on the combined ED and ACS measures. The differences in the performance on the combined ED and ACS measures are similar in kind, but not quite as dramatic as they are in Table 6.10 in which the highest and lowest quartiles on ED and ACS efficiency were compared on those specific measures. In the latter case in Table 6b.5, the ED and ACS measures differed by a factor of two to seven; in the table below, comparing performance on the second set of measures, the differences are factor of 1.5 to two. Thus, although the correlation of DEA input efficiency scores is highest among poor performers, high performance on diabetes monitoring also is predictive of relatively high performance on the ED/ACS measure, as well.

Table 6.10: Comparison of Cost and ED/ACS Measures for High and Low Performing TINs on Diabetes Monitoring (2008 five state data)

	Mean for high performers on diabetic monitoring (upper quartile)	Standard deviation for high performers	Mean for low performers on diabetes monitoring (lower quartile)	Standard deviation for low performers
Risk-adjusted cost	$4,751.21 below overall average	1,939.64	7,496.05 above overall average	9,907.94
ED visits per beneficiary	0.4969	0.2582	1.0186	0.5035
Non-emergent ED visits per beneficiary	0.0713	0.0551	0.1083	0.0660
Primary care treatable ED visits per beneficiary	0.0944	0.0571	0.1614	0.0785
Preventable or avoidable ED visits per beneficiary	0.0459	0.0350	0.1021	0.0673
Sample Size	1,718		1,699	

D. Peer Grouping – Adjusting for TIN Characteristics and Patient Sociodemographic Characteristics

The term “peer grouping” can be both ambiguous and controversial in health policy discussions. The concept behind peer grouping is to “level the playing field” when making performance comparisons across different physicians. Although the health services research field has not settled on formal definitions for the terms, “risk adjustment” usually refers to adjustment for the effects of patient clinical and demographic characteristics (e.g., age and sex) and perhaps sociodemographic characteristics, as well. “Peer grouping” usually refers to dimensions of similarity beyond patient clinical and demographic characteristics. For example, a physician practicing in a rural area might be compared only to other rural physicians.

Peer grouping is related to risk adjustment, because the reason that an analyst would propose that rural physicians be compared only to other rural physicians is that “rural” stands in for some variables that are beyond the physician’s control and unobserved by the analyst, that are likely to affect the physician’s performance measures. Beneficiaries in rural areas may have more difficulty accessing health care services. The care prescribed by their physicians or their hospital stays may reflect their longer travel times.

The selection of variables either for risk adjustment or peer grouping is, in part, a political choice, and for that reason, controversial. For example, in our analyses, as long as the physician meets the attribution rule for a beneficiary (i.e., is providing the plurality of the beneficiary’s non-hospital evaluation and management visits), we do not make any adjustment for the physician’s specialty. Thus, if a cardiologist is providing the plurality of a beneficiary’s E&M visits and the cardiologist’s bills are higher than those of a physician in family medicine, the cardiologist will have a lower performance score, holding the quality measures constant. Said another way, physicians with higher fee levels will need to provider higher quality care in order to remain competitive with physicians with lower fee levels. Obviously, that decision is controversial, though it isn't a necessary part of the method we have proposed. CMS could make a different choice.

However, the discussion above alluded to one principle that should be universally accepted, at least in theory – the risk adjustment or peer grouping variables should be variables that are not under the physician's control, i.e., they should be factors that are *exogenous* to the physician. But the operational definition of exogeneity also can be controversial. What factors truly are beyond the physician's control? Most physicians and most analysts would be comfortable saying that once a physician has located her practice in a rural area, that rural location is exogenous to the physician. Thus, if there are characteristics of rural locations, including the beneficiaries living in rural locations, that affect physician performance, then those factors should be controlled in performance assessments. But what about inpatient readmissions? How much control does the outpatient physician who is assigned a beneficiary under the plurality attribution rule have over the likelihood that a beneficiary needs to be rehospitalized? We put these questions to our beneficiary level TEP and the TEP expressed significant enthusiasm for a more expansive view of physician responsibility. That enthusiasm is reflected in our choice of risk adjustment and peer grouping variables that are described in greater detail later in this section:

- The HCC and demographic variables for risk adjustment.
- TIN and patient sociodemographic characteristics for peer grouping:
    - Income
    - Education
    - Urbanicity
    - State fixed effects

Earlier in this chapter we noted that the quality measures are adjusted by the inclusion of specifically eligible patients in the denominators of the quality measures such as HbA1c testing and breast cancer screening. However, it might be desirable to adjust, at some point in the analysis, for additional characteristics of either the TIN or the beneficiary.

Some patients are more difficult to manage than others. Some patients may not keep their appointments, and may be unwilling or unable to comply with recommended treatments. Some patients simply may view the emergency department as their usual source of care, affecting adversely the physician's cost and quality variables.

If some patient characteristics affect physician performance scores and *if those patient characteristics are beyond the physician's control* (an important qualifier), then failure to adjust for those characteristics will penalize physicians who treat difficult or vulnerable patients. However, any attempt to adjust for those additional variables will encounter several problems. First, beyond the broadest generalizations it is hard to define difficult-to-manage patients reliably and extremely difficult to identify them in administrative data. Second, even if one could identify them or their characteristics, there is a concern that adjusting for those additional characteristics could represent implicit acceptance of a lower standard of care for difficult populations. For example, if inappropriate use of the ED is more likely for more indigent populations, adjustment for that tendency will in effect sanction a different standard of care for the poor.

To our knowledge, this problem does not have a good solution at this time. Our beneficiary level TEP agreed with that assessment. If one were to adjust for additional peer grouping variables there are two places the adjustment might occur. The first is in the individual cost and quality measures and the second is in the performance score that combines the cost and quality measures.

If adjustments are to be made for additional peer grouping variables that run the risk of implicitly accepting a lower standard of care for vulnerable beneficiaries, then our recommendation at this point in the development of value based payment systems is to adjust the DEA performance score (or some other combination of cost and quality), rather than all the individual cost and quality ingredients to the DEA performance score. Adjusting the combined cost and quality performance score maximizes the number of ways in which the physician can improve her performance. A low performing physician can improve her performance either by reducing risk-adjusted cost or by improving the quality measures – separately or as a group. However, we emphasize that nothing in our system precludes adjusting either the individual cost or quality measures or the overall performance measures for beneficiary sociodemographic characteristics if that is a preferred policy alternative.

If the peer grouping variables are found to have a significant effect on performance scores, Medicare would have at least four payment policy options (in addition to making no payment adjustment):

- The performance scores could be adjusted for the peer grouping variables prior to computing the incentive payments. For example, a multiplier could be applied to the incentive payment for physicians with more vulnerable beneficiaries or who practice in geographic areas with a higher likelihood of containing difficult beneficiaries. Physicians with a vulnerable patient population who were not able to overcome the difficulties they face would receive no reward, but they would face an increased incentive to try innovative approaches to caring for vulnerable beneficiaries.

- Instead of adjusting the scores themselves or altering the “normal” incentive payment system, Medicare could provide a lump sum payment per attributed beneficiary to physicians with more vulnerable patients or who practice in geographic areas with a higher likelihood of containing vulnerable beneficiaries. The physician could determine how best to use the lump sum payment to improve his or her performance score. Of course, a physician simply could pocket the lump sum payment and still refuse to see difficult patients, even though his or her office was located in a geographic area filled with such patients. The relative amounts of the lump sum payment and incentive payments would have to be considered carefully to minimize the likelihood of such undesirable outcomes.

- Physicians could be allowed to exclude a certain number or percent of their patients from the computation of performance scores. Currently, physicians participating in the PQRS reporting system must report on only eighty percent of their patients eligible for a particular PQRS measure. However, such a system runs the risk of implicitly accepting a lower standard of care for the excluded patients.

All these proposals are, at best, very imperfect solutions to a very difficult problem. Before choosing among these policy alternatives, however, the first step is to test the effect of a sample of peer grouping variables on the performance scores. We investigate the effect of several possible peer grouping variables on the DEA performance scores in the next section.

E. The Effect of Peer Grouping Variables on Performance Scores

We investigate two types of peer grouping variables: TIN characteristics and beneficiary sociodemographic characteristics. The TIN characteristics we investigate are urbanicity and the state in which the practice is located. The beneficiary sociodemographic characteristics are education and income (measured at the county level).

The urbanicity variable is based on the zip-code level rural-urban community area (RUCA) codes developed by the Economic Research Service (ERS). RUCAs use Census tract population and work commuting information to develop a 10 point scale of rural and urban status. The most metropolitan end of the scale is 1 and the most rural end is 10. Our TIN-level measure of urbanicity is the average RUCA rating for the counties of residence among beneficiaries assigned to the TIN.[48] Urbanicity clearly could be related to exogenous beneficiary characteristics that could affect physician performance scores.

The role of the state variables is less clear, and we do not have a strong recommendation to include or exclude them from the peer group analysis. We would not recommend adjusting incentive payments for geographic variation, per se, unless that geographic variation is shown to result in better care and better health outcomes. We also are concerned about recent studies that show systematic geographic variation in coding of similar patients (Song, et al., 2010). We are aware of a current Institute of Medicine study that will examine geographic variation in cost in great detail, and our recommendation is to use the findings from that study to make better informed decisions about how to accommodate geographic variations in performance, if at all.

The education and income variables were defined as follows. For each beneficiary assigned to the TIN, we collected the median number of years of education and median income (expressed in $1,000s) for residents of the county in which the beneficiary lived. The data are from the 2000 Census and thus are quite dated, but more recent data soon will be available from the 2010 Census.

The first question, however, is the extent to which the TIN characteristics and beneficiary sociodemographic characteristics actually affect the performance score. We addressed that question by regressing the 2008 performance scores for the diabetes monitoring and ED/ACS measures on characteristics of the TIN and the beneficiary.

[48] Some analysts reduce the RUCA code to a “rural” binary variable. Rural equals one if the RUCA is greater than 6.

The regression results for the diabetes monitoring performance scores are shown in Table 6.11. The beneficiaries' level of education was positively associated with the TIN's input efficiency score, but income was not. TINs in Colorado had significantly higher efficiency scores than TINs in Florida.

Table 6.11: Regression of DEA Input Efficiency for Diabetes Monitoring on Characteristics of the TIN and the Beneficiary (2008 5 state data) N=11,637

Variable	Coefficient	Standard Error	t-statistic	P|t|>0
Constant	0.8053	0.0100	80.17	0.00
Median years of education (county)	0.0037	0.0009	4.09	0.00
Median income ($1,000s) (county)	0.0001	0.00008	1.46	0.14
Urbanicity (1-10) Urban=1, Rural=10	-0.0001	0.0006	-0.26	0.80
California	-0.0009	0.0015	-0.59	0.55
New Jersey	-0.0019	0.0018	-1.04	0.30
North Dakota	0.0140	0.0104	1.35	0.18
Colorado	0.0124	0.0029	4.29	0.00

We repeated the regression analysis for the ED/ACS measures. The results are shown in Table 6.12. In this case, input efficiency is positively associated with both income and education. TINs serving beneficiaries in more rural areas again have higher efficiency scores, but the coefficient is significant at only the 0.06 level. The mean efficiency scores for the states are not statistically significant from Florida, the omitted reference category.

Table 6.12: Regression of DEA input efficiency for the combined ED and ACS measure on characteristics of the TIN and the beneficiary (2008 5 state data) N=11,741

Variable	Coefficient	Standard Error	t-statistic	P|t|>0
Constant	0.7199	0.0119	60.76	0.00
Median years of education (county)	0.0100	0.0010	9.32	0.00
Median income ($1,000s) (county)	0.0004	0.0001	4.05	0.02
Urbanicity (1-10) Urban=1, Rural=10	0.0012	0.0007	1.86	0.06
California	0.0001	0.0017	0.07	0.95
New Jersey	-0.0017	0.0021	-0.82	0.41
North Dakota	-0.0103	0.0108	-0.95	0.34
Colorado	0.0049	0.0031	1.57	0.12

Because the relationship of these variables to efficiency scores could have direct bearing on payment policy, we investigated them further. As in the case of year-to-year differences, we asked whether the effects of each variable on the TIN's efficiency score was due to effects on cost, quality, or both (data not shown).

For the diabetic measures, TINs serving beneficiaries living in areas that had higher education levels and were more rural tended to have both lower risk adjusted cost and lower values of each quality measure. However, these two variables had a greater negative effect on cost than on quality, resulting in a net positive effect on efficiency. The same was true of several state variables, in contrast to Florida, the omitted reference state. All states except North Dakota, with its very small sample, had lower risk adjusted costs than Florida and lower values of the quality measures, but the *net* effect is that TINs in California, New Jersey and Colorado had significantly higher efficiency scores than those in Florida.

The opposite effect holds for income in the diabetes monitoring analysis. TINs serving beneficiaries in areas of higher income tended to have higher values of both cost and quality, but in this case, the effects appear to have offset each other so that the net effect of income on efficiency was not statistically different from zero.

For the combined ED and ACS measure, TIN's serving beneficiaries living in areas with higher education levels or more rural areas tended to have lower cost. Education level was positively associated with overall use of the ED, but negatively associated with unnecessary ED utilization. The implication is that negative effect of education on both cost and unnecessary ED utilization was sufficient to outweigh the positive effect of education on overall ED use. In addition to negative effects on cost, more rural areas tended to have lower overall ED utilization, but more avoidable ED utilization. However, the overall positive effect of urbanicity on efficiency suggests that the negative effects on cost outweighed the positive effects on ED/ACS utilization. The rural results highlight the need to calibrate the Billings algorithm to rural areas. It is possible that some of the ED utilization that is designated "avoidable" in New York City by the Billings algorithm is not as avoidable in rural areas. Income was positively associated with cost and overall ED use, but negatively associated with avoidable ED use, which explains the overall positive effect on efficiency.

This analysis suggests several concluding points regarding DEA input efficiency measures and payment policy:

- Although in general, DEA input efficiency is positively correlated with quality and negatively correlated with cost, there are some TINs that achieve high efficiency for some measures through lower cost, accompanied by lower quality. An example is diabetes monitoring by rural TINs.
- The effects of sociodemographic variables and TIN characteristics on DEA input efficiency is not uniform across measures. For example, income has no effect on the diabetes monitoring measure, but a significant positive effect on the combined ED and ACS measure.
- Improvement in some measures may be best achieved not through payment incentives to providers, but through changes in the FFS Medicare benefit package. For example, given the positive effect of income on overall ED use, Medicare might consider a larger out-of-pocket differential between avoidable utilization of the ED versus utilization of a physician's office, particularly in light of recent interest in limiting the ability of supplementary insurance policies to eliminate the effect of point-of-purchase cost-sharing.

F. Does Adjustment for Peer Grouping Variables Affect Quartiles of Performance?

If incentive payments were made to TINs in the upper quartile of DEA input efficiency, how much difference would it make if TIN characteristics and beneficiary sociodemographic characteristics were taken into account? We addressed that question by regressing the DEA input efficiency scores for the diabetic monitoring and ED/ACS measures on the same variables shown in the Tables 6b.6 and 6b.7. The residuals from that regression represent the DEA input efficiency scores, adjusting for the TIN and beneficiary sociodemographic variables.

We then compared membership in the four quartiles of efficiency from the model that corrects for those additional variables versus the model that does not. The results for the diabetes monitoring performance measure are shown in the Table 6.13.

Adjustment for the additional variables did not move any TIN more than one quartile. Of the 2,910 beneficiaries in the lowest quartile of *unadjusted* diabetes monitoring performance in 2008, 2,778 or 95 percent remained in the lowest quartile after adjustment for the additional variables. Of the 2,909 TINs in the highest quartile prior to adjustment for the additional variables, 2,773 or 95 percent remained in the highest quartile after adjustment. The largest movement was between the two middle quartiles. Further analysis showed that most of the movement was generated by the education, income, urbanicity, and the California state variable. TINs serving beneficiaries in areas with higher education, income and more rural beneficiaries tended to move down (towards greater inefficiency), while TINs in California tended to move up to a higher level of efficiency.

Table 6.13: The Effect on Quartile ranking of Diabetes Monitoring Performance Resulting from Adjustment for TIN Characteristics and Beneficiary Sociodemographic Characteristics (2008 five state data) Quartile 1 = lowest performance quartile

		2008 unadjusted performance quartile (1=lowest performance scores)				
		Total	Quartile 1	Quartile 2	Quartile 3	Quartile 4
2008 adjusted performance quartile (1=lowest performance quartile)	Quartile 1	2,911	2,778	133	0	0
	Quartile 2	2,907	132	2,524	251	0
	Quartile 3	2,910	0	251	2,523	136
	Quartile 4	2,909	0	0	136	2,773
	Total	11,637	2,910	2,908	2,910	2,909

We repeated the analysis for the ED/ACS measure. The results are shown in Table 6.14. The amount of movement resulting from the additional variables was similar to the results for the diabetes monitoring measure, although 14 of the 11,637 TINs moved more than one decile. Again, the largest movement was between the two middle quartiles.

Further analysis showed that most of the movement was generated by the education, income, urbanicity, and all the state variables except California. TINs serving areas with higher levels of education, income, and more rural beneficiaries tended to move to a lower quartile when the additional variables were added. TINs in New Jersey and North Dakota tended to move up, relative to Florida, while TINs in Colorado tended to move down.

Table 6.14: The effect of TIN and beneficiary characteristics on DEA input efficiency – combined ED and ACS measure (2008 five state data) Quartile 1 = lowest performance quartile

		2008 *unadjusted* performance quartile (1=lowest performance quartile)				
		Total	**Quartile 1**	**Quartile 2**	**Quartile 3**	**Quartile 4**
2008 *adjusted* performance quartile (1=lowest performance quartile)	**Quartile 1**	**2,935**	**2,673**	262	0	0
	Quartile 2	**2,935**	261	**2,195**	477	2
	Quartile 3	**2,935**	1	467	**2,045**	422
	Quartile 4	**2,936**	0	11	413	**2,512**
	Total	**11,741**	**2,935**	**2,935**	**2,935**	**2,936**

Analysis of these two performance measures shows that adjustment for these additional variables has different effects for different performance measures. As noted earlier, income does not matter for diabetes monitoring performance, but does matter for the ED/ACS measure. If CMS adopts our recommendation to give separate incentive payments for different performance measures, *adjustments for the peer grouping variables will need to be tailored to each performance measure*.

These results have mixed policy implications for CMS. Overall, the results of this analysis are somewhat reassuring from a policy perspective because they suggest that further adjustment for additional variables exogenous to the physician would not change the quartile rankings very much. However, that does not mean that the changes are inconsequential from a payment perspective. Movement among quartiles is not inconsequential for the TINs who move into or out of a level of performance that generates an incentive payment. Nor does adjustment have an inconsequential influence on the perceived fairness of the incentive payment system. *For that reason, our recommendation is that CMS adjust the final combined cost and quality measures for a carefully selected set of additional exogenous variables, even if the adjustment has little effect on the actual payments to TINs*.

G. Converting the Efficiency Scores into Payment Policy

There are several ways to use the input efficiency scores to compute incentive payments. First, the input efficiency score could be used as a continuous measure on which to compute the fee update. Like the RBRVU and DRG weights, this continuous variable (with an upper bound of 1) could be multiplied by a feasible dollar figure that conformed to the current budget constraint on increased Medicare spending on physician services, to determine the fee increase for individual TINs. In this manner, CMS would give a greater reward *to practices that are closer to the DEA "frontier" that defines the best practices*.

A second alternative is to divide the input efficiency scores into quantiles (e.g., quartiles) and base fee updates on the TIN's rank. This approach is simply a discrete-variable version of the first option. Medicare could choose a cutoff quantile and give increases to TINs that were at or above that quantile. [49]

Third, information from the DEA analysis could be combined with incentives linked to individual cost or quality measures – perhaps denying updates to TINs whose performance on a given quality measure is lower than a predetermined minimum, or giving extra incentive payments to TINs that perform well on specific measures.

As discussed in greater detail later in this chapter, any of these approaches results in information that is easily interpreted by providers. Providers do not have to be familiar with, or even aware of, the DEA methodology (though obviously their interest will increase if the DEA scores are linked to payment policy). The DEA score can be presented as a performance score that combines cost and quality, and each provider could be presented with its rank on the overall performance measure or on individual cost and quality measures. In that way, the provider could see immediately which cost or quality measure contributed the most to its overall efficiency ranking.

Regardless of the way in which the DEA results are linked to payment policy, it is important to emphasize that all of these approaches would reward practices that are producing relatively higher quality care at relatively lower cost using a methodology that:

- can include multiple measures of cost and quality;

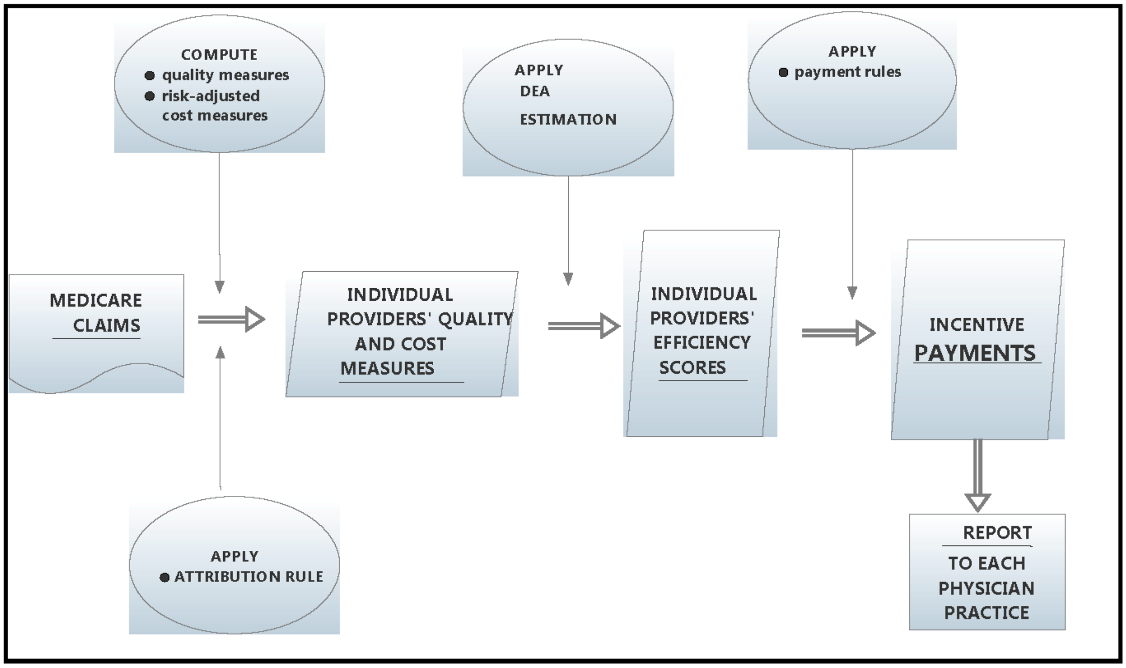
[49] If CMS desired instead to <u>avoid penalizing practices that aren't "too different" from the best practices</u>, it could adopt a rule allowing some distance from the frontier before penalties applied: e.g., give fee increases to practices whose input efficiencies were not statistically different from the frontier value of 1.0. Such a rule would treat DEA results in a manner similar to the hospital mortality and readmission data reported by CMS.

- permits complete flexibility in the measures chosen;
- is practical to implement from current data;
- permits reports to physicians/TINS that identify clearly why overall efficiency scores are high or low.

Figure 6.7 below shows the basic steps in an implementation of DEA analysis for computing the fee updates and reporting to physicians. There are five basic steps:

- Medicare claims are input to the process. The attribution rule is applied to them, and quality measures and risk adjusted cost measures are computed.
- These computations yield individual quality and cost measures at the TIN level, to which the DEA algorithm is applied.
- The result is input efficiency scores for each TIN, to which the payment rules are applied.
- The payment rules yield incentive payments to each TIN.
- A report is prepared so that the measurement system gives intelligible, actionable results to the TIN.

Figure 6.7: Basic Steps from Claims Data to Fee Updates Using DEA


COMPUTE
● quality measures
● risk-adjusted cost measures
APPLY DEA ESTIMATION
APPLY
● payment rules
MEDICARE CLAIMS
INDIVIDUAL PROVIDERS' QUALITY AND COST MEASURES
INDIVIDUAL PROVIDERS' EFFICIENCY SCORES
INCENTIVE PAYMENTS
APPLY
● ATTRIBUTION RULE
REPORT TO EACH PHYSICIAN PRACTICE

This would be a major improvement over current practices. It would also be a major improvement over simpler methods (e.g., tiers or quadrants) that have been used or proposed in the past. CMS would have defensible, sophisticated and flexible measure that presented providers with information that was fair, actionable and meaningful.

H. Additional Observations on DEA Input Efficiency Scores and Payment Policy

In this section we summarize some features of the DEA performance measures that have implications for payment policy. These features are not unique to the DEA approach, but apply to any method of combining cost and quality measures into a single performance score.

- Combining cost and quality into a performance score necessarily sacrifices information on the level of the cost and individual quality measures. An efficient TIN (e.g., a TIN with a high performance score that combines cost and quality) could be efficient at producing either low or high quality care at low cost. In our analyses, we have found that risk adjusted cost decreases and the quality measures increase as the DEA combined performance score increases, but that is not a necessary result.

- Measures that are uniformly low or high across all TINs will not be useful in distinguishing performance among TINs. For example, in our sample of TINs that had 20 or more diabetic beneficiaries, the mean of the quality variable representing recommended eye exams was 0.99 with a standard deviation of 0.03. Including eye exams among the diabetes quality measures would have no noticeable effect on the outcomes of DEA analysis or any other method of combining cost and quality into a performance measure. While that result may not seem troublesome for a measure that exhibits virtually full compliance, it is more troublesome for a measure that exhibits virtually full non-compliance. If CMS were to identify such measures and if improvement in those measures was determined to be a high priority, our recommendation would be to target those measures with special improvement rewards through a system separate from the "normal" system based on combined cost and quality measures.

- Uniformly reweighting the quality measures across all TINs does not change the results of the DEA analysis. In other words, multiplying the HbA1c measure by 5 and the LDL-C measure by 0.5 for all TINs will not affect the results of the DEA analysis, because DEA considers the distance to the frontier for each quality measure for each TIN *relative to the other TIN's in the sample*. If CMS wants to give greater weight to one diabetic quality measure than another in a way that would commensurately change the results, there are two possible approaches that apply to DEA or any other combination of cost and quality:

    - Create a new combined diabetic quality measure prior to the DEA analysis, applying the desired weights to the individual quality measures.

    - In addition to the incentive payments based on DEA performance measures, make supplementary payments for individual quality measures that are determined to be of higher priority. That approach preserves the notion of efficiency, but assigns greater priority to some quality measures than to others.

- It will be important for CMS to pay close attention to the developing literature on provider responses to incentive payments. An important issue is paying for achievement of a *level* of performance versus paying for *improvement*. A perennial concern is that if incentive payments are based on levels of performance that are too high, poor performers will give up trying to improve. Dowd, Feldman and Nersesian (2011) examined that issue in the context of generic drug prescribing in a physician network in the Twin Cities. Despite the fact that incentive payments in that system were based on levels of performance rather than on improvement, it appeared that poor performing physicians did not give up unless they were at exceptionally low levels of performance (and that result may be due more to the character of those particular practices than to any underlying behavioral problems with rewarding levels of performance). In fact, the data suggest that the performance levels triggering incentive payments could have been set higher. The best approach is likely to vary by quality measure. If the quality measures are uniformly low for all TINs, then it might be advisable to pay for *improvement*. Otherwise, setting realistic *levels* for incentive payments is a simpler and more transparent approach that may not carry a great risk that poor performers will give up.

- On a related point, little is known about the cost to physicians of improving different types of performance scores. Improving some types of measures could be relatively inexpensive, while others could be very costly. Marginal improvements for poor performers could be either more or less expensive than the same marginal improvement for good performers. Setting incentive payments intelligently will require a more developed literature on the physician’s cost of making improvements in cost and quality.

- Our system allows certain types of double counting. The same beneficiary can help a TIN earn an incentive payment for good care of diabetes and for avoidable utilization of services. Due to the unique assignment of beneficiaries to TINs under the attribution rule we have chosen, however, the same beneficiary cannot help two different TINs earn incentive payments under the beneficiary level approach. However, the same beneficiary could help either the same TIN or two different TINs earn incentive payments under the beneficiary level approach and again under the MS-DRG approach if the TIN included physicians who treated the beneficiary for inpatient care under the MS-DRG approach and also was assigned the beneficiary by the plurality rule under the beneficiary approach.

I. Reporting Performance Measures to Physicians

Although the development of a format to report performance results to physicians was not a formal part of this project, the reporting format can have an important bearing on whether the metrics that make up a value based payment system are actionable. The challenge here is to make clear to physicians what measures they need to improve in order to reach a higher level of performance and perhaps qualify for an incentive payment.

Table 6.15 shows the data that could be reported to a TIN on the two measures we have examined in detail: diabetes monitoring and the combination of ED visits and ACS admissions. The data are from an actual TIN in the 2008 5 state data. We selected a TIN that performed in the lowest quartile of DEA input efficiency for diabetes monitoring, but within that group of low performers, we purposely chose the TIN with the highest diabetes monitoring efficiency score to avoid inflating the correspondence between the two measures (see Figure 6b.6 above and accompanying discussion). The top part of the table shows the decile (1 through 10) rankings[50] for:

- overall performance for diabetes monitoring that combines cost and quality (the DEA input efficiency score);
- risk adjusted cost; and
- the diabetes monitoring measures.

The decile rankings for the diabetes measures are based on TINs that had 20 or more attributed diabetic beneficiaries. The lowest (worst) decile is coded 1 and the highest (best) decile is coded 10.

Next, for the same TIN, we computed the decile ranking for the combination of ED visits and ACS admissions, using the deciles computed across all physicians with 20 or more total attributed beneficiaries. The results are shown in the lower portion of the table. Again, we present the decile rank for the DEA input efficiency score and each cost and quality measure individually. The TIN performs much better on the combined ED/ACS measure – better than we would expect the average TIN in the lower quartile of diabetes to perform on this measure. However, it still does not fall in the upper quartile of efficiency on the ED/ACS measure, due mainly to its relatively high risk adjusted cost per beneficiary.

The results are interpreted as follows:

- When compared to other TINs with at least 20 attributed *diabetic* beneficiaries, this TIN falls in the lower 30 percent of TINs on the diabetic monitoring measures, which are computed only for diabetic beneficiaries.

However, when compared to other TINs with at least 20 attributed beneficiaries of *all* types, the TIN performs much better on measures of ED visits and ACS admissions, which are compute across all beneficiaries.

[50] We could have used quartiles (1 through 4) as in the earlier tables, but the decile (1 through 10) rank is closer to a continuous measure. One appealing feature of quartiles is that they could be converted easily into “stars” similar to those used in ratings of Medicare Advantage plans.

Table 6.15: An Example of a Physician Performance Report

Physician/TIN identifier	**Score**	**Rating: 10 = top 10 percent, 1=bottom 10 percent**
Diabetes monitoring Peer group: TINs with 20 or more diabetic beneficiaries		
Overall Performance (cost and quality)	87 (maximum = 100)	3
Risk-adjusted cost	$372 below average	3
Quality measures:		
Hba1c test	31 percent of diabetic beneficiaries	1
LDL-C test	35 percent of diabetic beneficiaries	2
Medical attention for nephropathy	22 percent of diabetic beneficiaries	1
Emergency department utilization and ambulatory care sensitive admissions Peer group: TINs with 20 or more beneficiaries		
Overall performance (cost and quality)	85 (maximum = 100)	4
Risk-adjusted cost	$372 below average	3
Quality measures:		
ED visits per beneficiary	56 visits per 100 beneficiaries	5
Non-emergent ED visits per beneficiary	6 visits per 100 beneficiaries	6
Primary care treatable ED visits per beneficiary	8 visits per 100 beneficiaries	7
Preventable or avoidable ED visits per beneficiary	3 visits per 100 beneficiaries	8
Ambulatory care sensitive admissions per beneficiary	8 visits per 100 beneficiaries	9

Chapter 7
Recommendations

In this chapter we present a summary of our recommendations for a value based payment system for physicians in the Medicare program. In order to provide a concise summary, we have not tried to include discussion of the reasoning and empirical results that underlie each recommendations. That material can be found in the body of the report. Without that context, the recommendations may be less helpful to a reader new to value based purchasing systems, but we hope that presenting the recommendations in this manner will provide the experienced reader with an easily accessible guide to our position on some of the most important and sometimes controversial issues in value based purchasing systems.

I. General

A. Value based incentive payments

1. We recommend that decisions regarding value based incentive payments be separated from any across-the-board changes in physician fee levels. Value based incentive payments should reward good performance, not general increases in the cost of doing business.

2. In order to encourage coordination of care, payments should be awarded at some organizational level greater than an individual physician, such as a tax ID number (TIN).

3. Incentive payments should be given only to physicians (TINs) who have enough attributed and eligible beneficiaries to produce a reliable estimate of cost and quality for each measure. Physicians with few attributed Medicare beneficiaries for a given measure would not be eligible for an incentive payments for that measure.

4. We recommend separate incentive payments for each category of quality measures rather than building a composite quality score that requires the physician to have adequate numbers of patients in many different categories. By "category" we mean, for example, monitoring of diabetic beneficiaries, breast cancer screening for eligible female beneficiaries, or potentially preventable rehospitalizations for hospitalized beneficiaries. Requiring a physician to have an adequate number of diabetics *and* female beneficiaries *and* hospitalized beneficiaries, etc., in order to have adequate sample sizes for a composite quality measure would greatly reduce the number of physicians eligible for an incentive payment. (This recommendation does not preclude using several different quality metrics to create a single performance index for a category measure, e.g., using multiple metrics of diabetes monitoring to form a single measure of diabetes monitoring as illustrated by our analyses in this report.)

B. Algorithms for risk adjusted cost and quality

We recommend that, to the extent possible, CMS use publicly available algorithms to compute risk adjusted cost and quality measures. We realize that publicly available algorithms facilitate strategic coding, but “black box” algorithms make it more difficult for well-intentioned physicians to improve their cost and quality measures. It would be politically difficult for CMS to say that it doesn’t know why a physician’s quality measure is low because the score is the result of a proprietary black box algorithm.

II. MS-DRG Approach

A. Defining Episode Window: We recommend using 60-day episodes over 30-day episodes because they capture more variation in physician cost and quality.

B. Organizational Unit: We recommend using the tax identification number (TIN) as the organizational unit that is held accountable for the beneficiary’s care and is rewarded for good performance.

C. Attribution: We recommend attributing cost and quality associated with an episode to all physician entities, represented by their TIN, who manage an episode. All costs associated with hospitalization up to 30-/60-days (excluding the DRG cost of the initial hospitalization) can be attributed to TINs based on a proportion of their Part B billing for the index hospitalization. We recommend that all of the quality associated with the episode be attributed to all TINs managing the episode without trying to attribute responsibility proportionately.

D. Risk adjustment

1. We recommend using a *prospective* risk adjustment model using claims one year before the index hospitalization, with appropriate risk adjustment for *severity* (number of prior hospitalizations and ED visits for the related condition). Based on recommendations of the TEP we also suggest that costs associated with trauma within the episode window be risk adjusted for using appropriate condition categories.
2. We recommend that episode level quality measures be risk adjusted using *prospective* risk adjustors.

E. Peer Grouping We recommend comparing TINs’ risk adjusted costs only *after* grouping TINs into peer groups based on the mean proportion of cost associated with the MS-DRG episodes on which they submitted bills.

F. Quality Measures

1. Based on recommendations from the April 2010 TEP we used both episode level and hospital level measures for quality. Hospital level quality measures could be used only when episode level measures are not readily computable from claims (as in the case of AHRQ patient safely indicators). The quality measures can be combined into single composite using weights obtained from physician technical expert panels.

2. We also recommend that both episode and hospital level quality measures be presented in a positive frame - where higher scores reflect better quality in the true sense, so that they highlight good performance and are favorably viewed by physicians.

G. Combining Costs and Quality: Across TIN peer groups, we recommend that a tiering approach be used to combine cost and quality. Within each peer group, TINs can be grouped into a meaningful number of tiers based on their quality and TINs that are below the mean peer-group cost within each quality tier be identified as low-cost TINs.

H. Weighting TIN Performance for Payment: We recommend using a three-step process to distribute total incentive payments for inpatient physicians, where the payment is distributed by (i) type of condition or MS-DRG, (ii) TIN's proportion of responsibility the episode- determined by peer group, (iii) and performance- with respect to both cost and quality, within a peer group.

III. Beneficiary Approach

A. Length of observation period: We recommend using a calendar year as the observation period for the beneficiary analysis.

B. Organizational Unit: We recommend using the tax identification number (TIN) as the organizational unit that is held accountable for the beneficiary's care and is rewarded for good performance.

C. Attribution: We recommend the plurality of non-hospital evaluation and management visits for attribution in the beneficiary-level approach. Better rules may become available in the future. The Physician Quality Reporting System is of particular interest.

D. Risk adjusted cost:

1. Based on results from the 2010 TEP, we recommend using *concurrent* risk adjustors in the beneficiary analysis. However, we recommend *not* adjusting inpatient costs for the condition categories (CCs) that are associated with ambulatory care sensitive (ACS) admissions.
2. We recommend using the *residuals* from risk-adjusted cost equations (combining the residuals from the inpatient and outpatient equations into a single cost measure) rather than a *ratio* of actual versus estimated cost. Residuals represent actual dollars while ratios express deviations from expected cost in percentage terms. We prefer residuals to ratios to ensure that higher cost beneficiaries and higher cost illnesses received greater weight.

E. Quality measures

1. For outpatient physicians, beneficiaries should be considered as whole persons to the extent possible, not divided into separate diseases or episodes of illness. This is particularly true of the Medicare population which is characterized by multiple co-morbidities. Proper risk adjustment can account for the presence of co-morbidities in cost analyses, and practice guidelines and quality of care measures *should* take account of multiple co-morbidities when appropriate before incorporating those measures into value based purchasing systems.

2. If a beneficiary has two conditions for which quality measures exist, then it is appropriate to reward the physician for appropriate care of each condition. Similarly, if the beneficiary is eligible for two different quality measures, the physician can be rewarded for good performance on each quality measure even though the same beneficiary's experience contributed to both measures. This rule would apply both to measures within the beneficiary approach, and across the beneficiary and MS-DRG approaches.

3. The Billings algorithm for avoidable emergency department visits has several shortcomings:

- The algorithm was based on the experience of an emergency department in New York City, and thus may not be entirely transferrable to other locations – e.g., less urban or rural.

- The algorithm does not consider the cumulative or interactive effects of multiple diagnoses.

We recommend that CMS consider addressing these shortcomings if the Billings algorithm is to be incorporated into a value based purchasing system. Our use of the algorithm in its current form thus should be considered illustrative.

F. Combining cost and quality measures

We recommend that DEA efficiency scores be considered as a means of combining cost and quality measures in the Beneficiary Approach. The DEA approach allows risk adjusted cost to be compared to multiple individual quality measures within a group of measures such as diabetes monitoring or avoidable utilization. However, if CMS identifies high priority *individual* measures within any *group* of measures, CMS could target those measures with special improvement rewards through a system separate from the "normal" system based on combined cost and quality measures.

G. Adjusting incentive payments for additional variables

1. CMS should adjust the final combined cost and quality measures for a carefully selected set of additional exogenous variables, even if the adjustment has little effect on the actual payments to TINs. If such adjustments run the risk of implicitly accepting a lower standard of care for vulnerable beneficiaries, then our recommendation at this point in the development of value based payment systems is to adjust the DEA performance score (or some other combination of cost and quality), rather than all the *individual* cost and quality ingredients to the DEA performance score. Adjustments for the peer grouping variables will need to be tailored to each performance measure.

2. Adjustment for geographic variation (as opposed to urbanicity) is controversial. We would not recommend adjusting incentive payments for geographic variation, per se, but instead, adjusting for lower cost, better care, and better health outcomes that might or might not be associated with geographic areas. We are concerned about recent studies that show systematic geographic variation in coding of similar patients. We are aware of a current Institute of Medicine study that will examine geographic variation in cost in great detail, and our recommendation is to use the findings from that study to make better informed decisions about how to accommodate geographic variations in performance, if at all.

References

Agency for Health Research and Quality (AHRQ). Algorithm dowloadable from http://wagner.nyu.edu/faculty/billings/acs-algorithm.php (last accessed May 31, 2011).

Agency for Healthcare Research and Quality. "Talking Quality: Framing Scores as Positive or Negative" available at https://www.talkingquality.ahrq.gov/content/create/scores/framing.aspx (Last accessed March 3 2012)

Andes, S., Metzger, L.M., Kralewski, J. and Dave Gans. "Measuring efficiency of physician practices using data envelopment analysis," *Managed Care* 11:11 (November 2002) 48-54.

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

Buntin, Melinda Beeuwkes, and Alan Zaslavsky. "Too Much Ado About Two-part Models and Transformation? Comparing Methods of Modeling Medicare Expenditures" *Journal of Health Economics* 23 (2004) 525-542.

Center for Medicare and Medicaid Services. Generating Medicare Physician Quality Performance Measurement Results (GEM) Project 2007 specifications http://www.cms.gov/GEM/Downloads/GEMMeasureFunctionalSpecs2007.pdf (accessed May 31, 2011)

Dowd, Bryan E., Feldman, Roger and William Nersesian. "Setting Pay for Performance Targets: Do Poor Performers Give Up?" forthcoming in *Health Economics* (2012).

General Accounting Office (GAO). Medicare Advantage: CMS Should Improve he Accruracy of Risk Score Adjustments for Diagnostic Coding Practices." (January 2010). Available at http://gao.gov/assets/590/587637.pdf.

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

Mehrotra, Ateeve, Adams, John L., Thomas, J. William and Elizabeth A. McGlynn. "The Effect of Different Attribution Rules on Individual Cost Profiles," *Annals of Internal Medicine* 152:10 (May 18, 2010) 649-654.

Mehrotra, Atteve, Adams, John L., Thomas, J. William and Elizabeth A. McGlynn. "Is Physician Cost Profiling Ready for Prime Time?" Rand Corporation: Santa Monica, California (2010). Available at http://www.rand.org/content/dam/rand/pubs/research_briefs/2010/RAND_RB9523-1.pdf

Manning, W.G. and John Mullahy. "Estimating Log Models: To Transform or Not to

Transform," *Journal of Health Economics* 20 (2001) 461-494.

Mathematica Policy Research Inc: Standardized Unit Costs for Use in Resource Use Reports (RURs). Report Submitted to Centers for Medicare and Medicaid Services, under Contract Number HHSM-500-2005-00025I. Mathematica Policy Research Inc. Washington DC. July 2008

Mutter, Ryan L., Rosko, Michael D., Greene, William H., and Paul W. Wilson. "Translating Frontiers into Practice: Taking the Next Steps Toward Improving Hospital Efficiency," *Medical Care Research and Review* Supplement to 68(1) (January 2011) 35-195.

National Quality Forum. *NQF Endorsed Standards.* http://www.qualityforum.org/Measures_List.aspx (last accessed January 25, 2012).

Pope, G.C., Ellis, R.P., Ash, A.S., et al.: Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment. Final Report to the Health Care Financing Administration under Contract Number 500-95-048. Health Economics Research, Inc. Waltham, MA. December, 2000b.

Pope, G.C., Kautter, J., Ellis, R.P., et al.: Risk Adjustment for Medicare Capitation Payments Using the CMS-HCC Model. *Health Care Financing Review* 25(4):119-141, Summer, 2004.

Silber, Jeffrey H., Rosenbaum, Paul R., Brachet, Tanguy J., Ross, Richard N., Bressler, Laura J., Even-Shoshan, Orit, Lorch, Scott A. and Kevin G. Volpp. "The Hospital Compare Mortality Model and the Volume-Outcome Relationship," *Health Services Research* 45:5 Part 1 (2010) 1148-1167.

Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science 30; 211(4481):453-8. Jan 1981.