WORKING
PAPER

Assessment of Pay-for-Performance Options for Medicare Physician Services: Final Report

MELONY E.S.SORBERO, CHERYL L.DAMBERG, REBECCA SHAW, STEPHANIE TELEKI, SUSAN LOVEJOY, ALISON DECRISTOFARO, JAKE DEMBOSKY, CYNTHIA SCHUSTER

WR-391-ASPE

May 2006

Prepared for the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services

HEALTH

PREFACE

In recent years, pay-for-performance (P4P) programs have been developed as a strategy for driving improvements in the quality and cost-efficiency of health care. The Centers for Medicare & Medicaid Services (CMS), is actively considering P4P for Medicare physician services, viewing this policy strategy as one way to increase physician responsibility for efficiently providing high quality care to beneficiaries of the Medicare program.

With an interest in learning more about P4P programs targeted at physicians, the Assistant Secretary for Planning and Evaluation (ASPE), within the U.S. Department of Health and Human Services, contracted with the RAND Corporation to help in its assessment of whether P4P can be effectively implemented in the Medicare physician service delivery and payment environment.

This report presents the results of this study, including a review what is known about P4P and the empirical evidence about its effectiveness, a description of the characteristics of current P4P programs, lessons learned from currently operating P4P programs about how to design and implement these programs, key P4P program design components and an assessment of the options for each component, a framework for guiding the development of a P4P program, the challenges CMS can expect to face in designing and implementing a P4P program for Medicare physician services, as well as steps that CMS could take to prepare for building and supporting a national P4P program for physician services.

The contents of this report will be of interest to national and state policymakers, health care researchers, health plans and providers and others interested in having a better understanding of P4P programs. The study findings also should be useful for individuals and organizations developing P4P programs.

This work was sponsored by ASPE under by Task Order No. HHSP23300001T under Contract No. 100-03-0019, for which Susan Bogasky served as project officer.

CONTENTS

FIGURES

TABLES

SUMMARY

Pay-for-performance (P4P), the practice of paying health care providers differentially based on their quality performance, emerged in the late 1990s as a strategy for driving improvements in health care quality. The Centers for Medicare & Medicaid Services (CMS), the largest purchaser of health care services in the United States, is actively considering P4P for Medicare physician services, viewing this policy strategy as one way to increase physician responsibility for efficiently providing high quality, outcome focused care to beneficiaries of the Medicare program.

In September 2005, the Assistant Secretary for Planning and Evaluation (ASPE), within the U.S. Department of Health and Human Services, contracted with the RAND Corporation to help in its assessment of whether P4P can be effectively implemented in the Medicare physician service delivery and payment environment. RAND's tasks, within the scope of the project, were to

1. Review what is known about P4P and the empirical evidence about its effectiveness.
2. Describe the characteristics of current P4P programs
3. Assess whether features of these programs could help inform development of a P4P program for Medicare physician services.

This report summarizes what we, the RAND study team, did and what our findings were. Specifically, we describe

- Evidence from empirical studies about the effects of P4P.
- Lessons learned from currently operating P4P programs about how to design and implement these programs.
- Key P4P program design components and options.
- A framework for guiding the development of a P4P program.
- Challenges that CMS will face in designing and implementing a P4P program for Medicare physician services.
- Steps that CMS could take to prepare for building and supporting a national P4P program for physician services.

WHY IS MEDICARE INTERESTED IN PAY-FOR-PERFORMANCE?

Medicare's current interest in P4P is motivated by a range of concerns, especially continuing deficits in quality of care, rising health care costs, and the current Medicare fee schedule's inability to control costs.

The Quality Problem

A variety of studies document substantial deficiencies in the quality of care delivered in the United States (Asch et al., 2006; Institute of Medicine, 2001; Schuster et al., 1998; Wenger et al., 2003). A national examination of the quality of care delivered to adults found that they received only about 55 percent of recommended care on average and that adherence to clinically recommended care varied widely across medical conditions (McGlynn et al., 2003).

The Health Care Cost Problem

Health care costs continue to rise at a rapid pace; they are expected to account for nearly 19 percent of gross domestic product by 2014 (Heffler et al., 2005). In 2006, the Federal government will spend $600 billion for Medicare and Medicaid; by 2030, expenditures for these two programs are expected to consume 50 percent of the federal budget, jeopardizing funding for other, discretionary programs (McClellan, 2006). CMS Administrator Dr. Mark McClellan has stated publicly that if the United States is to continue funding these programs, it will need to redesign existing policies and practices.

The Current Medicare Payment Policy Problem

Approximately 484,000 physicians regularly bill for providing Medicare Part B services (MedPAC, 2006). Medicare's FFS payments for physician services follow a resource-based relative value fee schedule (RBRVS). The annual update to the fee schedule is determined by three factors: (1) the rate of change in the Medicare Economic Index (MEI), (2) a price index measuring changes in the costs of maintaining a physician practice, and (3) a sustainable growth rate (SGR) expenditure target. The annual update factor to the physician fee schedule is adjusted based on a comparison of cumulative past actual expenditures with the SGR.

Although the SGR was established as an expenditure control mechanism, the SGR target has been routinely exceeded because it is applied at the national level and treats all physicians the same regardless of their individual performance. Congress has protected physicians from negative updates resulting from expenditures exceeding the SGR targets

by the Medicare Modernization Act (MMA, 2003), which provided 1.5 percent updates for 2004 and 2005, and the Deficit Reduction Act (S. 1932), which provided 0 percent updates in 2006. Without modification, Medicare Part B expenditures will continue to exceed the SGR. Furthermore, the FFS payment system does not reward high quality care and often pays physicians more for treating complications that arise from poor quality of care.

Pay-for-Performance as a Means of Addressing These Problems

To close the gap between the care that is recommended and the care that patients receive, the Institute of Medicine (IOM) recommended reforms to the health system, one of which is the reform of current payment policies to create stronger incentives for providing high quality, efficient health care services (IOM, 2001). In response, a number of system reform experiments have been carried out in both the public and the private sector that offer financial and sometimes non-financial incentives to providers with the explicit goal of stimulating improvements in health care quality, provider accountability, and efficiency (Rosenthal et al., 2004; Epstein et al., 2004).

In 2005, the Medicare Payment Advisory Commission (MedPAC), which advises the U.S. Congress on issues related to Medicare, recommended that P4P be implemented for hospitals, home health agencies, and physicians (MedPAC, 2005). Congress has also shown interest in P4P, as evidenced through the multiple bills it has put forth. For example, the Medicare Value Purchasing Act (S.1356) proposed that a portion of Medicare reimbursement for physicians, hospitals, health plans, end-stage renal disease providers, and home health agencies be tied initially to the reporting of performance measures (either 2006 or 2007, depending on provider type) and then to actual performance (ranging from 2007 to 2009).

The Deficit Reduction Act of 2005(passed on February 8, 2006) does not include provisions for physician level P4P. It does, however, require MedPAC to submit a report by March 1, 2007, on mechanisms that could be used to replace the SGR system. Furthermore, the Deficit Reduction Act calls for hospital P4P to be implemented in fiscal year 2009, thereby setting the stage for future legislative activity to embed P4P in the reimbursement formula for Medicare physician services.

Important groundwork is being laid through a variety of CMS demonstrations. One of these, the Medicare Physician Group Practice (PGP) demonstration, is providing financial incentives to 10 physician group practices based on their quality and cost-efficiency performance. In addition, the Physician Voluntary Reporting Program (PVRP),

started in January 2006, will provide internal comparative performance feedback to providers but will not involve public reporting. CMS has started signaling its anticipated policy direction in public forums—first, by engaging in voluntary reporting of performance by physicians; then by moving to financially incentivize reporting; and then by implementing P4P (McClellan, 2005a; McClellan 2006). Finally, CMS is collaborating with the Ambulatory Care Quality Alliance in conducting a series of pilot projects around the country to test the feasibility of aggregating data across multiple payers and then scoring physicians and/or physician practices on a range of performance measures.

WHAT IS THE EMPIRICAL EVIDENCE FOR THE EFFECTIVENESS OF PAY-FOR-PERFORMANCE?

Neither the peer-reviewed literature on P4P programs nor ongoing evaluations of such programs provide a reliable basis for anticipating the effects of P4P in Medicare. Our examination of the peer-reviewed literature on P4P yielded 15 published studies whose goal was to determine the effect of directing financial incentives for health care quality at physicians, physician groups, and/or physician practice sites. All of these studies evaluated experiments that occurred in the late 1990s or in the early 2000s.

These studies do not, separately or in total, provide a clear picture of how P4P affects performance. The following is a breakdown of our findings:

- **The seven most rigorously designed studies (i.e., those using randomized controlled trials) provide an ambiguous message**: four show **mixed results** (Fairbrother et al., 1999, Fairbrother et al., 2001; Kouides et al., 1998; Roski et al., 2003), and three report **no effect** (Grady et al., 1997; Hillman et al., 1998; Hillman et al., 1999).
- **The two quasi-experimental studies report mixed findings** (Rosenthal et al. 2005; Levin-Scherz et al., 2006).
- **The least rigorously designed studies tend to report positive results** for at least one aspect of the programs examined (Francis et al., 2006; Greene et al., 2004; Amundson et al., 2003; Armour et al., 2004; Fairbrother et al., 1997; Morrow et al., 1995).

Drawing conclusions from the published literature about how P4P affects health care quality is problematic for a number of reasons:

- The interventions evaluated were small, and most were of very short duration, thus limiting the likelihood that an impact would be observed.

- The interventions typically occurred in one location with selected characteristics (e.g., targeting Medicaid providers), thus limiting the ability to generalize from the studies' findings.
- Many of the studies lacked control groups, thus making it difficult to distinguish the effects of P4P from the effects of other factors in the environment (e.g., medical group quality improvement interventions, public reporting of performance scores).
- The studies provide no information about the various design features that may have played a role in an intervention's success or failure, such as level of engagement and communication with providers and what share of a physician's practice the intervention represented (i.e., the dose effect).

In addition to the studies' methodological limitations, most of the programs evaluated in these studies do not resemble the P4P programs operating today in terms of size (i.e., number of measures or number of providers), duration, and magnitude of rewards. Thus, it is impossible to generalize from the findings in the published literature in order to estimate the effects of the newer generation of P4P programs. Some of these newer programs are being evaluated, but the results are just starting to emerge, and much of the new literature speaks only to lessons learned about the implementation process.

Furthermore, these new P4P programs are real-world experiments and, as such, suffer from some of the same methodological problems (i.e., lack of control groups, lack of random assignment to and not to incentives) as the studies evaluated in the peer-reviewed literature. These shortcomings will, of course, limit what the evaluations can reveal about how P4P affects performance. They do not, however, mean that these programs can offer no useful lessons for CMS.

WHAT CAN BE LEARNED FROM REAL-WORLD PAY-FOR-PERFORMANCE PROGRAMS?

As of December 2005, approximately 157 P4P programs were operating across the United States. These programs were sponsored by 130 organizations—including individual health plans, coalitions of health plans, employer groups, Medicare, and Medicaid—and they covered over 50 million health plan enrollees (Med-Vantage, 2005).

Because there is no published literature describing lessons learned in these programs, RAND held discussions with 20 private-sector P4P programs that target individual physicians or groups of physicians, as well as with six of the 10 medical groups participating in the CMS Physician Group Practice (PGP) P4P demonstration. The

PGP P4P demonstration is a three-year program to implement a P4P program for group practices with at least 200 doctors who care for FFS Medicare beneficiaries. The program's goal is to improve care for beneficiaries with chronic medical conditions by rewarding physician groups that manage patients across the continuum of care in a cost-effective, high quality manner. These discussions provided insights that could be useful to CMS as it embarks on a P4P program for Medicare physician services.

Some common themes emerged from our discussions with participants in the private-sector P4P programs:

- **P4P is not a panacea.** It is not, by itself, a solution for poor quality and rising costs. P4P needs to be implemented as part of a multi-dimensional set of strategies designed to change physician behavior so as to achieve quality and cost goals.
- **Physician involvement and engagement are critical to successful program implementation**. Sponsors of these programs found communicating with physicians to be a challenge, particularly in markets lacking sizable group practices or strong local physician leadership or organization. Traditional methods of communication, such as newsletters and mailings, were insufficient for raising awareness about the program and engaging physicians in quality improvement activities.
- **Health care remains local, and a one-size-fits-all approach may not work**. Discussions with P4P program participants revealed no consensus about the best way to design a P4P program (this lack of agreement is reflected in the design variations across existing programs). Variations occurred as a function of differences in the goals of individual P4P sponsors, the type of insurance product, and how physicians were organized within a geographic market.
- **It is essential to pilot test the implementation of measures and other implementation processes (e.g., audit, feedback reporting) at each step of the program**. Participants in programs that had not conducted pilot tests indicated that the omission was a serious mistake and strongly advised that all aspects of program design and implementation be tested. Two items repeatedly mentioned as being necessary were a willingness to be flexible and change, and the recognition that program development will involve some trial and error.
- **The accuracy and reliability of data underlying the measures must be ensured, and there must be a fair and equitable process for appeals.** These

two items were mentioned repeatedly as being essential for addressing providers' concerns about their performance scores and data accuracy.

- **Ongoing evaluation is needed.** Monitoring is needed to track programmatic effects and the process of implementation and to provide information that can be used to adjust program design and implementation.
- **Programs should start small, and success should be demonstrated.** These are seen as ways to build trust among program stakeholders.
- **Substantial infrastructure is required to support program operations.** Core operational functions needed are data warehousing; data aggregation; programming and analysis; data auditing; appeals management and data correction; performance feedback, such as report cards; communication with, engagement of, and support of physicians; measures maintainance; and payout computation and distribution. To support these operations, additional infrastructure investments in the form of both people and information technology will be needed, and sufficient resources must be allocated to support program operations.
- **Alignment of programs with the measures being used and physician requirements is vital.** This type of coordination is essential for reducing both confusion and the burden placed on providers who contract with multiple payers. Without it, providers may have to cope with inconsistent program requirements and measure specifications. Alignment among P4P sponsors within a market also strengthens the behavior-change signal to providers, increasing the likelihood that providers will in fact change their behavior.
- **Physicians need support for successful program participation**. Examples of the types of support being provided by programs are patient registries, technical support, and education.
- **To motivate physicians to change behavior, performance information must be actionable**. Providing rates is not sufficient; physicians must be able to act upon the information provided. For example, since most physicians in P4P programs continue to operate in an environment of paper records, they could be provided with specific lists of patients who need recommended care rather than being expected to start accessing population-based data and using it to track the provision of services and/or identify specific clinical areas in which there are less costly treatments that yield the same clinical outcomes as costly treatments for most patients (e.g., the use of ACE inhibitors instead of ARBs to treat

hypertension). Such lists would provide information in a manner that facilitates behavior change.

Early lessons that have emerged from the first year of the CMS PGP demonstration include the following:

- **Participation in the PGP demonstration was a key driver of performance improvement in the physician group organizations**. Four of the six physician groups noted that participating in the demonstration enabled them to implement changes (particularly to information systems) that had been discussed internally for years but never put into place. Once the demonstration was in place, changes began to happen or happened much more quickly than before.
- **Capital investments are required to support measurement and quality improvement work.** The physician groups told us they believe that the infrastructure investments needed to support P4P management and measurement will be "enormous." An influx of capital of this size suggests that vendors likely to fill infrastructure needs will have to be closely scrutinized.
- **Participation in P4P can prompt improved sharing of ideas to promote better care for the population.** Several physician groups mentioned their surprise that providers participating in the demonstration have embraced the concepts of population management underlying their case management strategies. The demonstration has improved the physicians' sharing of ideas for promoting better care within the organizations.
- **The support provided to PGPs was a critical feature of program design.** All physician groups mentioned at least one instance in which they contacted either CMS or its support services contractor, Research Triangle Institute (RTI), to comment on how a measure should be specified, to appeal the inclusion of beneficiaries in the group's target population under the attribution algorithm, to question the inclusion of patients in the denominator of measures, or to request help managing data.

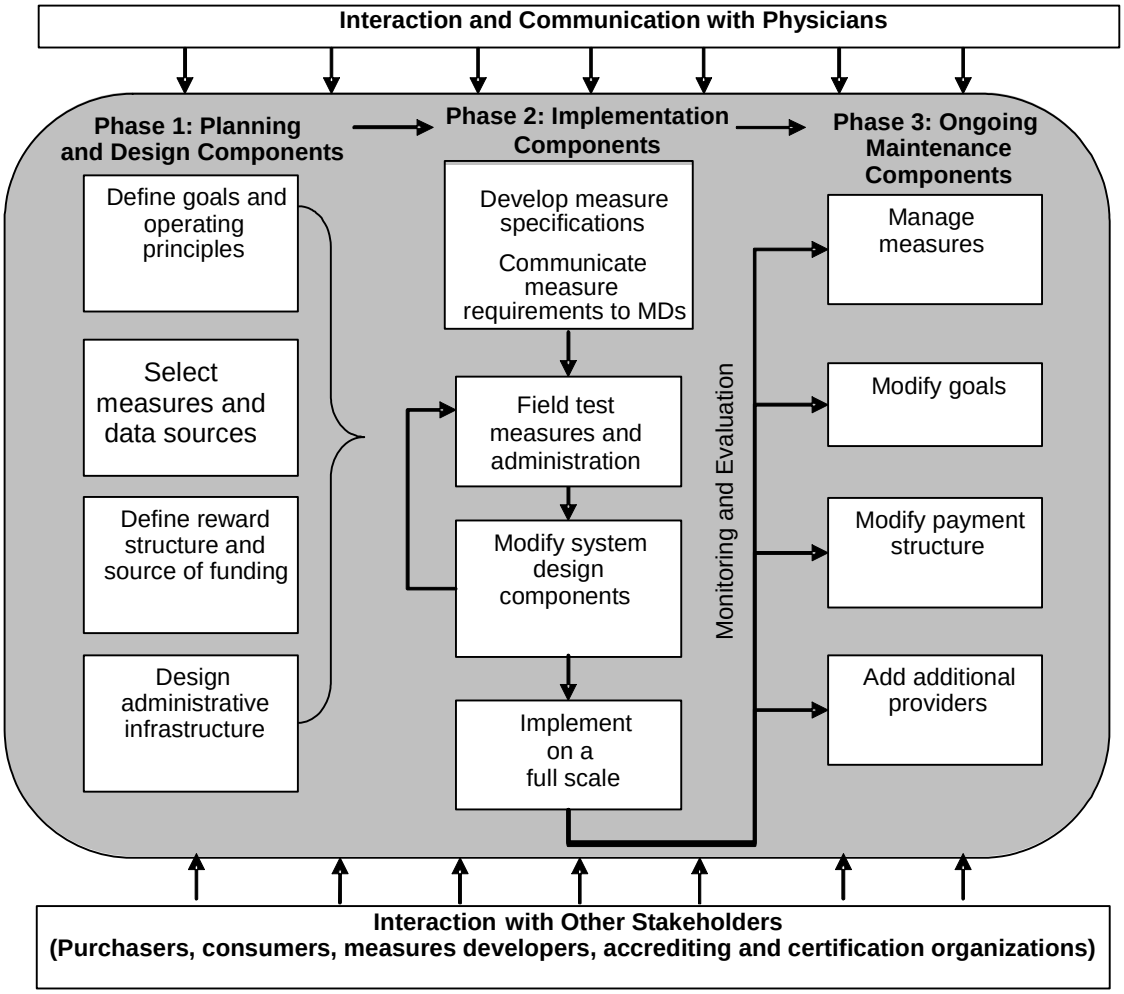
The organizations we spoke with, both in the private sector and in the CMS demonstration program, are firmly committed to P4P and believe that their programs, often in conjunction with other quality improvement activities, are resulting in care that is both of a better quality and more efficient. However, few of these programs are being evaluated in a rigorous manner, and only a handful have attempted to compute the return on investment (ROI) from their efforts.

WHAT DESIGN ISSUES AND OPTIONS NEED TO BE CONSIDERED?

Currently, there is no single strategy for designing and implementing a P4P program, so a great deal of experimentation and refinement is occurring as programs learn lessons along the implementation path. While all programs have key design components—such as attribution rules, payout structures, and measures selection—very little is known about the best form for these components or the relative importance of different components for achieving the program's goals. In some cases, newer programs are adopting the design components of more-mature programs, but there is substantial variation across P4P programs in terms of their approach to designing their programs. Programs are generally customized to address specific characteristics of the local health care market (e.g., organization of physicians, existence of physician leaders in the community), and little attention is paid to what theory suggests might be the best options of various program components to adopt. At this stage, absent empirical evidence to support one design approach over another, the variation in P4P experiments will allow opportunities for testing various design strategies.

Development of a P4P program is a complex undertaking involving many moving, interrelated parts. In addition, P4P programs are not static in terms of design. As a guide for those planning P4P programs, we used the steps identified in our discussions with P4P sponsors to construct a framework of core steps associated with developing and operating a P4P program. This framework, shown in Figure 1, displays the array of decisions that have to be made by any P4P program developer; it also highlights the interactions among various steps in the process.

Figure 1
A Framework to Guide Development of a Pay-for-Performance Program



Choosing among the various options typically reflects considerations of whether the approach helps to achieve programmatic objectives and what consequences may occur as a result. For example, if an explicit goal is to accelerate implementation of information technology (IT), the program developer may elect to include measures on the provider's IT capabilities. However, as was underscored in our discussions with P4P program developers, P4P program development is largely experimental in many respects, and the impact of various design components has not been studied and is not well understood.

Our review of the literature and discussions with a broad cross-section of existing P4P programs in the private sector revealed a host of options for the design components

that need to be addressed when developing a P4P program. The design issues that are of specific interest to ASPE and that we thus assessed are

- How should the initial performance areas that are to be subject to P4P be identified?
- What role should physicians and other stakeholders play in developing a P4P program for Medicare?
- What are appropriate measures for a P4P program?
- What unit of accountability should CMS measure and reward?
- Given the geographic variation in practice of care, should CMS pursue a national or a regional approach to implementation?
- How should patients be matched to individual physicians or group practices to ensure accuracy of measurement?
- How should rewards be structured?
- What should CMS be considering with regard to program infrastructure, including measure selection and specification, pilot testing, data collection and management, support to physicians, reporting and feedback, and monitoring?

We present information helpful in understanding the consequences or challenges associated with choosing particular design options. However, the effects of choosing one option over another are in many cases not known.

IS IT POSSIBLE TO IMPLEMENT A PAY-FOR-PERFORMANCE PROGRAM FOR MEDICARE PHYSICIAN SERVICES?

A P4P program for Medicare physician services can be implemented. However, in designing and implementing a P4P program, CMS will face significant challenges that include

- **The absence of an existing organizational infrastructure within CMS with which to manage the myriad components associated with running a P4P program.** This is particularly the case given a program of the size and scope necessary to measure and reward all or most physicians in the Medicare Part B program. To support a P4P program's operations, many systems will have to be designed, built, tested, and maintained, an endeavor that will require dedicated and sustained resources.
- **The absence of a P4P program comparable in size and scope to a P4P program for Medicare physician services**. There is no P4P program of comparable size from which to draw lessons.

- **The absence of infrastructure (personnel and information systems) at the individual doctor-level to support a P4P program's requirements**. For example, the majority of physician offices have neither electronic health records nor sufficient staff to perform the chart abstractions that might be required to provide information needed to construct the performance measure.
- **The difficulty of communicating with and engaging individual physicians in the program to achieve the desired behavior changes.** Organized medical groups have the staff and structure to facilitate communication between a P4P program sponsor and front-line physicians. At the level of the individual physician, however, there is no "local physician leadership" or point person who can help to facilitate communication with physicians about the program, engagement of physicians in the program, and assistance with behavior change.
- **The shortness of the timetable for ramping up a national operation.** Given Congress's mounting pressure for action, CMS is unlikely to have time to pilot a P4P program in multiple sites. There likely will be pressure to roll out a national program in a short period of time.
- **Physician resistance to transparency (public reporting) of performance data.** Some people have asserted that public transparency and accountability are valuable additions to P4P programs because they drive behavior change among physicians. However, physicians have expressed concerns about public reporting of performance results, especially about problems with data inaccuracies and failure to account for differences in patient populations served.

TAKING THE FIRST STEPS TO IMPLEMENT A P4P PROGRAM FOR MEDICARE PHYSICIAN SERVICES

There are several steps that CMS could undertake immediately and in the near term, as well as in the longer term, to prepare itself for designing and implementing a P4P program for Medicare physician services. The actions presented here, if taken, would provide information to guide program planning, would help generate awareness and engagement among physicians, and would begin to build the program infrastructure needed to support a P4P program.

Near-Term Steps (6 to 18 months)

Model critical design components using existing data.

CMS could start laying the groundwork for structuring a P4P program by modeling various program design components using existing Medicare claims data. Some of the critical design issues to be addressed in modeling the components are (1) the implications of different attribution rules, (2) the number of measures that can be scored today using claims data, (3) the number of physicians whose performance can be reliably scored using measures based on administrative data, and (4) the increase in the number of physicians that CMS could score if scores were based on composite measures versus individual indicators of performance.

Monitor the experiences of the Physician Voluntary Reporting Program and consider how to address emerging lessons in the design of P4P for Medicare physician services.

Implementation of the PVRP, a program started in January 2006 that will provide internal comparative performance feedback to providers on a starter set of 16 measures, offers CMS a potential foundation on which to build a P4P program. The lessons being learned in the PVRP will provide CMS with valuable information; in particular, the monitoring of physician participation and growth in participation over time will provide indications about the readiness of physicians nationally to provide information on the selected measures. Interviews with physicians could give CMS valuable insights about why physicians did or did not agree to participate. Participating providers could describe the challenges they experience with the data collection and reporting process, as well as their reactions to performance feedback reports. Non-participating providers could help to identify barriers to participation and actions needed to address them. Information gained from physician interviews could be useful in determining how to modify the program going forward as a stepping stone to full P4P. The interviews also would allow CMS to build communication channels with physicians before a P4P program is implemented and would constitute an important step in soliciting physician input on program design.

Mid-Term Steps (18 to 36 months)

Create incentives for participation in the PVRP as a way to help physicians move toward understanding performance measurement, to build systems to support measurement, and to work toward performance transparency.

Low participation in PVRP may suggest the need to provide inducements for participation, such as pay-for-reporting. Participation in PVRP is important in that it offers physicians the opportunity to gain experience with submitting data and receiving performance feedback, well in advance of P4P. PVRP participation also allows physicians time to see performance scores in a confidential manner, giving them the opportunity to improve systems for data capture and to identify and correct quality problems in advance of public reporting. This is an important step for CMS to take on the path to public transparency.

Expand the PVRP measurement set and administrative collection of measures.

CMS could also continue to expand the PVRP 16-measure set so that it is consistent with P4P program design decisions about which measures to reward to drive improvements. In addition, to support the administrative reporting of data to produce performance measures, particular attention should be paid to modifying the HCFA 1500, the form physicians use to submit claims to Medicare, to capture administratively the data elements needed to support performance measurement (e.g., working with the AMA to develop Current Procedural Terminology [CPT] supplemental codes).

Plan for program evaluation and collect baseline data.

It is also very important for CMS to build into its P4P design the continuous evaluation of program implementation and effects. Ongoing evaluation will give CMS critical information that it can use to adjust the program. Assessment of program effects will require that CMS collect baseline information about performance. If CMS expects to compute the return on investment, it will need to track program costs.

Longer-Term step (36 months and beyond)

Scale up incrementally and continue to build infrastructure capacity.

As the PVRP matures, CMS could scale up the program incrementally by adding measures and physician specialties and continuing to build infrastructure to accommodate

the program’s increasing size. By building gradually on successes, CMS will help to build trust within the provider community and will gain experience along the way.

ACKNOWLEDGMENTS

We gratefully acknowledge the sponsors of the pay-for-performance programs, who willingly made the time to participate in individual discussions with us in which they offered valuable information and insights about their experiences in designing and implementing pay-for-performance programs. We thank the members of our advisory panel—Steve Asch, Melinda Beeuwkes Buntin, Howard Beckman, Elizabeth McGlynn, Barbra Rabson, Meredith Rosenthal, and Barbara Wynn—for carefully reviewing the discussion guidelines to help ensure that pertinent topics and issues were addressed, for providing input into the applicability of practical elements analysis, and for reviewing and commenting on this report in its final stages. In addition, we appreciate the assistance that Geoff Baker and Beau Carter of Med-Vantage provided; they helped us identify and narrow the list of candidate pay-for-performance programs for our discussions. Finally, we thank Susan Bogasky, Assistant Secretary for Planning and Evaluation Project Officer, for her time spent reviewing our work and for her guidance on the project.

GLOSSARY

Abbreviation	Definition
AAFP	American Academy of Family Physicians
ACG	Adjusted Clinical Group
ACOVE	Assessing Care of Vulnerable Elders
AHIP	Association of Health Insurance Plans
AHRQ	Agency for Healthcare Research & Quality
AMA	American Medical Association
AMGA	American Medical Group Association
ANOVA	Analysis of Variance
AQA	Ambulatory Care Quality Alliance
ASO	Administrative Services Only
ASPE	Assistant Secretary for Planning and Evaluation
BCBSA	Blue Cross/Blue Shield Association
BCC	Blue Cross of California
CABG	Coronary Artery Bypass Graft
CAD	Coronary Artery Disease
CAHPS	Consumer Assessment of Healthcare Providers and Systems
CAPG	California Association of Physician Groups
CHCF	California Health Care Foundation
CHF	Congestive Heart Failure
CMS	Centers for Medicare & Medicaid Services
CPT	Current Procedural Terminology
CRC	Colorectal Cancer
DCG	Diagnostic Cost Group
DPRP	Diabetes Physician Recognition Program
DRG	Diagnosis Related Group
E&M	Evaluation and Management
EHRS	European Histamine Research Society
ESRD	End Stage Renal Disease
ETG	Episode Treatment Group
FFS	Fee-for-Service
GP	General Practitioner
HCFA	Health Care Financing Administration
HCPCS	Healthcare Common Procedure Coding System
HEDIS	Health Plan Employer Data and Information Set
HF	Heart Failure
HMO	Health Maintenance Organization
HQA	Hospital Quality Alliance
HQI	Hospital Quality Initiative
ICSI	Institute for Clinical Systems Integration

IFMC	Iowa Foundation for Medical Care
IHA	Integrated Healthcare Alliance
IOM	Institute of Medicine
IPA	Independent Practice Association
IT	Information Technology
JCAHO	Joint Commission on the Accreditation of Healthcare Organizations
KDOQI	Kidney Disease Outcomes Quality Initiative
LDL	Low Density Lipoprotein
LVF	Left Ventricular Failure
MedPAC	Medicare Payment Advisory Commission
MEI	Medicare Economic Index
MHQP	Massachusetts Health Quality Partners
MMA	Medicare Modernization Act
MMR	Measles, Mumps, and Rubella
NCQA	National Committee for Quality Assurance
NKF	National Kidney Foundation
NMS	Net Medicare Savings
NQF	National Quality Forum
NVHRI	National Voluntary Hospital Reporting Initiative
OTC	Over-the-Counter
P4P	Pay-for-Performance
PCHI	Partners Community Healthcare, Inc.
PCPI	Physician Consortium for Performance Improvement
PGP	Physician Group Practice
PMPM	Per Member Per Month
POL	Physician Office Link
POS	Point-of-Service
PPO	Preferred Provider Organization
PVRP	Physician Voluntary Reporting Program
PY	Program Year
QA Tools	Quality Assessment Tools (RAND)
QI	Quality Improvement
QIO	Quality Improvement Organization
RBRVS	Resource-Based Relative Value Fee Schedule
ROI	Return on Investment
RTI	Research Triangle Institute
RWJF	Robert Wood Johnson Foundation
SCIP	Surgical Care Improvement Project
SGR	Sustainable Growth Rate
UK	United Kingdom
VHA	Veteran’s Health Administration

1. BACKGROUND AND CONTEXT

The practice of paying health care providers based on their quality performance, which is referred to as pay-for-performance (P4P), emerged in the late 1990s as a strategy for driving improvements in the quality and value of health care. The prevalence of P4P programs has increased significantly in the past five years, and many of the early programs have evolved to cover more providers and a broader set of measures (e.g., cost-efficiency, information technology capability). Currently, the Centers for Medicare & Medicaid Services (CMS), the largest purchaser of health care services in the United States, is actively considering P4P for Medicare physician services. CMS is viewing this policy strategy as one way in which physician reimbursement can be restructured to more clearly incentivize physicians to provide care to Medicare beneficiaries that is both high quality and cost-effective.

In September 2005, the Assistant Secretary for Planning and Evaluation (ASPE), within the U.S. Department of Health and Human Services, contracted with the RAND Corporation to help in its assessment of whether P4P can be effectively implemented in the Medicare physician service delivery and payment environment. Specifically, RAND was tasked to

1. Review the empirical evidence on the effectiveness of P4P.
2. Identify the characteristics of current P4P programs.
3. Assess aspects of current approaches to determine how well they will transfer to the development of a P4P program for the Medicare physician fee schedule.

In laying out the scope of work for the six-month contract with RAND, ASPE emphasized its particular interest in having the following design issues examined:

- How should the initial areas subject to P4P be identified?
- Depending on consensus definitions of P4P, what are the appropriate measures (outcomes, process, quality, financial, etc.)?
- Does geographic variation in service delivery or practice patterns affect scope (regional or national) of implementation?
- How are patients matched to individual physicians and/or group practices to ensure accuracy of measurement?
- How should accuracy of measurement be validated?
- What performance is being rewarded and how are rewards structured?

- How will loss and profit sharing be incorporated into the approach (e.g., should poor performance be reflected in reduced payments)?
- How will the issues of data collection and information infrastructure be addressed?

This chapter builds the foundation for subsequent chapters of this report by defining P4P and its dimensions and by providing the policy context underlying the rationale for investing in P4P as a strategy to create system change. We conclude our background discussion with a summary of principles and recommendations that have been offered by a range of national organizations seeking to influence the design of P4P programs in both the public and the private sector.

DEFINING PAY-FOR-PERFORMANCE

P4P is defined as the differential payment of health care providers based on their performance on a set of specified measures. The term *provider*, which we use throughout this report, encompasses individual physicians, physician practices, medical groups and integrated delivery systems, and hospitals. P4P programs seek to align payments to providers with a program sponsor's goals, such as provision of high quality care, improved cost-efficiency, and delivery of patient-centered care. For example, if a program sponsor is seeking to improve quality of care, the program will include clinical measures, such as the provision of screenings and immunizations or the provision of disease-specific services. If that program sponsor seeks to improve the cost-efficiency of care as well, the program may also include the use of generic medications or the risk-adjusted costs of treating patients with a particular condition. P4P programs are intended to financially reward providers who already perform in line with the program sponsor's identified goals or who modify their behavior and re-engineer their practice to achieve those goals.

Three other policy levers are also designed to change physician and/or consumer behavior through the use of financial and/or non-financial incentives. We specifically excluded these three mechanisms from our examination of P4P programs. The first of these, **provider profiling (or report cards)**, is an internal reporting activity through which a health plan or other organization supplies providers with comparative performance information on a set of measures. The second mechanism, **public reporting,** makes provider performance information available to external stakeholders, such as purchasers and patients/consumers, in an effort to hold providers accountable for their performance and to inform patients' choice of providers. The third mechanism, **the tiered**

provider network, sorts hospitals, physician groups, or physicians into differential categories on the basis of costs and/or quality and provides consumers with financial incentives in the form of lower out-of-pocket costs (i.e., lower co-payments or deductibles) to use providers in the high performing tier.

THE POLICY CONTEXT FOR PAY-FOR-PERFORMANCE

Impetus

A variety of studies document substantial deficiencies in the quality of care delivered in the United States (Asch et al., 2006; Institute of Medicine [IOM], 2001; Schuster et al., 1998; Wenger et al., 2003). In a national examination of the quality of care delivered to adult patients, McGlynn and colleagues found that patients received on average only about 55 percent of recommended care and that adherence to clinically recommended care varied widely across medical conditions (McGlynn et al., 2003). Wenger and colleagues found similar results for vulnerable community dwelling elders; they also found that performance was worse for geriatric conditions (Wenger et al., 2003).

Health care costs continue to rise at a steady pace and are anticipated to account for 18.7 percent of gross domestic product by 2014 (Heffler et al., 2005). In 2006, the federal government will spend $600 billion for Medicare and Medicaid, covering approximately 87 million beneficiaries; by 2030, expenditures for these two programs are expected to consume 50 percent of the federal budget, putting funding for other discretionary programs in jeopardy (McClellan, 2006). CMS Administrator Dr. Mark McClellan has stated publicly that to be able to continue funding these programs, CMS will have to redesign existing policies and practices.

Mechanisms for paying providers—e.g., fee-for-service (FFS), capitation, and salary—have been shown to influence the behavior of physicians, such as their use of expensive tests and procedures. However, none of these payment mechanisms encourages providers to deliver high quality and cost-effective care. Reviews of the evidence on the relationship between payment method and quality fail to show a clear and consistent pattern of one type of payment resulting in better quality of care (Dudley et al., 1998; Miller and Luft, 1994). To close the quality gap, the Institute of Medicine (IOM) recommended reforms to the health system, including the reform of current payment policies to create stronger incentives for quality and the efficient delivery of health care services (IOM, 2001).

Public and Private System Payment Reforms

In response to the IOM's call for payment policy reform, a number of experiments in health system reform have been developed in both the public and the private sector. These reforms offer providers financial and, in some cases, non-financial incentives explicitly aimed at stimulating improvements in health care quality, provider accountability, and efficiency (Rosenthal et al., 2004; Epstein et al., 2004). As P4P gains traction at the hospital, health system, and medical group levels, associated changes are occurring in performance-based compensation at the individual physician level. For example, early evaluation findings from the California Integrated Healthcare Alliance (IHA) P4P program show that physician groups have responded at the group level by restructuring their physician compensation formulas to include performance-based measures, with between 5 percent and 10 percent of a physician's salary at risk for quality performance (Damberg and Raube, 2006).

In September 2004, the IOM initiated a project called Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement. This project's committee is tasked with recommending options for redesigning Medicare provider payment policies and performance improvement programs to encourage and reward improvements in the delivery of health care (IOM, 2006). The first report from this project focuses on provider performance measurement (IOM, 2005); two additional reports, on payment incentives and quality improvement initiatives, are expected before the end of 2006.

The Medicare Payment Advisory Commission (MedPAC), which advises the U.S. Congress on issues related to Medicare, reports to Congress on Medicare payment policy annually. In 2004, MedPAC recommended P4P for Medicare Advantage plans and dialysis providers. In 2005, this recommendation was broadened to include hospitals, home health agencies, and physicians (MedPAC, 2005).

Existing Payment Policy for Medicare Part B Physician Services

Approximately 484,000 physicians regularly bill for providing Medicare Part B services (MedPAC, 2006). Medicare's FFS payments for physician services follow a resource-based relative value fee schedule (RBRVS). The annual update to the fee schedule is determined by three factors: (1) the rate of change in the Medicare Economic Index (MEI), (2) a price index measuring changes in the costs of maintaining a physician practice, and (3) a sustainable growth rate (SGR) expenditure target. The annual update

factor to the physician fee schedule is adjusted based on a comparison of cumulative past actual expenditures with the SGR.

Although the SGR was established as an expenditure control mechanism, the SGR targets are routinely exceeded because the SGR is applied at the national level and treats all physicians the same regardless of individual performance. Congress has used the Medicare Modernization Act (MMA, 2003) to protect physicians from the negative updates that would have resulted from the SGR targets being exceeded. The act provided 1.5 percent updates for 2004 and 2005, and the Deficit Reduction Act (S. 1932) provided 0 percent updates in 2006. Without modification, Medicare Part B expenditures will continue to exceed the SGR. It is questionable whether across-the-board decreases in the fee schedule would control costs, since physicians might increase the volume of services provided to maintain gross revenues.

Furthermore, the FFS payment system ignores quality and efficiency of care. In fact, payment is structured to pay physicians more for treating complications that arise from poor quality of care.

The Growing Policy Problem and Movement Toward Pay-for-Performance for Medicare

The sustained growth in volume—in part a by-product of the current, FFS payment system—has combined with evidence of substantial unnecessary variation in practice patterns and deficiencies in quality of care to generate within CMS an interest in designing financial incentives that encourage increased quality and efficiency by putting a portion of a physician's Medicare payments at risk for performance (MedPAC, 2005; McClellan, 2005). Important groundwork is being laid through a variety of CMS demonstrations; for example, the Medicare Physician Group Practice (PGP) demonstration, CMS's collaboration with the American Medical Association (AMA) and physician specialty organizations on measure development, and the Physician Voluntary Reporting Program (PVRP). Begun in January 2006, the PVRP will provide internal comparative performance feedback to providers and will not involve public reporting. CMS has started signaling in public forums its anticipated policy direction, beginning by engaging physicians in voluntary reporting of performance, then moving to financially incentivized reporting, and finally implementing P4P (McClellan, 2005; Straube, 2005).

The Medicare Value Purchasing Act (S. 1356), introduced in Congress on June 30, 2005, proposed that a portion of Medicare reimbursement for physicians, hospitals, health plans, end-stage renal disease providers, and home health agencies be tied first to

reporting on performance measures (either 2006 or 2007, depending on provider type) and then to actual performance (ranging from 2007 to 2009). The bill eventually passed, although provisions for physician-level P4P were removed from the final version of the legislation (Pear, 2006; Endocrine Insider, 2006). This Act set the stage for future legislative activity to embed P4P in the reimbursement formula for Medicare physician services.

Principles for Pay-for-Performance Programs

In response to increased interest in and growth of P4P programs, a number of organizations have put forth design principles for P4P programs in the hopes of influencing how CMS and other P4P sponsors decide to structure their P4P programs (see Appendix A). Among these organizations are MedPAC, the Joint Commission on the Accreditation of Healthcare Organizations (JCAHO), employer coalitions, the AMA, the American Academy of Family Physicians (AAFP), and the American College of Physicians.

The principles are varied, and at times the recommendations made by the different organizations directly oppose one another. The major areas of disagreement about P4P design issues are as follows:

- Should P4P programs, especially in Medicare, be budget neutral?
- Should P4P programs include negative financial incentives for participating providers?
- Should P4P programs include efficiency measures?
- Should the measures used be stable or change over time?

Should P4P programs include public reporting? Organizations also vary in what they explicitly include in their statements. For example, physician organizations frequently include these principles: voluntary participation, no link between rewards and the ranking of physicians relative to one another, reimbursement of physicians for administrative burden of collecting and reporting data, and physician involvement in program design.

There are, however, areas of consensus among the organizations. The following principles/recommendations are endorsed by nine or more organizations:

- P4P programs should use accepted, evidence-based measures.
- Risk-adjustment methods should be used to prevent deselection or avoidance of patients who are more difficult to treat (i.e., sicker or non-compliant).
- Incentives should be aligned with the practice of high quality, safe health care.

- Physicians should be provided with positive incentives for adopting and using information technology.

ABOUT THIS REPORT

The remainder of this report presents the findings of RAND's assessment of P4P options for Medicare physician services. Chapter 2, provides a review of the published empirical literature on the impact of P4P that targets physicians. Chapter 3 contains information derived from our discussions with private-sector P4P program sponsors nationally, and with group practices participating in the CMS PGP demonstration. The emphasis in these discussions was on defining the design components of currently operating P4P programs across various settings, developing a framework to guide P4P program development and implementation, and identifying key lessons learned.

Chapter 4 discusses key P4P design components and options, drawing on the experiences of current P4P program sponsors and participants; it also assesses the applicability of various options for design components to the development of a P4P program for the Medicare physician fee schedule. Chapter 5 concludes by summarizing key P4P lessons, identifying challenges that CMS will need to address should it decide to develop a Medicare physician services P4P program, and outlining a set of actions CMS could take to prepare for P4P.

2. A REVIEW OF THE EVIDENCE ON PAY-FOR-PERFORMANCE

This chapter provides a review of the empirical literature on the effect of P4P targeted at physicians on health care quality. Currently (April 2006), there are few peer-reviewed published studies on the effect of financial incentives for quality, and the empirical studies largely address small-scale, financial-incentive demonstrations of limited duration. As a result, it is difficult to use the findings of most of these studies to generalize to the current generation of P4P programs, which differ from the earlier ones not only in that they are sustained efforts, but also in that they are of a substantially larger scope in terms of number of performance measures being rewarded, number of providers exposed to them, and dollars at risk.

METHODS

We limited our review of the published literature to studies that examine the effect of P4P programs specifically targeting individual physicians, medical groups, and/or physician practice sites, our goal being to inform policy discussions on applying this work to Medicare physician services. We excluded studies that targeted hospitals or other institutions.

We searched for articles published between January 1995 and April 2006 in the Medline/PubMed, ABInform, PsycInfo, and CINAHL databases; we used various combinations of search terms—i.e., P4P, pay for quality improvement, quality improvement, financial incentive(s), monetary incentive(s), reimbursement, bonus, reward, provider payment, performance improvement, and quality initiative. This search generated 1,066 peer-reviewed articles published in English, most of which focused on quality improvement without inclusion of a financial incentive.

Once we had identified publications that fit our search goals, we examined their citation lists for additional, relevant publications. We also consulted experts in the field of P4P, and reviewed recent summaries published on this topic to ensure that our scan was comprehensive (Dudley et al., 2004; Rosenthal and Frank, 2006). We retained only articles that reported empirical findings related to the effect of paying for quality, specifically excluding articles that focused on incentives to increase productivity or on the relationship between different payment structures, such as fee-for-service (FFS) and capitation and quality.

RESULTS OF PAY-FOR-PERFORMANCE STUDIES

We identified 15 published studies that examined the effects of directing financial incentives for quality at physicians, physician groups, and/or physician practice sites. As Table 1 shows, the design features of the P4P programs included in the studies that we evaluated differed. For example, some programs tested the effect of offering cash bonuses as the reward structure (Grady et al., 1997), others tested the effects of enhanced fee schedules (Kouides et al., 1998), and one tested both approaches for constructing the reward (Fairbrother et al., 1999). The studies also varied in terms of the level receiving the incentive (e.g., individual physician versus medical group). Half of the studies (n=7) focused on programs that targeted incentives at individual physicians (Fairbrother et al., 1999; Fairbrother et al., 2001; Grady et al., 1997; Francis et al., 2006; Greene et al., 2004; Armour et al., 2004; Fairbrother et al., 1997); five focused on incentives to practice sites/medical clinics (Hillman et al., 1998; Hillman et al., 1999; Kouides et al., 1998; Roski et al., 2003; Morrow et al., 1995); two focused on incentives to medical groups (Rosenthal, 2005; Amundson et al., 2003), and one focused on providing incentives to an integrated delivery system physician network (Levin-Scherz et al., 2006).

Table 1: Summary of the Design Features of the Empirical Studies

Study	Focus		Types of Measures			Type of performance target			Form of financial incentive			Incentive Target			Study Design			Results
	Single Clinical Area	Type pf Care	Process	Efficiency	Patient Satisfaction / Experience	Absolute	Relative	Service delivery (no target)	Bonus	Withhold	Enhanced fee schedule	Primary Care Provider only	PCP + Specialist	Practice site / Medical group	Randomized	Comparison Group	Pre- Post- Test	Positive, No Effect, Mixed
Amundson, et al. (2003)		P	I/O			X			X				X	X			X	+
Armour, et al. (2004)	X	P	I/O			NR	NR	NR	X			X					X	+
Fairbrother, et al. (1997)		P	I/O					X			X	X					X	+
Fairbrother, et al. (1999)	X	P	I/O			X	X		X		X	X			X	X	X	+/-
Fairbrother, et al. (2001)	X	P	I/O			X	X		X		X	X			X	X	X	+/-
Francis, et al. (2006)	X	T	I/O	I	I		X			X			X				X	+/-
Grady, et al. (1997)	X	P	I/O					X	X			X			X	X	X	-
Greene, et al. (2004)	X	T	I/O	I	I		X			X			X				X	+
Hillman, et al. (1998)		P	I/O				X		X				X	X	X	X	X	-
Hillman, et al. (1999)		P	I/O			X	X		X				X	X	X	X	X	-
Kouides, et al. (1998)	X	P	I/O			X			X			X			X	X	X	+/-
Levin-Scherz, et al. (2006)		P/T	I/O			X			X	X			X	X		X	X	+/-
Morrow, et al. (1995)		P	I/O	I			X		X			X					X	+
Rosenthal, et al. (2005)		P	I/O			X			X				X	X		X	X	+/-
Roski, et al. (2003)	X	P	I/O			X			X				X	X	X	X	X	+/-

N/A = Not applicable O=Study Outcome (dependent variable) P=Preventive Care + = Positive Results +/- = Mixed Results

NR = Not reported I=Part of Incentive Determination T=Treatment of Clinical Condition - = No Effect

Drawing conclusions from the existing published literature about the effects of P4P on health care quality is problematic. This is partly attributable to the weak designs of the studies, which limit the ability to rule out other factors that may have contributed to the observed effects. The ability to generalize from the findings of these studies is limited by the fact that interventions typically occurred in a single location with unique characteristics, so what was observed may not apply to other locations. The variations across these studies in types of performance targets, forms of financial incentives, types and levels of providers targeted, and clinical areas also complicate the ability to discern which factors are contributing to the observed effects and whether the results can be replicated in other settings under different conditions. The empirical studies yield little information on whether and, if so, how the design features of incentive programs (e.g., target of incentive, types of measures selected, difficulty in complying with program, frequency of providing incentive, amount of incentive, incentive program development process) impact the likelihood of achieving a positive result.

The strength of study designs varied across the 15 evaluations:

- Seven of the 15 were randomized controlled trials (Fairbrother et al., 1999; Fairbrother et al., 2001; Grady et al., 1997; Hillman et al., 1998; Hillman et al., 1999; Kouides et al., 1998; Roski et al., 2003).
- Two employed a quasi-experimental case/control design (Rosenthal et al., 2005; Levin-Scherz et al., 2006).
- Six were pre-/post-test studies with no control group (Amundson et al., 2003; Armour et al., 2004; Fairbrother et al., 1997; Francis et al., 2006; Greene et al., 2004; Morrow et al., 1995).

In terms of how quality was affected by the various incentives examined, the findings are mixed, and the types of results varied by study design type (see Tables 1 and 2):

- **The seven most rigorously designed evaluations (i.e., the randomized controlled trials) send an ambiguous message**. Four had **mixed results** (Fairbrother et al., 1999; Fairbrother et al., 2001; Kouides et al., 1998; Roski et al., 2003), and three reported **no impact** (Grady et al., 1997; Hillman et al., 1998, Hillman et al., 1999).
- **The two quasi-experimental studies also reported mixed results** (Rosenthal et al. 2005; Levin-Scherz et al. 2006).
- **All six of the least rigorously designed studies reported positive results** for at least one aspect of the incentive programs they examined (Francis et al., 2006;

Greene et al., 2004; Amundson et al., 2003; Armour et al., 2004; Fairbrother et al., 1997; Morrow et al., 1995).

Table 2 describes each of the 15 studies, including their study designs, program targets, incentives, results, and limitations.

Table 2: Empirical Studies of Pay-for-Performance Programs Directed at Individual Physicians or Groups of Physicians

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Randomized Controlled Trials					
Fairbrother et al., 1999	Randomized controlled trial at physician level 4 study arms: 1) feedback only 2) bonus plus feedback 3) enhanced fee-for-service plus feedback 4) control – no money and less detailed feedback Multistage stratified cluster sampling Repeated cross-sectional independent samples of patients 8-month study period (3 measurements, 4 months apart) Key outcome: immunization rates Pediatric Medicare population	Individual physicians (n=60 total physicians, with 15 in each arm of the study)	2 different incentives tested: 1) Cash bonus ($1,000 for a 20% improvement from baseline, $2,500 for a 40% improvement, and $5,000 for reaching 80% coverage regardless of baseline), vs. 2) Enhanced fee-for-service schedule ($5 for each vaccine administered within 30 days of its coming due, $15 for any visit where more than one vaccine was due and all vaccines were provided). Incentives provided every 4 months along with performance feedback.	The intervention group receiving the cash bonus and feedback showed a 25% increase in up-to-date immunization status for their patients ($p<.01$). No other study groups showed a significant change, relative to the control group.	Small sample size in each study arm Short study period Some improvements observed may be due to better documentation.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Fairbrother et al., 2001	Randomized controlled trial at physician level 3 study arms: 1) bonus plus feedback, 2) enhanced fee-for-service plus feedback, 3) control – no money and less detailed feedback Multistage stratified cluster sampling Repeated cross-sectional independent samples of patients Approx. 16-month study period (4 measurements, 4 months apart) Key outcome: immunization rates Pediatric Medicare population Note: This is a continuation of the 1999 study (second period of observation) with a 12 month break between the 2 years of the study during which the physicians did not know that the incentives would resume. There was	Individual physicians (n=57, with 24 in bonus group, 12 in enhanced fee-for-service, and 21 in control)	2 different incentives tested: 1) Cash bonus ($1,000 for a 30% improvement from baseline, $2,500 for a 45% improvement, $5,000 for reaching 80% coverage regardless of baseline, and $7,500 for reaching 90% coverage regardless of baseline), vs. 2) Enhanced fee-for-service schedule ($5 for each vaccine administered within 30 days of its coming due, $15 for each visit at which all due vaccines were provided). Incentives were provided every 4 months along with performance feedback.	Physicians in the two incentive groups increased their patients' up-to-date immunizations rates significantly, relative to the control group (bonus group: 5.9% increase, $p<.05$, enhanced fee-for-service: 7.4% increase, $p<.01$). However, increases primarily were due to better documentation, not to better immunization practices.	Small sample size in each study arm Short follow-up period for second round of incentives (12-month break between studies reduced likelihood of a carry-over effect from first study. Low response rate in follow-up study. Improvements observed primarily due to better documentation.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
	some, but not complete overlap of physicians participating in two studies, with 8 of the original physicians dropping out (distributed across the study arms) and an additional 12 physicians recruited to the control group. The feed-back only group from the original study was combined with the bonus group. The improvement required to obtain the incentive increased in this second study.				
Grady et al., 1997	Randomized controlled trial at practice level Prospective, longitudinal 3 study arms: 1) education-only control (breast cancer statistics, importance of mammography), 2) education plus cue enhancement (posters & chart stickers), 3) education plus cue enhancement plus peer comparison feedback & "token rewards"	Individual physicians (primary care) (n=95 physicians in 61 practices, with 23 in control, 18 in arm 2 & 20 in arm 3);	Financial incentive was a check based on the percentage of patients referred for mammography during each audit period (i.e. $40 for a 40% referral rate; $50 for a 50% referral rate). Payments were provided quarterly, but did not begin until second half of first year.	Across the entire sample, at both the practice and individual physician levels, mammography referral and completion rates increased in the first quarter in which the intervention began, and then steadily declined in the second, third and fourth quarters. Also at the practice and individual physician levels, compliance rates increased in the first quarter; then, in subsequent quarters, the two experimental arms showed continued increases in rates, while the control's rate essentially remained unchanged. At the practice level, repeated measures analysis of variance (ANOVA) confirmed overall differences among the study arms for mammography referral and completion:	Small sample size in each study arm Short study period Unclear how much of change was due to each part of the multifaceted intervention. Small reward.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
	1-year study period Key outcomes: 1) Referral rate: number of women referred divided by all women due for a mammogram, 2) Completion rate: number of women who received a mammogram divided by all due for one, 3) Compliance rate: number of women who received a mammogram within 14 months prior to end of each quarter divided by entire eligible patient sample			$F(2,58) = 3.99$, $p<.05$, and $F(2,58) = 4.32$, $p <.05$ respectively; contrasts showed that the differences were between the control group and the cue and the cue plus reward groups. At the physician level, the annual rates across study arms were not different for referrals but were significantly different for completions: $F(2,94)=3.7$, $p<.05$) While both interventions resulted in improved outcomes measured compared to the control group, the addition of feedback and financial reward did not improve outcomes beyond cueing alone.	

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Hillman et al., 1998	Randomized controlled trial at practice level, plus physician survey 2 study arms: 1) control, 2) physician-specific feedback every 6 months and a site-level financial bonus Repeated measurement (baseline, and 6 month intervals thereafter for 1.5 years) 18 month study period Key outcome: Compliance with cancer screening guideline; physician awareness of program also tracked with a survey	Practice sites (primary care), stratified by solo vs. group practice type (n=53, half randomized to each arm)	Financial bonuses paid to the 3 intervention sites with the highest compliance scores; these sites received "full bonuses" of 20% of capitation for eligible patients. The 3 next highest scorers and the 3 sites improving the most from the previous audit paid a "partial bonus" of 10% of their capitation amount.	Screening rates in both the intervention and control groups doubled over the study period (from 24% to 50%) with no significant differences detected between intervention and control groups. Bonuses ranged from $570 to $1,260 per site; average of $775 per audit; 17 of 26 sites received at least one bonus during study. Regarding physician awareness of program, of the 18 responding sites, 12 (67%) were aware of the study after the second mailing.	Short study period Small sample size in each study arm Unclear how much of change was due to each part of the multifaceted intervention in arm 2. Unclear how sites used the incentive and whether it was shared with individual physicians.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Hillman et al., 1999	Randomized controlled trial at practice level, plus physician survey 3 study arms: 1) control, 2) physician-specific feedback every 6 months, 3) feedback plus site-level financial bonus Repeated measurement (baseline, and 6 month intervals for 1.5 years) 18-month study period Key outcome: Compliance with pediatric preventive care guidelines	Practice sites (primary care) (n=49, with 15 in control, 15 in arm 2, & 19 in arm 3)	Eligibility for bonus was based on total compliance score, which had to be a minimum of 20% for each indicator. The 3 intervention sites w/ highest total compliance scores were paid "full bonus" of 20% of site's total 6-month capitation for pediatric patients up to 7th birthday; 3 next highest scoring sites were paid "partial bonus" of 10% of capitation; 3 sites showing most improvement also were paid 10% if total compliance score increased by at least 10%.	Compliance with pediatric preventive care guidelines improved dramatically during study period in all 3 study groups re: total compliance scores (56 to 73%), immunization scores (62 to 79%) and preventive care scores (54 to 71%), but no significant differences observed between either intervention group and control, and no interaction (group by time) effects found. Bonuses ranged from $772 to $4,682 per site; average of $2,000; 13 of 19 sites received at least one bonus during study, and 6 received two bonuses during study. Regarding physician awareness of program, of the 27 responding intervention sites, 15 (56%) were aware of the study after the second mailing.	Short study period Small sample size in each study arm Unclear how sites used the incentive and whether it was shared with individual physicians.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Kouides et al., 1998	Randomized control trial at practice level, unblinded 2 study arms: 1) control (poster to track flu shots), 2) poster plus financial incentive 2-year study period Key outcomes: Influenza immunization rate, & improvement in influenza rate	Practice sites (solo and group) of physicians participating in the 1990 Medicare demonstration immunization project (n=54, with 27 in each arm)	Physicians reaching a 70% influenza immunization rate for their practice received an additional 10% reimbursement ($.80) per immunization above the $8 administration fee paid within the Medicare Demonstration Project, & an additional 20% ($1.60) if they reached 85%.	The financial incentive resulted in a 7% increase (from baseline) in influenza immunization rates in the incentive group, relative to the control group (p=.05). The mean immunization rate for the incentive group was 68.6%, compared to 62.7% for control group (p=.22).	Study participants were not blinded. All participants were part of the Medicare Influenza Vaccine Demonstration Project (but had not received incentive reimbursement), which reduces generalizability, because of community effort to increase use of flu shots. Small incentive

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Roski et al., 2003	Randomized controlled trial at medical clinic level 3 study arms: 1) control (distribution of smoking cessation guidelines), 2) financial incentive to clinic-only 3) financial incentive plus computerized patient registry linked to telephone counseling 1-year study period Key outcomes: Physician documentation of smoking status and of advice to quit smoking, as well as patient smoking cessation rates after 1 year	Medical clinics (n=40, with 15 in arm 1, 15 in arm 2, and 10 in arm 3)	Financial incentives based on meeting the following fixed performance targets: 1) documented smoking status for 75% of all patients 18 years & older, 2) documentation of advice to quit smoking at last visit for 65% of current smokers. Clinics with 1-7 physicians were eligible for up to a $5,000 bonus; those with 8 or more were eligible for up to $10,000 bonus. Clinics that reached or exceeded only 1 of 2 performance goals were eligible for half these amounts (i.e., $2,500 & $5,000, respectively). The clinics were free to allocate incentive payments as desired.	Identification of patients' tobacco use status significantly improved in the incentive-only group (by 14.4%) and in the incentive-plus-registry group (by 8.1%), compared to the control group (6.2%) (p=.009). Clinical practice rates for advising smokers to quit, and for providing smokers with assistance to quit did not differ significantly between experimental conditions. No significant impact on smoking cessation rates.	Small number of clinics in each arm.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
QUASI-EXPERIMENTAL DESIGN					
Levin-Schertz et al, 2006	Pre-/post-test with external comparison Statewide HEDIS scores (without PCHI) served as control 3-year study period Key outcomes: HEDIS measures including: 1) HbA1c Screening, 2) Diabetic LDL Screening, 3) Diabetic Eye Exam, 4) Diabetic Nephropathy Screening, 5) Asthma Controller Use	Physician Network	Financial incentive: withholds of about 10% of physician fees as well as bonuses of $3000 - $5000, based on explicit target, usually previous year's state or national 90th percentile for HEDIS measures. System support accompanied P4P in one of the health plans in the network - PCHI, involving a computerized patient registry that was able to identify patents that had not received particular health care services to enable proactive outreach by non-clinical staff to encourage patients to visit their health care provider to receive indicated care.	Scores on adult diabetes measures for PCHI improved significantly between 2001 and 2003. There was not a significant improvement in the childhood asthma control measure. On the diabetes measures, PCHI experienced greater improvements than both the state of MA and national HEDIS benchmarks. PCHI experienced greater improvements, significant at $p<.05$. PCHI did not experience greater improvements than MA for asthma measure, but PCHI baseline performance was approximately the national 90th percentile for the measure and was higher than the state average performance. The physician network met their P4P contracts' target goals for diabetes and asthma measures for each of the study years.	Unclear whether incentive went to individual physicians. Lack of comparable control group State comparison contaminated by other groups in states striving for improvement on same set of measures, some of which were also receiving incentives. Unable to distinguish impact of incentive from impact of system support.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Rosenthal et al., 2005	Quasi-experimental, case/control 2 study groups: 1) control (public report card only), 2) (public report card plus financial incentive) 2.5-year study period Key outcomes: performance on 3 process measures of quality (cervical cancer screening, mammography, & HbA1c testing)	Medical groups (n=205, with 163 in intervention, 42 in control group)	Participants were eligible for a quarterly bonus of about $0.23 PMPM or about 5% of the professional capitation amount, based on their performance on 10 clinical and service quality measures.	Significant improvements in all 3 clinical quality scores were observed for the intervention group: cervical cancer screening (5.3 percentage points, p<.001); mammography (1.9 percentage points, p=.04); and HbA1c (2.1 percentage points, p=.02), while significant improvements in the control group were only observed for cervical cancer (1.7 percentage points, p=.03). Compared to the control, the intervention group showed greater quality improvement only in cervical cancer screening (3.6 percentage points difference, p=.02). For all 3 measures, groups with baseline performance at or above the performance threshold for receipt of a bonus improved the least but received the largest share of bonus payments. The mean quarterly bonus payment to each medical group during first year increased from $4,986 (July 2003) to $5,437 (Aug 2004).	No random assignment Physicians in control group started at higher base level performance Unclear how much of change was due to incentive vs. secular or regional effects. Relatively short time period examined. Plan accounted for on average only ~15% of practice revenues.
Pre/Post-Test Design Without Controls					

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Amundson et al., 2003	Pre/post-test No control group 3-year study period Key outcomes: Tobacco usage tracking and advice to quit smoking rates	Medical groups (n=20), varying in size from 16-500 physicians	Incentive tested had the following components: 1) Bonus: Medical groups paid from a bonus pool for meeting performance target of asking 80% of patients about their smoking status and advising 80% of smokers to quit. Groups could receive between $6,650 and $43,750 in annual bonuses. 2) Feedback: Groups received feedback on results for each medical group by name at baseline and 1-year intervals. 3) Public report: High-performing medical groups publicly recognized. Note: Tobacco bonus was 1 of 4 clinical quality measures that comprised the recognition program.	From 1996 to 1999, average rates of asking patients about their smoking status (24% increase, $p<.001$) and advising smokers to quit (21% increase, $p<.005$) increased significantly across all medical groups. None of the medical groups increased their rates enough in the first year to receive the bonus. Four groups received the bonus in the second year of the study and 8 groups received the bonus in the third year of the study.	No control group Unable to separate effect of feedback and public reports from incentive. Unclear how much of change was due to incentive vs. secular effects. Some improvements observed may be due to better documentation.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Armour et al., 2004	Pre-/post-test (multivariate logistic regression used to assess association between CRC screening and physician bonus eligibility) No control group 2-year study period Key outcome: Colorectal cancer (CRC) screening rates	Individual physicians (n not reported)	Annual financial bonus provided for improving the rate of colorectal cancer screening. Details of bonus program not reported, given proprietary nature.	Colorectal cancer screening increased by 3% (from 23.4% to 26.4%, p<.01) in the year after bonuses were implemented. Patients whose physicians received a financial incentive for providing colorectal cancer screening were more likely to have received the screening (p<.01).	No control group Bonus program details not presented (proprietary) Unclear how much of change was due to incentive vs. secular effects
Fairbrother et al., 1997	Pre-/post-test No control group 3-year study period Pediatric Medicaid population Key outcome: vaccination and other screening rates	Individual physicians (n=23)	Incentive tested had the following components: 1) Free vaccines and 2) Increased vaccine administration fee of $17.85 (vs. $2 standard fee) for every vaccination provided to children.	Up-to-date rates for all vaccinations combined increased significantly after the implementation of the incentive program (24.3% increase, p<.05). Other significant increases include: tuberculosis screening (by 28.8% , p<.05); lead screening (assuming all children are at low risk for lead exposure) (by 23.4%, p<.05); and well-child visits (by 6.6%, p<.05). No change in the rates of missed opportunities to immunize.	No control group Small number of participating physicians. Some improvements observed may be due to better documentation. Unclear how much of change was due to incentive vs. secular effects.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Francis et al., 2006	Pre/post-test Multifaceted intervention including: 1) provider and patient education, 2) provider profiling reports (including patient-specific data and 1-on-1 actionable feedback), and 3) financial incentive to encourage adherence to an acute otitis media (AOM) guideline. 3-year study period HMO patient population in IPA Key outcome: adherence to otitis media guideline	Individual physicians (pediatricians, internists and family practitioners) (n=486)	Financial incentive: variable withholds based on patient satisfaction (20%); efficiency (40%), and clinical quality (40%) for a variety of conditions. Results across 3 measurement categories combined to obtain total score. The capitated IPA kept a percentage of the capitation in reserve ("withhold") to accommodate for increases in utilization. From 1999-2001, withhold amount was 15% on each physician service. In 2000, withhold decreased to 10% for the top 5% of performers and increased to 20% for bottom 5% of performers, based on total score.	Pediatricians and internists significantly reduced overall exceptions to the guidelines per 1,000 episodes from pre to post intervention period (18.2% and 14.7%, respectively, p<.000). Family practitioners' reductions were not statistically significant. Additionally, all 3 specialties significantly decreased their use of less effective/ inappropriate antibiotics (41.5% reduction among pediatricians; 22.1% reduction among internists; 14.7% reduction among family practitioners). Across the 3 specialties, no statistically significant reductions were detected in: use of first-line antibiotics before second-line, decreased antibiotic prescriptions prior to office visits; or use of appropriate radiology procedures. When taking all pathways together, overall adherence improved.	No simultaneous control group Unclear how much of change was due to incentive vs. secular effects. Unclear how much of change was due to each part of the multifaceted intervention.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Greene et al., 2004	Pre-/post-test cohort Multifaceted intervention including: 1) education (training, tool kit w/care pathway), 2) profiling reports, 3) financial incentive 3-year study period HMO patient population in IPA Key outcomes: Adherence to guideline and improved antibiotic use for treatment of acute sinusitis	Individual physicians (primary care) (n=~900)	Financial incentive: variable withholds based on patient satisfaction (20%); efficiency (40%), and clinical quality (40%) for a variety of conditions. Adherence to sinusitis guidelines accounted for up to 50% of the quality component. Results across 3 measurement categories combined to obtain total score. The capitated IPA kept a percentage of the capitation in reserve ("withhold") to accommodate for increases in utilization. From 1999-2001, withhold amount was 15% on each physician service. In 2000, withhold decreased to 10% for the top 5% of performers and increased to 20% for bottom 5% of performers, based on total score.	Exceptions to guidelines per-1000 episodes decreased 20% ($p<.005$) from 326 to 261. Decreased use of less effective/inappropriate antibiotics accounted for most (32%) of the change (from 199 to 136 exceptions per 1000 episodes). Inappropriate radiology use decreased 20%, from 15 to 12 per 1000 episodes. All changes significant at $p<.005$.	No simultaneous control group Unclear how much of change was due to incentive vs. secular effects. Unclear how much of change was due to each part of the multifaceted intervention.

Author, year	Study design	Incentive target	Description of incentive	Results	Study Limitations
Morrow et al., 1995	Pre/post-test, cohort Multifaceted intervention including: 1) peer review, 2) feedback, & 3) financial incentives. 3-year study period Key outcomes: rates of MMR immunization, cholesterol screening, and charting adequacy	Practice site (n=418 for MMR immunization audits; 271 for cholesterol screening audits, & 1,607 for charting adequacy)	Financial incentive: Physician reimbursement varied based on utilization of services (e.g., hospital days) and quality elements (e.g., chart audits, member surveys). Each of the utilization and quality elements were assigned a numerical value, and the score earned determined the amount of capitation and the dollar amount and frequency of distribution of additional funds paid to the primary physician.	Over 3 years, offices meeting MMR vaccination standards increased from 78% to 96% ($p<.05$, all 3 audits); those meeting cholesterol screening standards increased from 92% to 95% ($p<.05$ for first 2 of 3 audits only). Average charting accuracy scores rose from 87% to 92% ($p<.05$, all 3 audits). The percentage of practices not in compliance with a standard of 90% decreased as follows: from 57% to 12% for MMR vaccination ($p<.05$, all 3 audits); from 21% to 11% for cholesterol screening ($p<.05$ only for first 2 of 3 audits); and from 53% to 29% for charting adequacy ($p<.05$, all 3 audits).	No control group Unclear how much of change was due to incentive vs. secular effects. Unclear how much of change was due to each part of the multifaceted intervention.

The researchers who conducted the 15 studies suggested several possible explanations for the observed results:

- **Size of incentives.** A common explanation for the lack of effect was that the amount of the incentive used in the program was probably too small a percentage of revenue to influence behavior (Rosenthal et al., 2005; Hillman et al., 1999; Hillman et al., 1998; Roski et al., 2003). This is especially true for measures that require substantial practice investment to secure improvements (Greene et al., 2004). There is some evidence to suggest incentives need to be a minimum of 5 percent of practice revenues to influence behavior (Hillman et al., 1991).
- **Length of study period.** A factor cited as a potential reason for the modest effects was that the study period was too short for positive program effects to be seen (Hillman et al., 1998; Grady et al., 1997; Rosenthal et al., 2005). The empirical studies in most cases measured the impact of interventions that were a year or shorter in duration, and in limited cases the intervention ran for 18 to 36 months. Implementing practice changes may take more time than these studies provided.
- **Improved documentation.** A possible explanation for the increase in performance scores that was observed in several studies, and confirmed in two, is that a portion of the observed program effect resulted from improvements in medical record documentation and charting, rather than from actual changes in performance (Amundson et al., 2003; Armour et al., 2004; Fairbrother et al., 1997; Fairbrother et al., 1999; Fairbrother et al., 2001). Because improvement in documentation typically occurs in the early years of a P4P program, studies seeking to gauge programmatic impact need to measure improvements after the initial implementation period.
- **Other factors.** Performance monitoring, whether it is tied to financial incentives or not, can have the overall effect of improving performance (Roski et al., 2003), which raises the possibility that all or some of any observed positive effects were the result not of financial incentives, but of increased monitoring in the practice environment. One study whose findings were positive had used a combination of P4P, system support in the form of a patient registry to identify patients for which health care services were indicated, and internal reporting of performance scores for physician groups (Levin-Scherz et al., 2006). These positive results may be partially explained by the effects of the system support and performance reporting, rather than by the P4P component of the program. In several other

studies, the observed improvements were also observable in national trends, such as increases in childhood immunization rates that may be associated with national efforts to improve HEDIS measures (Roski et al., 2003; Rosenthal et al., 2005; Hillman, 1999; Hillman et al., 1998). Another potential external factor is the pressure within a health care organization to focus on other aspects of care, which could have distracted physicians who were participating in the P4P study (Roski et al., 2003). In at least three studies, researchers believed that low awareness of the intervention contributed to the studies' results (Roski et al., 2003; Hillman et al., 1999; Hillman et al., 1998). Low awareness could have been caused by a number of things: the number of issues competing for the physicians' attention, the program sponsor not having invested sufficient resources in publicizing the intervention, or the intervention failing to capture attention because it represented too small a share of the physician's or practice's business. In the case of incentives being provided at the group level, individual physicians may not have been adequately educated about the measures or incentives, or incentives may not have been shared with individual physicians.

LIMITATIONS OF EMPIRICAL STUDIES ON PAY-FOR-PERFORMANCE

The existing body of peer-reviewed empirical literature on P4P programs is small, and most of the studies evaluated financial incentive experiments that occurred in the late 1990s or early 2000s. These interventions were small in scale, and most were very short in duration, which limited the likelihood of seeing an effect. The key limitations of these studies are as follows:

- **Lack of concurrent control groups.** The most common limitation of the reviewed studies is the lack of concurrent control groups. When there is no control group, there is a possibility that the observed program effects were at least partially the result of other factors. This is especially the case in these studies, since a number of the targeted performance measures were being addressed by other quality improvement activities (e.g., anti-smoking campaigns, physicians being exposed to quality improvement efforts within their organizations, plan HEDIS measurement and improvement activities, and public reporting). Furthermore, with the exception of one study (Rosenthal et al., 2005), the pre-and-post-intervention comparisons did not control for pre-intervention trends in performance already occurring. In at least one study, similar increases in

immunization rates were seen in the general population at the same time as the study period (Fairbrother et al., 1997).

- **Poor generalizability.** The results of the published studies are difficult to generalize to other populations and settings given that they typically occurred at a single site or dealt with unique physician populations (i.e., targeted Medicaid providers, who typically receive the lowest reimbursement rates and may respond differently than physicians treating commercial or Medicare patients). The published studies also focus on testing the effect on one or only a small handful of clinical quality measures—where it may be easier to generate behavior change; whereas today's programs are rewarding a large number of performance measures across multiple domains (patient experience, clinical quality, cost-efficiency).
- **Limited use of and unblinded randomized control trials.** Fewer than half of the studies used randomized control trials (RCTs); and two of them suffered from other limitations, including data collectors, study personnel, or participants not being blinded to group assignments (Fairbrother et al., 1999; Kouides et al., 1998). However, both of these studies attempted to determine whether this particular type of limitation introduced bias in the study results and concluded that it more than likely did not.
- **Other limitations.** One study that had a small sample also had groups that differed significantly at baseline, although the researchers attempted to control for these differences in analysis (Fairbrother et al., 1999). A number of the studies examined multifaceted interventions (e.g., physician education plus feedback reports plus a financial incentive); however, these studies often were not designed so that the degree to which each individual aspect of the intervention affected the outcome(s) could be examined. One study that involved system supports—in this case, a patient registry to identify patients who had not received needed services—suffered from a crucial limitation: it was unable to isolate the effect of the P4P program from that of the patient registry on the observed improvements in performance (Levin-Scherz et al., 2006). A number of the studies measured a small number of providers, so the small sample size would have required a large difference in scores between intervention and control groups to conclude that the effect was statistically significant and not due to chance. The short intervention follow-up periods (e.g., one year or less) for some studies reduced the likelihood of seeing behavior change, since most P4P programs require time for physicians

to become aware of the program, understand the incentive system, accept it (Beckman 2006), develop an improvement strategy, and change their practices.

SUMMARY OF FINDINGS

Taken together, the findings of this literature review suggest that it is still too early to determine the effect of physician-focused P4P programs. The published literature provides an ambiguous set of results. Although several of the studies yielded positive results, their designs are often lacking the rigor to separate the effect of the incentive from the effect of other factors occurring in the environment. Importantly, not one of these studies examines the more comprehensive types of P4P programs that are emerging rapidly across the country today. The studies also provide no information on the various design features that may or may not affect the likely success of the intervention, such as level of engagement and communication with the providers, bonus size, intervention's length of time in operation, and share of a physician's patient panel that is represented by the intervention.

Finally, most of the programs evaluated in the 15 studies do not resemble the type of P4P programs in operation today—not in size (the number of measures or providers), duration, or magnitude of rewards. This makes it nearly impossible to generalize from the findings of these evaluations to what might occur as a result of the interventions in operation today. In addition, today's P4P programs that are through commercial health plans are the product of contract negotiations between the physician contracting entity and the health plan, which means that those negotiations determine the measure that will be included, the thresholds for receipt of the incentive, and other program characteristics. This aspect of the programs may also have a bearing on P4P program effects.

Some of the current larger, sustained P4P programs are currently being evaluated, but results from these studies are just starting to emerge, and most of these studies focus on implementation experiences. The current P4P programs reflect real-world experiments, often with multiple interventions (e.g., financial incentives and public reporting), rather than controlled trials. In consequence, they suffer from some of the same methodological problems that the earlier studies did (e.g., lack of control groups), which limits the ability to isolate and draw definitive conclusions about the effect of P4P on performance. To understand the true impact of P4P on performance improvement, the interventions and the evaluations of the interventions must be carefully designed.

3. A REVIEW OF EXISTING PAY-FOR-PERFORMANCE PROGRAMS

This chapter describes key design features of current P4P programs that are operating nationally in both the private and the public sector. It also highlights some of the important lessons that have been learned in these efforts.

The vast majority of P4P programs targeted at physicians or groups of physicians are occurring in the private sector. However, our review also examines early lessons that are just emerging from the recently begun CMS Physician Group Practice (PGP) demonstration program. Publicly available program information is supplemented with insights and perspectives gathered through our discussions with an array of P4P program sponsors and six medical groups participating in the PGP demonstration.

METHODS

Private and Public Pay-for-Performance Programs

Information on private and public P4P programs was compiled based on several sources:

- A review by Rosenthal and colleagues (2004) of 37 P4P programs.
- A 2004 Med-Vantage[1] study by Baker and Carter (2005) that identifies 84 health plan P4P programs.
- The 2005 Med-Vantage national survey of P4P programs (Baker personal communication, 2005).
- The Leapfrog Compendium of incentive and reward programs (The LeapfrogGroup, 2005).
- Discussions with major professional organizations whose members either sponsor or are the target of P4P programs—i.e., Association of Health Insurance Plans (AHIP), the American Medical Association (AMA), the American Medical Group Association (AMGA), Blue Cross/Blue Shield Association (BCBSA), California Association of Physician Groups (CAPG).
- A review of the CMS website (www.cms.gov).

[1] Med-Vantage is a health informatics company that focuses on P4P.

- A Lexis/Nexis search of major U.S. newspapers, a broad Google-based Internet search, and a search of relevant trade journals.[2]
- The expertise of RAND project staff who have been directly involved in evaluating several of the national Rewarding Results P4P demonstrations.
- Input from our project's advisory panel, some of whom currently operate or are involved with P4P programs.

Since most publicly available information on P4P programs provides little detail on design elements, implementation processes, and lessons learned, we conducted semi-structured discussions with project staff from a subset of identified P4P programs to gain a more comprehensive understanding of these issues. We used the following criteria in selecting P4P programs for the discussions:

- Inclusion of physicians or physician groups as targeted program participants.
- Representation of various types of sponsors (i.e., single sponsors, coalitions, members of BCBSA, commercial health plans) and plans (i.e., health maintenance organizations [HMOs], preferred provider organizations [PPOs], and administrative services only [ASO]).
- Two or more years in operation. We wanted programs to have experience making payouts and working through a range of design and implementation issues.

We did hold discussions with some programs that had been in operation for less than two years, but only if they had unique characteristics warranting further exploration.

After reviewing the criteria and our final candidate list with our project advisory panel, we invited 24 organizations with P4P programs to participate in the study. We held discussions with 20 of these program sponsors between January 2006 and March 2006.

We also held discussions with six of the 10 medical group sites participating in the CMS PGP demonstration program, which is one of seven ongoing CMS P4P demonstration projects (see Appendix B) and summarize some of the emerging early lessons. The discussions occurred between December 2005 and January 2006, when these

---

[2] The journals searched were *Managed Care*, *Hospitals and Health Networks*, *Modern Healthcare*, *Managed Health Care Executives*, *Healthcare Intelligence Network*, *Medical Economics*, *Managed Care Weekly*, *Modern Physician*, *Business Insurance*, *California Healthline*, *Managed Care Online*, and *Managed Care Magazine*. The search terms used included pay for performance, pay for quality improvement, financial incentive, bonus, reward, provider payment, performance improvement and quality initiative.

projects were three-quarters of the way through the first year of program implementation. The six sites that participated were chosen to represent diversity on a number of characteristics, including geography, rural versus urban practice settings, academic affiliation, mix of primary versus specialty care, and size of practice.

National Overview of Current Pay-for-Performance Programs

We identified 157 P4P programs sponsored by 130 organizations[3] covering over 50 million beneficiaries as of December 2005. The 130 sponsoring organizations in 2005 represent a 67 percent increase over the 78 sponsors identified as of November 2004 in the annual Med-Vantage survey. A similar increase year-over-year was observed in the number of covered beneficiaries (Baker and Carter, 2005). In addition, CMS currently has seven P4P demonstrations under way. Of the P4P programs we identified, the vast majority are sponsored by health plans (80 percent). Employers or employer coalitions sponsor about 6 percent of all programs; Medicare sponsors 4 percent. Of the 33 programs that serve Medicaid beneficiaries, we identified 11 in which the state provides incentives to Medicaid managed care organizations. The remainder are sponsored by health plans.

P4P sponsors are most likely to have programs through their HMO products (88 percent). About 25 percent of sponsors have programs in PPO products, and 24 percent have programs in their ASO products. Health plans frequently develop different programs according to differences in

- The organizational structure of physicians in various health care markets.
- The unit of accountability (medical group versus individual physician) by insurance product.
- Payment structures (capitation versus FFS).
- Avenues available for engaging and communicating with physicians.

[3] If an organization has multiple distinct programs, we counted each program separately. Three collaboratives (i.e., Integrated Health Care Association, Local Initiative Rewarding Results, and Massachusetts Health Quality Partners) represent multiple incentive sponsors (i.e., health plans). We included the participating plans but not the umbrella organization in the count.

Summary of Findings from the Med-Vantage Survey

Med-Vantage, Inc., conducts an annual survey of P4P programs in the United States. The following list highlights some of the descriptive findings on P4P program characteristics from Med-Vantage's 2005 survey of 82 respondents (Med-Vantage, 2005):

- **Motivation.** The primary reason sponsors say they implement P4P is to generate improvements in clinical outcomes.
- **Incentive recipients.** All physician P4P programs included incentives for primary care physicians; 52 percent of programs included specialists, and 37 percent included hospitals. Part of the observed evolution of incentive programs is that they initially target primary care physicians for the incentive and then expand to include specialists. Programs including both primary care physicians and specialists became more common in 2005 compared with 2003 and 2004.
- **Specialists included.** P4P programs that included specialists most frequently measured and rewarded obstetrics-gynecology (70 percent), cardiology (58 percent), and endocrinology (47 percent).
- **Level of performance measurement.** About two-thirds of all P4P programs (64 percent) measure individual physician performance.
- **Measures used.** Clinical quality measures are the most common (91 percent) measures used in the programs studied, followed by cost-efficiency or resource utilization (50 percent) and information technology (IT) adoption (42 percent). The use of patient satisfaction measures declined from 79 percent in 2003 to 37 percent in 2005. Med-Vantage suggests that the decline may be because of the cost of conducting patient surveys to obtain information at the level of the individual physician.
- **Weighting of domains.** Domains of measures (e.g., clinical quality, cost-efficiency, patient experience) have differing importance or weight in the formulas used to determine provider rewards. The average weighting for clinical measures (52 percent) was higher than that for either resource use/cost-efficiency measures (35 percent) or IT adoption (26 percent). Thus, on average, clinical information accounts for approximately half of the providers' rewards. Large variations exist across programs in the weighting of categories of measures. The reported weighting for each domain (clinical quality, efficiency, and IT adoption) ranged from 5 percent to 100 percent.

- **Appeals process.** Approximately three-quarters of program respondents stated that they have an appeal process in place to address inaccurate data, performance scores, and other concerns.
- **Form of incentive.** Bonuses are the most common form of incentive payment in physician P4P programs (88 percent).
- **Incentive amount.** Across respondents, the average maximum bonus a physician could earn was 9 percent of total revenue from the sponsor, but the maximum for individual programs ranged up to 20 percent of revenue from the sponsor.
- **Non-financial incentives.** Of all the programs, 46 percent provide incentives in the form of simplified or alleviated administrative requirements, such as pre-certification for certain procedures. Thirty percent of programs use public performance reporting and a publicized "physician honor roll," and 30 percent provide performance feedback to providers via internal, web-based reports at a frequency ranging from quarterly to annually.

SUMMARY OF FINDINGS FROM DISCUSSIONS WITH PAY-FOR-PERFORMANCE PROGRAMS

Of the 20 P4P program sponsors we held discussions with, nine had started their programs by 2000, nine had started their P4P programs between 2001 and 2004, and two had just formed their program or were in the pilot stage. Program planning typically started at least a year prior to the program's actual implementation; one P4P sponsor said a minimum of two years was necessary for planning. Five sponsors reported partnering with other organizations around their P4P programs. These partnerships were varied in nature and could be between groups of health plans, between a health plan and an independent practice association (IPA), or between employers. Eight of the sponsors reported their organization as having more than one P4P program. An organization could have different programs based on region, product line, provider type, or program focus (quality versus efficiency); and in some cases, the programs were quite dissimilar from one another. Here is the breakdown:

- Sixteen programs included commercial HMO and point-of-service (POS) populations.
- Ten programs included PPO populations.
- Five included self-insured populations.
- Four included Medicare populations.
- Three included Medicaid populations.

Although some sponsors had hospital P4P programs, we focused our discussions only on their physician P4P programs.

Below, we summarize information on program design features, program evolution, and efforts to engage physicians. While we did not conduct a random sample of all P4P programs, the findings from our discussions with 20 programs are consistent with the findings from the Med-Vantage survey of a broader set of programs in which topics overlapped.

Selection of Areas Subject to P4P

- **Program goals and objectives.** All of the P4P sponsors reported that improving or maintaining quality of care or the health of their membership was a goal of their program. Twelve sponsors reported cost savings (i.e., improved cost-efficiencies) as a goal. Other program goals included improved patient satisfaction and experience (n=5), providing recognition to outstanding providers (n=2), improved patient safety (n=2), and decreased variation in patterns of care (n=2). Goals mentioned by single respondents included increased ease of working with medical groups, changing provider behavior, engaging physicians and getting them to work together, encouraging physicians to think about processes of care, improved provider satisfaction, and promoting the Institute of Medicine's (IOM's) six aims for care delivery.
- **Factors influencing program design.** The selection of clinical areas was driven by the prevalence of disease conditions (e.g., heart disease, asthma, diabetes), the existence of evidence-based measures, the opportunity for improvement in care (high variability in treatment), and an expected return on investment (ROI) based on evidence from the literature. The availability of evidence-based measures was the most commonly stated factor influencing the design of P4P programs and their measurement areas (n=15), whereas the endorsement of measures by national organizations such as NCQA, Ambulatory Care Quality Alliance (AQA), and the Institute for Clinical Systems Integration (ICSI) influenced five sponsors. Six program sponsors explicitly mentioned that they targeted areas identified as having opportunities for improvement or variations in quality of care delivered; five sponsors mentioned prevalence of conditions, and five mentioned acceptance of measures by local doctors. Ease of data collection (n=6) and data accuracy (n=1) also factored into the selection of measurement areas.

Role Providers Played in Development of P4P programs

- **Provider involvement and awareness.** Thirteen of the 20 P4P programs involved providers in the selection of performance measures, and four involved providers in overall program design. Six had a mechanism for obtaining ongoing input from providers, including focus groups, input from quality committees, meetings with key providers, information gathering through account managers, and physician surveys.

Measures, Data, and Risk Adjustment

- **Measures used.** Clinical quality measures were used by all 20 programs. Health Plan Employer Data and Information Set (HEDIS) measures or variations thereof are most commonly used for PCPs, in part to align physician incentives with what health plans are being held accountable for in NCQA accreditation requirements. There was less agreement around measures for specialists, with almost half of the programs that scored specialists having internally developed clinical measures, largely as a result of the lack of available measures for specialists. Resource use or cost-efficiency measures were included in 15 of the programs; patient experience was measured in nine programs. The overall high level of patient satisfaction, lack of variation across providers, and costs of conducting the surveys were cited by a couple of programs as reasons for no longer including patient experience measures. Administrative measures were included in nine programs; IT measures were included in six programs. Sixteen of the 20 sponsors reported that measures were piloted prior to being included in the P4P program; the two most frequently mentioned methods were starting programs in a a small group of physicians and a one-year measurement-and-reporting period prior to paying out on the measure. Programs that did not originally pilot their measures stated that this was a costly mistake and that they now pilot all measures prior to including them in the P4P program.
- **Data used.** Claims data and other administrative data were used by 19 of the 20 sponsors to construct at least some of their measures. Ten sponsors used data from medical charts to supplement or in lieu of claims data for clinical measures. Programs differed as to whether the physician or the program sponsor reviewed charts. Programs also used physician self-reported data for measures (n=7) and patient surveys (n=11). While claims data were used to construct measures for all

patients meeting the eligibility criteria for a performance measure, chart reviews and surveys were generally conducted with a sample of eligible patients.

- **Risk adjustment.** Eleven sponsors reported that they used risk adjustment on at least one of their cost-efficiency or health outcome measures. Methodologies mentioned included Diagnostic Cost Groups (DCGs), Episode Treatment Groups (ETGs), and Adjusted Clinical Groups (ACGs). A number of program sponsors acknowledged that patient noncompliance was a potential issue for P4P programs that include patient health outcome and intermediate outcome measures (e.g., blood pressure control). One program sponsor stated that 6 percent of patients with chronic conditions do not visit their primary care physician and that 1 to 2 percent do not visit any physician in a given year. None of the sponsors, however, had examined the extent to which this affected individual physicians' scores or whether non-compliant patients had experienced decreased access to care since the initiation of the P4P program. While utilization measures were frequently adjusted for patient demographics and case mix, intermediate outcome measures (e.g., LDL levels) typically were not risk adjusted.

Pay-for-Performance Program Participants and Eligibility Criteria

- **Incentive recipients.** All of the 20 P4P sponsors included primary care physicians in their P4P programs, and 14 included specialists. Reasons cited for not including specialists included the difficulty of attributing patient care to specific providers and the lack of a robust set of existing performance measures for specialists.
- **Eligibility criteria.** To determine which physicians to include for scoring, P4P programs used varying eligibility criteria, such as a minimum number of enrollees with the sponsor (n=9 programs), a minimum number of encounters with enrollees (n=4 programs), a minimum dollar amount for evaluation and management (E&M) visits (n=3 programs), participation in a certain type of product (n=2 programs), affiliation with a participating physician organization (n=3 programs), submission of electronic claims above a threshold (n=2 programs), participation in the sponsor's physician network at the time of reward payout (n=2 programs), and strong group leadership (n=2 programs). Some programs used multiple eligibility criteria. A number of the criteria were devised to ensure that the number of events used to score a physician or a physician group would be sufficient to produce an accurate and stable estimate of performance. In

addition to using criteria for determining eligibility for the program overall, many programs also had minimum requirements for individual performance measures, which ranged from 10 to 50 individuals qualifying for a measure's denominator..

Attributing Patients to Doctors

- **Attribution methods.** P4P program sponsors described five different methods they were using to attribute patients to primary care physicians. Among those reporting their attribution method, the most common method was to use the assigned PCP (n=10), which works only with HMO populations. Other assignment methods included (1) the highest volume of E&M visits (n=5), (2) the greatest share of patient costs (allowed amounts) either overall or on E&M services (n=4), (3) the highest volume of preventive care (n=1), and (4) all physicians who "touch" the patients regardless of specialty (n=1). The methods used for attribution to specialists were similarly diverse and, in some cases, varied by the specific measure or specialty. Methods reported were (1) E&M service volume (n=6), (2) greatest share of patient costs (n=4), (3) all physicians of the relevant specialty who "touch" the patient (n=2), and (4) all physicians who "touch" the patients regardless of specialty (n=1).

Reward Structure

- **Level of performance measurement and payment.** Slightly more than half of the programs (n=11) measured primary care physicians at the individual level. Programs were slightly more likely to measure specialists at the medical group or practice site level (n=8), in part because of concerns about having a sufficient number of events to score a physician, as well as challenges with attributing care to specialists. Some programs combined individual measurement and group or practice site measurement, the combinations depending on sample size, structure of practices, and ease of reporting.
- **Weighting of measures in payment formulas.** Among the programs willing to share information on this topic, there was considerable variation in the weighting of measure domains (n=13). Clinical measures had the highest average weight, at 50 percent (range of 25 percent to 100 percent). We combined the weights for cost-efficiency and pharmacy use measures (e.g., generic prescribing rate, adherence to formulary); this combined category had an average weight of 23 percent (range of 0 to 65 percent). Patient satisfaction and experience had an

average weight of 10 percent (range of 0 percent to 30 percent); IT measures had an average weight of 4 percent (range of 0 percent to 20 percent); administrative measures had an average weight of 5 percent (range of 0 percent to 20 percent).

- **Form of incentive.** P4P program sponsors most frequently used bonus payments (n=16). Withholds (n=4), shared savings (n=6), and modified fee schedules (n=3) were also used. Two programs reported their intent to move from a bonus payment to a modified fee schedule. Incentive payouts usually occurred on an annual (n=8) or semi-annual basis (n=6), the reason being to simplify program administration.
- **Financing pay for performance programs.** A variety of not mutually exclusive methods were used to fund P4P programs: (1) premium increases (n=7), (2) sharing of savings generated through increased cost-efficiencies (n=7), (3) reallocation of existing resources (n=3), (4) withholds (n=4), and (5) direct payments from employers (n=1). Seven of the sponsors reported using "new money" to finance their programs. However, one of them likened the financing of programs to squeezing a balloon and said, "It's not like there is new money; you just have to move dollars from other activities where premium dollars would have been spent. P4P sponsors have a set amount of money available from premium increases, and it's just a matter of how these funds are distributed."
- **Performance target used to determine basis for reward.** Programs used a variety of performance targets to determine whether a physician or physician group was eligible for reward: (1) absolute thresholds (n=11), (2) percentile or other relative threshold (n=7), (3) improvement over time (n=4), (4) participation only (n=2), and (5) the achievement of group specific goals (n=1). In some programs, the target depended on the measure; for example, a program might use performance improvement targets for clinical measures and relative thresholds for cost-efficiency measures. In other programs, an individual measure could have a mix of target types for different levels of payout.
- **Factors driving the structure of the financial reward.** Among the factors mentioned in this category were a reward large enough to capture physicians' attention, the ability to include (and reward) as many physicians as possible, an easily understood reward formula, and timely award of the reward. No sponsors reported explicitly linking the amount of a reward to the estimated cost of reengineering a practice to successfully participate in the program and improve performance.

- **Number of physicians receiving payout.** The programs varied substantially in the share of participating providers that received a payout, ranging from 20 to 100 percent. Five sponsors reported that 90 percent or more of their eligible providers received at least some reward. The extent to which programs provided rewards to a small number of providers versus many providers reflects philosophical differences among program sponsors. The perspectives ranged from choosing to focus on rewarding only truly excellent physicians to wanting to engage all physicians and encourage quality improvement across the board. The share of physicians receiving payouts also may be affected by whether rewards are determined using a composite score, which either sums or averages performance across multiple measures, or piecemeal on the basis of individual measures. Fewer physicians tend to receive incentives when rewards are based on composite scores. Furthermore, the use of relative thresholds versus absolute thresholds may also affect the number of physicians receiving rewards.
- **Incentive amount.** The average amount of the payout varied across programs and in some cases was capped. Of the 20 programs that we spoke to, 15 were willing to share information about incentive amounts. The average incentive-through enhanced fee schedule for E&M visits was $6 for primary care physicians, with caps ranging from $9 to $18; it averaged $12 to $15 for specialists, with a maximum cap of $18. Other programs reported the range of payouts to individual physicians as $500 to $5,000. Some programs discussed incentives in terms of a percentage of revenue, with the averages ranging from one to 10 percent of a physician's revenues. Still other programs calculated incentives on a per-member-per-month (PMPM) basis, with an average payout of about $2 PMPM and caps of $3 PMPM or higher.

Program Infrastructure

- **Data auditing.** Not all programs audit the data used to produce measures. Of the 20 programs, 14 said they either audited data or reserved the right to audit data. One program cited cost as a reason for not auditing data.
- **Appeals process.** More than half of the programs (n=13) reported that providers were able to see their performance data prior to receiving their payout and that they had a mechanism in place by which providers could raise concerns about the accuracy of both their performance scores and the underlying data used to produce them. Some programs allowed physicians to submit additional data or

otherwise correct data prior to payout; other programs corrected identified issues prior to the next payout cycle.

- **Comparative performance feedback (either confidential or public).** Seventeen of the 20 program sponsors gave providers performance feedback. This which usually compared an individual physician's or medical group's performance with the mean performance of peers, but a subset of programs provided the full distribution of performance scores for providers. In most of these cases, peer data was blinded; in others, it was not. Sponsors stressed that the performance feedback information supplied to providers needs to be actionable.[4] Six of the 20 sponsors used public reporting, which typically occurred at the group or practice site level.
- **Program evolution.** Of the 18 P4P sponsors whose programs had existed for more than a year, 17 reported having made program changes over time. The most common changes were retiring or adding new measures (n=12), expanding participants (n=12), modifying payouts (e.g., making incentives larger) (n=6), modifying thresholds (e.g., increasing the performance needed to receive an incentive) (n=3), and revamping the overall program (n=4).
- **Support for physicians.** All but one program sponsor reported that it supplied participating providers with some form of support: education (n=12), technical assistance (n=8), patient registries (n=8), facilitation of best-practices sharing among physicians (n=4), reminder mailings to patients eligible for measures (n=2), and reminders to physicians about patients not yet receiving care (n=2). A number of sponsors reported that they also provided other support, such as programs to encourage and facilitate the adoption of IT, which were separate from their P4P activities.
- **Inducements for participation.** Only one sponsor reported using a specific inducement for participation in its P4P program. This took the form of data collection assistance for the participating provider's first reporting period in the program.
- **Return on Investment (ROI).** The 20 sponsors we spoke to universally felt that their programs were improving the care delivered. Many of them had not

[4] Information should not be provided only in the form of rates. It needs to be in a form that can be acted upon—e.g., a list of patients that have not received services covered by measures in the program.

performed ROI calculations, however, because of the difficulty involved. Four sponsors felt the methods for conducting ROI for P4P programs were not developed enough to be sound. Sponsors that did attempt to perform ROI calculations for at least part of their programs reported estimates ranging from $2 to $5 saved for every $1 spent. When an ROI or cost-benefit calculation was performed, the savings estimates were computed by projecting expenditures based on spending trends prior to implementation and comparing them with actual expenditures. Estimates of program costs were based on approximations of staff time to operate the program, as well as the cost of incentives. Administrative costs of operating the program typically were not formally calculated. Other sponsors reported that they had observed improved performance on quality measures, flat pharmaceutical expenditures, and reduced disease-specific costs, and that participants performed better than non-participating providers. However, these sponsors' programs did not have control groups, and the observed effects could have been produced by other factors in the environment.

Several other issues came up during the discussions that are worth noting:

Reservations About Public Reporting

In their effort to drive improvements in quality and cost, some P4P programs use non-financial incentives in the form of transparency—or reporting of performance results—combined with financial incentives. There was broad consensus among P4P sponsors that financial incentives alone will not solve quality and cost problems, and that additional mechanisms need to be employed in conjunction with dollars held at risk for performance. Two private-sector P4P programs that measure at a higher level of aggregation than the individual physician (i.e., medical group or practice site)—the Integrated Healthcare Association (IHA) and Massachusetts Health Quality Partners—do make or plan to make group-level or practice site performance scores fully transparent to the public; they view public transparency and accountability as an additional means to drive improvements. Additionally, CMS is making hospital performance results publicly available within the Hospital Quality Alliance (HQA) voluntary reporting program and in the Premier hospital P4P demonstration, where hospitals scoring at the 50th percentile of performance or higher have publicly released performance results.

Some P4P program sponsors expressed discomfort with public reporting. The reasons they cited for not publicly disclosing physician-specific scores were the potential

for physician backlash and fear of lawsuits associated with possible data inaccuracies. They also expressed concern that current P4P efforts are still somewhat experimental, and that a move to broader transparency at this stage may thus be premature. Program sponsors felt that the problems associated with public reporting could be minimized in various ways—for example, by engaging physicians in the measurement process and by allowing physicians to review and correct data prior to its being made public in order to ensure the integrity of the measurement process. There appeared to be general recognition among P4P program sponsors that physician-level measurement and public reporting were inevitable, and that if done in a fair way, transparency could serve as an important stimulus to quality improvement. Among programs that do make performance scores transparent, full disclosure typically did not occur until after at least one or two cycles of internal data collection and reporting.

Empirical Evidence on Effect of Public Reporting on Performance

There is little published literature that addresses the impact of public reporting, in part because this activity is relatively new. However, a recent study by Hibbard and colleagues (2003) examined the effect of public reporting of hospital quality performance results. The objective of the study was to see whether there was a differential impact on investments in quality improvement and on actual quality improvement when performance information was transparent versus when no feedback information was provided or only internal confidential feedback was provided. Hospitals whose results were made public responded by making much larger and statistically significant increases in the number of quality improvement activities they engaged in and improvements in their obstetric quality of care performance than hospitals that received either no performance information or only internal (relative to their peers) feedback on their performance. Similar results were seen for poorly performing hospitals in the three groups. NCQA also found that among the plans it measures on an array of quality indicators, poorer performers were less likely than high performers to report and had less improvement in performance over time (McCormick et al., 2002; Thompson et al., 2003). Concerns have been raised that public reporting may have unintended consequences, such as the avoidance of sick patients by physicians in order to improve quality rankings, the discounting of patient preferences, and the discouragement of clinical judgment (Werner and Asch, 2005).

Stability and Alignment of Measures

A key issue that arose in our discussions with program sponsors is the need for stability in the measures over time, since providers need time to make investments in data systems, staffing, and other quality improvement actions that address the specific clinical problems being measured. Another issue that some program sponsors expressed concern about is that CMS, as the largest payer, could dramatically shift the focus away from their own local efforts to measure and provide incentives for doctors if CMS does not choose measures that are aligned with measures already in use by local-level stakeholders. To the extent that alignment exists, there was acknowledgment among local P4P sponsors that having CMS engage in P4P would significantly help to leverage their local measurement and P4P activities. The vast majority of individual P4P sponsors represent only a small fraction of most providers' patient panels and find it challenging to garner the attention of providers. Some of the P4P program sponsors we spoke with believe that by coming to the table, CMS may create a tipping point for physicians to actively engage in efforts to stimulate quality and cost improvements in the system.

SUMMARY OF KEY LESSONS LEARNED FROM PRIVATE-SECTOR P4P PROGRAMS

There was a great deal of consistency in what the P4P program sponsors saw as challenges and as lessons learned for program development, implementation and maintenance:

- **Programs need to engage physicians during program development**. The most universal lesson stated by program sponsors was the need to interact with and engage providers from the beginning of program development. Many programs did this by involving physicians on committees to select measures and review measure specifications. A number of programs had mechanisms in place to obtain provider input throughout the life of the program and viewed being open to physician suggestions as critical for buy-in. Newsletters and other high-volume methods of communication with no interaction were viewed as not sufficient for engaging physicians; programs using these methods noted that engagement was a struggle for them. Tapping into physician leadership in the community was a successful strategy for programs.
- **Programs need to be piloted and will change over time.** The need to pilot test the implementation of measures and other implementation processes (e.g., audit, feedback reporting) at each step of the program was another common lesson.

Programs that had not initially built in pilot testing indicated that this was a serious mistake and strongly advised that all aspects of program design and implementation be tested. Two items mentioned repeatedly as being necessary were a willingness to be flexible and change, and the recognition that program development will involve some trial and error. Also mentioned repeatedly were the need to ensure the accuracy and reliability of claims data and other data underlying the measures, and the need for a fair and equitable process for appeals by providers concerned about their performance scores and data accuracy.

- **Physicians require support to successfully participate in P4P programs.** Most of the programs provided some form of support to providers as part of the P4P program. The provision of patient registries, technical support, and education were most common. The most important aspect is that the information provided be actionable. The provision of support enhances the provider's engagement and buy-in.
- **One size does not fit all.** Sponsors with programs in multiple markets noted the need to tailor each program to its market and that a one-size-fits-all approach does not work. The ways in which sponsors tailor programs include allowing regional selection of measures from a menu of measures and using different units of measurement (group versus individual physician) based on the organization of physician practices in the geographic market.
- **Administering P4P programs requires time and resources.** Multiple sponsors raised the issue that P4P programs take more time and more resources to manage than were initially anticipated. Infrastructure, both IT and personnel, is needed to support data warehousing capabilities; data aggregation; programming and analysis; data auditing; processes for appeals and data correction; provision of performance feedback; communication with and engagement and support of physicians; measures maintenance; and modification of data collection processes. One approach taken by programs is to start small, in terms of number of physicians or number of measures included, and gradually build the infrastructure to support the P4P program.

CMS PHYSICIAN GROUP PRACTICE DEMONSTRATION

In January 2005, CMS announced the initiation of the Physician Group Practice (PGP) demonstration, a three-year program to implement a P4P program for physicians who care for FFS Medicare beneficiaries. The program's goal is to improve care for

beneficiaries with chronic medical conditions by rewarding physician groups that manage patients across the continuum of care in a cost-effective, high quality manner. Ten PGPs across the United States, each with 200 or more physicians, were selected through a competitive process to participate in the program. Participants must meet efficiency and quality thresholds to be eligible to receive bonus payments. Annual payment cycles began in April 2005 and end in March 2008; January 1, 2004 through December 31, 2004 is the base year for comparison purposes.

Calculation of Bonus Payments

To be eligible for bonuses, PGPs must demonstrate cost savings, regardless of the group's level or improvement in scores on the quality measures. Risk-adjusted per capita expenditure targets for the PGPs are determined by applying to the PGP's base year spending the rate of expenditure growth in the population of non-participating, FFS beneficiaries in the PGP's community. Any amount that exceeds a 2 percent reduction below the expenditure target represents Net Medicare Savings (NMS). Program savings less than 2 percent are retained by Medicare, and the PGP is not eligible for a bonus. The demonstration capped potential bonus awards at a 5 percent reduction in expenditures below the target, with any additional savings being retained by Medicare. Thus, the groups are eligible to receive a portion of the savings of 2 percent to 5 percent below their targets.

Twenty percent of the NMS is set aside as Medicare program savings. A portion of the remaining 80 percent is forwarded to the group as cost savings, and a portion is eligible to be earned by the group as a quality bonus. The proportion of the total NMS devoted to cost and quality changes over time: 70 percent is eligible as cost savings and 30 percent as quality bonus in program year one (PY1), 60 percent/40 percent in PY2, and 50 percent/50 percent in PY3. Once a payment amount has been determined, 25 percent is set aside until the end of the three-year demonstration program as a stop-loss to protect against increased expenditures in subsequent years of the demonstration.

Attributing Beneficiaries to PGPs

Beneficiaries are assigned to a medical group according to an algorithm that uses claims data to determine which beneficiaries received the preponderance of their E&M care from providers in the participating groups. If E&M charges for an individual beneficiary are the same for two or more group practices, then that beneficiary's charges for all Part B claims—not just E&M services—are used to make the assignment to a

single group practice. Claims data are similarly used to assign beneficiaries to a comparison group.

Quality Measurement and Quality Performance Payment Determination

A set consisting of 32 quality measures for six clinical conditions was selected. Some of the measures are constructed using claims data; others require chart review. The measures are phased in cumulatively over time, with diabetes measures (i.e., hemoglobin A1c testing and control, lipid measurement, foot exam) included in PY1; congestive heart failure (CHF) measures (i.e., weight measurement, beta blocker treatment, flu vaccine) and coronary artery disease measures (i.e., beta blockers after heart attack, antiplatelet treatment) added in PY2; and hypertension measures (i.e., blood pressure screening and control) and screening for colorectal and breast cancer measures added in PY3.

Beneficiaries falling into the numerator and denominator of measures based on claims are identified for the groups by the project's support contractor, Research Triangle Institute (RTI), based on the measure specification. Eligibility for a measure is determined using all of the beneficiary's claims, not just those for services provided by the PGP. The numerator for claims-based measures is calculated in similar fashion using claims data; but it can be augmented by other data sources, such as medical records on a random sample of eligible beneficiaries, if the group elects to do so.

The denominator for measures based on medical records is either a random sample of 411 beneficiaries who meet the relevant criteria or 100 percent of such beneficiaries. The numerator for measures based on medical records can be determined by (1) using medical record data only, abstracted using a data collection tool developed and managed by the Iowa Foundation for Medical Care (IFMC), or (2) pre-populating the tool using the group's own data systems (e.g., disease registry) and then supplementing this information with a medical record review. Each measure is weighted in computing the overall performance score: one point for measures that are medical record based; four points for measures that are primarily claims based. The total number of available quality points in a given measurement year is the sum of all weights for the measures designated for that year. Quality points for each measure are earned if the group meets any one of three targets: (1) the higher of 75 percent or the national Medicare HEDIS average performance (if applicable); (2) 10 percent or greater reduction in the gap between the group's baseline and 100 percent compliance; (3) achieving at least the 70th percentile Medicare HEDIS level. The ratio of quality points earned to total available quality points

determines the proportion of the bonus set aside for quality performance that the group will earn.

EARLY EXPERIENCES FROM THE PGP DEMONSTRATION

Discussions were held with six of the 10 PGP demonstration sites that ranged in size from approximately 300 to 1,000 physicians. The organizational structures vary, ranging from an integrated delivery system to a network model IPA. Two of the PGPs administer health plans that include physicians within and outside the medical group. All of the physician groups are affiliated with at least one hospital and often with multiple hospitals. The groups' experience with P4P prior to participation in the demonstration was varied but generally limited. Across the groups, there was almost a complete absence of capitation in their contracts with payers.

At the time of our discussions, CMS had assigned base year (January to December 2004) patient populations to each of the PGPs in order to calculate expenditure targets. Comparison groups of Medicare beneficiaries were also assigned for each community in which a demonstration site was located. Base year risk-adjusted per capita expenditures for the PGPs and their comparison populations had been calculated. The groups had completed medical record abstraction for the set of diabetes quality measures for their base year and were preparing to receive results on the quality measures from RTI sometime early in 2006. The following paragraphs highlight the experiences of participants during the first year of implementing the program.

PGP Rationale for Participation. Each of the groups indicated that the main reason for deciding to participate in the demonstration was the potential to be reimbursed for providing care the way they thought it should be provided to elderly patients with chronic illness. The groups also indicated that their medical management-level physicians believe that P4P will become the way of the future in health care, representing an important philosophical shift from encounter-based care and reimbursement to population health management, and thus hoped to gain early experience with P4P.

Strategies for Achieving Cost Savings and Improving Quality. All of the groups are using some form of case management as part of their strategy for the P4P program (e.g., involvement of nurses and an emphasis on continuity of care and reducing gaps between the care that chronically ill patients need and the care they receive). We observed that the strategies for and rationales behind achieving cost savings varied considerably among the groups. The general sense among all respondents is that the best way, and for several groups the only way, to achieve cost savings within the three-year

demonstration period is to reduce hospitalizations, specifically among CHF patients. Some groups expect case management to result in cost savings, in addition to improving quality. Other groups either are less optimistic that case management will achieve savings in the short term or are using other strategies to reduce expenditures while focusing case management primarily on quality improvement. While five of the six PGPs have rolled out their strategies for diabetic patients (the sole clinical focus of PY1), one group's case management strategy is based entirely on coordinating care for CHF patients, because that program is explicitly linked to its cost-savings strategy. All participants noted the evidence from the literature indicating that active case management for patients with diabetes, coronary artery disease, and hypertension may actually lead to cost increases in the short term, with savings not being realized until five to 10 years, which would be after the demonstration was completed. This led to some skepticism about the likelihood of receiving bonuses at all, particularly in light of the 2 percent minimum saving threshold, which was viewed by one PGP as "a huge hurdle." Two groups, however, have reason to believe that they already provide care more efficiently than the surrounding community does. To them, operating at their current level affords an opportunity to keep cost increases below their neighbors', thus yielding an automatic cost savings. They view the demonstration as a way to use health coaching to provide higher quality care for some of their most vulnerable patients, and they hope to recoup some of the start-up costs by operating efficiently across the board.

Anticipated Distribution of Incentive Payments. The PGPs varied as to how they planned to use any bonuses obtained through the demonstration. Two of the groups entered the demonstration with explicit strategies to return at least some portion of any bonus they receive to their physicians. Two other groups' strategies are to funnel any bonus payments back into program costs. The other groups have adopted more of a wait-and-see approach, deferring the decision in part because of their skepticism about receiving a bonus.

Information infrastructure. Most medical groups faced challenges identifying patients who are eligible for the quality measures (denominators) and tracking them to ensure they receive needed care (numerators). In some cases, the groups lack the necessary information; in other cases, the information is not easily available to physicians in an actionable format that they can use to improve the provision of care. Some of the smaller groups that have less robust internal data systems stated that having access to the claims data from RTI gave them new information about their patient populations that was quite useful. Getting the maximum benefit from such data, however, proved challenging

for some of the PGPs and required internal programmers and/or support from RTI at a level that may not be feasible for a program implemented on a larger scale.

SUMMARY OF INITIAL LEARNING EMERGING FROM PGP DEMONSTRATIONS

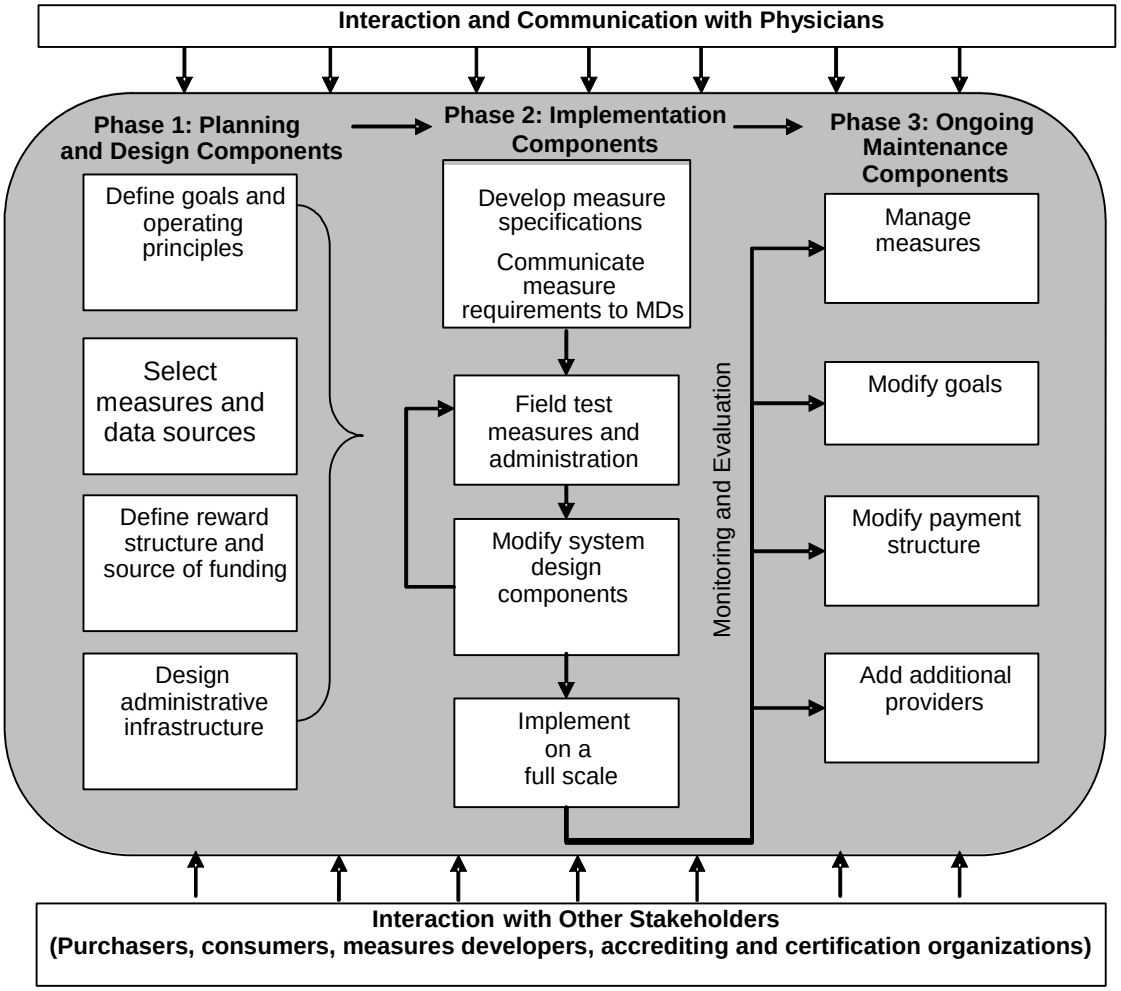
- **Participation was a key driver of performance improvement change in their organization**. Four of the groups mentioned that participation in the demonstration enabled them to finally implement changes, particularly to information systems, that had been discussed internally for years but never implemented. Once the demonstration was in place, changes started happening or happened much more quickly than before.
- **Capital investments are required to support measurement and quality improvement work.** Each of the respondents believed that the capital investments needed to improve infrastructure within medical groups to support P4P management and measurement will be "enormous," and that an influx of capital of this size will require close scrutiny of the vendors likely to fill the needs. Examples of infrastructure improvements include electronic health records, patient registries, and improved methods for data entry (Damberg and Raube, 2006). These capital improvements are one reason the ROIs for P4P programs are likely to be viewed very differently by the physicians and the program sponsors.
- **Improved sharing of ideas to promote better care for the population.** Several groups specifically mentioned their surprise that providers participating in the demonstration have embraced the concepts of population management underlying their case management strategies. The demonstration has improved the sharing of ideas for improving care within the organizations. The possibility exists to improve care coordination by creating an incentive for groups to seek information from community physicians about patients who receive much, but not all, of their care from the group.
- **Support of PGPs was a critical feature of program design.** Each respondent mentioned at least one instance in which the group communicated with either RTI or CMS on an individual basis in order to give input on measure specification, to appeal the inclusion of beneficiaries in the group's target population under the attribution algorithm, to question the inclusion of patients in the denominator of measures, or to request help managing data. These encounters highlight the importance of engaging with physicians and the value of giving providers a

mechanism for voicing concerns and providing input on the P4P program's design.

A FRAMEWORK FOR PAY-FOR-PERFORMANCE PROGRAM DEVELOPMENT

We constructed a framework that identifies core steps associated with P4P program development and operation (Figure 2). These steps were identified through our discussions with P4P sponsors. The framework can serve as a useful guide when thinking through the myriad steps involved in planning and operationalizing a P4P program. It also serves to organize our discussion of design components and options that follows in Chapter 4.

Figure 2
A Framework for Pay-for-Performance Program Development



We conceptualize P4P program development as having three major phases: (1) planning and design, (2) implementation, and (3) ongoing maintenance and improvement. At each phase, the interaction with physicians and other stakeholders (such as consumers, purchasers, accrediting and certification organizations, measures developers) will influence decisions about the structure and operation of the program. Based on our conversations with private-sector P4P programs, the timetable for moving through these stages tends to run from one to three years for the planning and design phase and approximately two years for the implementation phase in order to get the program up and running and stabilized. This timetable is likely to be longer with a program of the size that would be involved in implementing a P4P program nationally in Medicare physician services.

SUMMARY OF FINDINGS DRAWN FROM PAY-FOR-PERFORMANCE PROGRAMS

P4P programs are growing more prevalent in the private sector, and early experiments are developing in the public sector. Programs are being sponsored by individual health plans, coalitions of health plans, employer groups, Medicare, and Medicaid.

P4P sponsors stated that the two primary goals of P4P were to improve quality of care and reduce variation in resource utilization as a means to control health care costs. While the organizations we spoke with are firmly committed to P4P and believe that their programs—often in conjunction with other quality improvement activities—are resulting in better quality and more efficient care, few programs were being evaluated in a rigorous manner, and only a handful had attempted to compute the ROI from their efforts. Given the point reached so far in establishing the relationship between quality measures and cost-effective care, and the lack of knowledge about the size of the incentive needed to drive physician behavior change, it is unclear what the impact of current P4P programs will be on costs and quality. Some program evaluations are maturing to the point where results will be available within the next year to three years. However, given that all of these P4P programs are uncontrolled, natural experiments, P4P will continue to be just one of the many factors in play within various health care markets that could affect cost and quality.

Currently, there is no single strategy for designing and implementing a P4P program, so a great deal of experimentation and refinement is occurring as programs learn lessons along the implementation path. While all programs have key design

components—such as attribution rules, payout structures, and measures selection—very little is known about the best form for these components or the relative importance of different components for achieving the program's goals. In some cases, newer programs are adopting the design components of more-mature programs, but there is substantial variation across P4P programs in terms of their approach to designing their programs. Programs are generally customized to address specific characteristics of the local health care market (e.g., organization of physicians, existence of physician leaders in the community), and little attention is paid to what theory suggests might be the best options of various program components to adopt. At this stage, absent empirical evidence to support one design approach over another, the variation in P4P experiments will allow opportunities for testing various design strategies.

Although there is a lack of empirical data to demonstrate the impact of P4P in its current form, many of the existing P4P programs have learned valuable lessons about the process of implementation. CMS should seek to understand this process prior to embarking on a P4P program for Medicare physician services, since successful implementation is necessary for the program to have the desired effect. Successful implementation of a program is not, however, sufficient by itself for improving the quality and cost-efficiency of care. To bring about these improvements will require additional evidence on what constitutes a well-designed program.

CRITICAL PAY-FOR-PERFORMANCE LESSONS

1. **There was universal agreement that P4P alone is not a panacea for today's health care problems**. P4P needs to be implemented as part of a multi-pronged set of strategies designed to change physician behavior.
2. **Physician involvement and engagement in all steps of the process are necessary for successful implementation of a P4P program.** And beyond structured involvement, a mechanism is needed by which physicians can raise questions and provide input.
3. **Pilot testing of all aspects of program design and operation is critical.** Moreover, programs need to be open to making revisions based on what is learned during pilot testing. Trial-and-error is common in the creation and implementation of P4P programs.
4. **Starting small and demonstrating success helps to build trust among the physicians involved in the program.**

5. **There is a need for flexibility in design and potential customization**. Health care remains "local," and the variation in organization of physicians across geographic markets, as well as in the P4P programs already in play, suggests a need for flexibility in design and potential customizing of the program.
6. **A commitment to building and maintaining the operational infrastructure for P4P programs is necessary**. This commitment is needed to address such functions as data warehousing; data aggregation; programming and analysis; data auditing; processes for appeals and data correction; performance feedback; communication with and engagement and support of physicians; measure maintenance; and modification of data collection processes. Both monetary and personnel resources will be required.
7. **There must be alignment among the various program sponsors.** To reduce confusion and the burden placed on providers, it is important for program sponsors to be aligned on the set of measures used to report on and incentivize physicians.
8. **Providers need support to successfully participate in P4P programs**. This may take the form of patient registries, technical support, education, etc.
9. **Feedback information to physicians needs to be actionable.** Providing feedback in the form of rates alone does not assist physicians in improving the care delivered. Physicians benefit from information that can be acted upon—e.g., lists of patients who are in need of services to comply with measures.
10. **Continuous evaluation of the program operations and effects is essential.** It provides critical information for adjusting the program and creates increased physician trust and engagement in the program.

4. STRUCTURING A PAY-FOR-PERFORMANCE PROGRAM FOR MEDICARE PHYSICIAN SERVICES: DESIGN ISSUES AND OPTIONS

Our review of the literature and discussions with a broad cross-section of existing P4P programs in the private sector revealed a host of options for the design components that need to be addressed when developing a P4P program (see Figure 2, Chapter 3). This chapter explores a number of these key options for the design components, and draws from the experience of private-sector P4P programs to understand whether and how various options might apply to a Medicare physician P4P program. For each design component, there are frequently several options that could be pursued.

Choosing among the various options typically reflects considerations of whether the approach helps to achieve programmatic objectives and what consequences may occur as a result. However, as was underscored in our discussions with P4P program developers, P4P program development is largely experimental in many respects, and the impact of various design components has not been studied and is not well understood. Partly as a result of the lack of evidence about what does and does not work in P4P, the choice of design options tends to be influenced by a variety of factors, such as the P4P program goals, available funding, data constraints, expected consequences resulting from the choice, and stakeholder preferences. It is therefore important to clearly articulate these factors at the outset, and to be mindful of them when considering various design options. Where lessons have been learned over the past five years, we draw from them to help CMS understand the potential implications of various design options. For our analysis, we drew on discussions with P4P programs, discussions with our advisory panel members, relevant literature, and the experience of our research team.

The key design issues addressed in this chapter are as follows:

- How should the initial performance areas subject to P4P be identified?
- What role should physicians and other stakeholders play in developing a P4P program for Medicare?
- What are appropriate measures for a P4P program?
    - What types of measures should be included?
    - Should existing measures be used or new measures developed?
    - How many measures should be rewarded?
- What unit of accountability should CMS measure and reward?

- Should CMS pursue a national or a regional approach to implementation, given geographic variation in practice of care?
- How should patients be matched to individual physicians or group practices to ensure accuracy of measurement?
- What performance should be rewarded and how should rewards be structured?
    - What is the basis for reward?
    - How large should the incentive be?
- What should CMS be considering with regard to program infrastructure/implementation, including specifying measures, pilot testing, data collection and management, providing support to physicians, reporting and feedback, and monitoring?

DESIGN ISSUE: HOW SHOULD INITIAL PERFORMANCE AREAS SUBJECT TO PAY-FOR-PERFORMANCE BE IDENTIFIED?

Several factors will drive decisions about where initially to focus program measurement and incentive efforts:

- Goals selected by CMS for the Medicare physician P4P program.
- Prevalence of conditions in the Medicare population.
- Availability of existing performance measures upon which to draw.
- Availability of data to produce performance measures.
- Cost and burden associated with implementing the measures.

Provider performance is multi-dimensional in nature, and a P4P program may elect to work on multiple areas of performance based on the program sponsor's priorities. Examples of the types of goals that CMS could choose are

- Improve clinical quality (i.e., effectiveness of care as measured by process of care and outcomes).
- Reduce medical errors.
- Improve patient experience with care.
- Reduce costs or improve the cost-efficiency with which care is delivered.
- Stimulate investments in structural components or systems factors believed to be associated with improving quality and/or efficiency (e.g., IT capability, care management tools and processes).

The current impetus for P4P within the Medicare program is a need to reign in health spending and improve the quality of care received by Medicare beneficiaries.

Thus, in the near term, Medicare is likely to focus on addressing improvements in clinical quality and improving the cost-efficiency with which care is delivered, perhaps through a focus on quality measures thought to reduce costs. Because program goals affect a host of design choices, such as selection of performance domains and measures, they should be the starting point for P4P program development.

Appendix A includes a summary of design principles articulated by an array of policy and stakeholder groups, including MedPAC, purchaser groups, JCAHO, and physician organizations. A review of the design principles provides insights into the priorities of the various stakeholders that will either be at the table or seek to influence the design of a Medicare P4P program for physicians. These design principles can provide CMS with an understanding of the likely reactions from various stakeholders to decisions made during the design phase.

DESIGN ISSUE: HOW SHOULD PHYSICIANS AND OTHER STAKEHOLDERS BE INVOLVED IN DEVELOPING A MEDICARE PHYSICIAN P4P PROGRAM?

Involvement of Physicians

The importance of including physicians in the program design process was a key lesson cited by the P4P program sponsors in our discussions. This involvement was viewed as necessary to overcome provider resistance to being externally measured and held accountable, and to secure provider buy-in and engagement in the process of measurement and improvement. There is an absence of empirical research to show the differential impact associated with varying types and amounts of provider involvement. Physicians are interested in ensuring that the measures are evidence based and can be measured accurately, and that mechanisms exist to address data errors. In addition, physicians are interested in whether and how their results will be displayed or shared publicly. P4P programs with more-collaborative approaches indicate that they successfully worked through design and implementation challenges, that provider input frequently identified alternative approaches and increased provider awareness about the program, and that the stakeholder environment was less contentious. While programs have been able to achieve this on a local level, it has not always been an easy or straightforward process. Engaging with physicians on a national level will be an even greater challenge, particularly because CMS does not have established relationships with doctors in local markets.

Provider engagement is also critical during implementation. Our discussions with P4P programs revealed the important role played by physician leadership in medical groups and organized delivery systems in helping to explain the program requirements, engaging physicians in the performance feedback process, and working locally on quality improvement and building system supports (e.g., instituting patient registries, electronic data capture). One P4P program reported contracting with physicians to create local physician champions to perform these functions in an area dominated by solo and very small practices. This is a very real problem for Medicare on the FFS side, where there is an absence of organized medical groups with staff, resources, and infrastructure (e.g., data systems) to provide logistical support to individual physicians. We discuss this issue further in the implementation section.

Inclusion of Other Stakeholders

Stakeholders such as private purchasers and consumers generally are interested in the type and scope of measures selected and in ensuring access to comparative performance results. CMS will need to consider what other stakeholders need to be involved in the process of designing the Medicare physician P4P program, as this program will interface with the work of specialty societies and their recertification processes, with measures developers and accrediting bodies that will want to align their efforts, and with private purchasers and health plans that will be interested in having Medicare's policies be in synch with their own efforts to measure, reward, and promote transparency among the commercially insured population. One possible role for the various "outside" stakeholders is to serve in an advisory capacity, making recommendations and providing feedback on Medicare's overall design and possibly in vetting measures. Many of these external stakeholders have a substantial amount of experience with measurement, P4P, and public reporting, and could provide useful guidance and feedback during the development process.

DESIGN ISSUE: WHAT ARE APPROPRIATE MEASURES OF PERFORMANCE?

The measures that CMS selects for the Medicare physician services program should be based explicitly on programmatic goals, such as Medicare's desire to reduce or constrain growth in spending and to improve the quality of care for beneficiaries (i.e., clinical effectiveness, access to care, and coordination of care). Matching the measures to the program goals should be the first step in selecting appropriate measures of

performance. Measures selection should also be consistent with the criteria put forth by a number of national organizations—e.g., the Institute of Medicine (IOM), the National Quality Forum (NQF), the Agency for Healthcare Quality and Research (AHRQ), and the National Committee for Quality Assurance (NCQA) (see Appendix E)—including that a measure be (1) important (i.e., low performance and/or wide variation), (2) methodologically sound and evidence-based, (3) feasible to implement, and (4) have an impact (i.e. significant burden of disease in population). When choosing measures, CMS will need to consider several other issues, such as

- The specific type of measures to include.
- Whether to use existing measures or to develop new measures.
- How many performance measures to track and reward.
- Whether to seek alignment with existing performance measurement.

Below, we discuss each of the issues and describe the associated design choices.

What Types of Measures Should Be Included?

Alternative types of performance measures can be chosen for reward. Choices made among these alternatives will depend on program goals and feasibility/costs of data collection. Measure options include clinical effectiveness (process and outcomes), patient safety, patient experience, resource use, and systems/structural measures.

Option 1: Clinical Process-of-Care Measures

These measure whether the patient received a recommended test or service. Care processes typically are recorded either in the chart or in claims data for billing purposes and thus are easier to track than outcomes, which are not typically tracked over long periods of time. A key advantage of choosing process measures is that many of the measures can be produced from existing administrative data sources, making them relatively inexpensive to generate. Compared to outcome measures (discussed below), process-of-care measures are much more feasible to implement. Looking ahead, CMS should seek to promote the electronic capture of information for process measures that currently cannot be produced from administrative sources. CMS can do this by modifying the HCFA 1500 billing form to capture intermediate outcomes. Examples of additional information that could be captured are blood pressure and HbA1c levels.

Another significant advantage of focusing on process measures is that there are substantial numbers of existing evidence-based process-of-care measures that have been

well vetted and that physicians agree are important to provide. Nearly all existing P4P programs in the United States use some or all of the 40 HEDIS effectiveness-of-care measures, which are predominantly process measures (26 of the HEDIS measures are applicable to the Medicare-age population). Moreover, there are other measures sets available—such as the RAND QA Tools, the set that was used in the Community Quality Index Study (McGlynn et al., 2003) and is the basis of the UK P4P program; the ACOVE measures, a set designed for use with the vulnerable elderly, as well as measures available through a wide variety of vendors. These provide measurement on a wider array of clinical conditions and patient populations and could be readily applied by CMS.

Table 3 compares the measures included in these and other measure sets for a subset of clinical conditions relevant for the Medicare population. While many of these measures have been applied at the health plan and medical group level, there is less experience applying them at the physician level. Although there is limited published literature on how many events are required to produce a reliable estimate of performance when examining at the individual indicator level (see “The Problem of Small Numbers” discussion, below), work is currently being conducted on this issue, and it is being tested through the AQA pilots. The concern about producing reliable performance estimates is one reason the use of composite scores is attractive; it is easier with composites to have enough people in the denominator produce reliable estimates.

Further, the evidence upon which these measures are based varies in robustness, and most have not been designed specifically for an elderly population (the notable exceptions being ACOVE, PVRP, and most CMS measures). Thus, some existing measures may not be good predictors of outcomes in the elderly. An additional disadvantage of process measures is that they do not encourage the use of innovation to improve patient outcomes.

Table 3
Common Clinical Measures Relevant to the Medicare Population that are Included in National Measurement Sets

Clinical Condition	NCQA	AMA/ PCCI	PVRP	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
Depression/Behavioral Health										
• Follow-up after diagnosis and/or treatment	X				X	X	X	X		
• Medication during acute phase	X		X		X		X	X	X	X
• Medication during continuation phase	X									X
• Suicide risk assessment		X					X	X		
• Screening for alcohol misuse					X		X	X		
Bone Conditions										
• Osteoarthritis—OTC medications		X					X	X		X
• Osteoarthritis-Exercise recommended		X					X	X		
• Osteoarthritis-Functional and pain assessment		X					X	X		X
Diabetes										
• A1C Screen	X	X			X	X	X	X	X	X
• A1C Control	X	X	X		X	X	X	X	X	X
• Blood Pressure Control	X	X	X		X	X		X	X	X
• Lipid Screen	X	X			X	X	X	X	X	

Clinical Condition	NCQA	AMA/ PCCI	PVRP	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
• LDL Control	X	X	X		X	X		X	X	X
• Urine Protein Screening	X	X				X	X	X		X
• Eye Exam	X	X			X	X	X	X	X	X
• Foot Exam	X	X			X		X	X		X
• Smoking Status	X	X				X				X
• Smoking Cessation		X				X				
• Aspirin Use		X				X		X		
End Stage Renal Disease (ESRD)										
• Dialysis Dose			X	X						
• Hematocrit Level			X	X						
• Receipt of Autogenous AV Fistula			X	X						
Heart Disease										
• CAD: Antiplatelet therapy		X				X	X	X		X
• CAD: Drug Therapy for Lowering LDL Cholesterol		X		X			X		X	X
• CAD: Beta Blocker for prior AMI patients	X	X	X				X		X	
• CAD: Lipid Profile	X	X				X	X			X
• CAD: LDL Cholesterol Level	X			X				X		X
• CAD: Smoking Cessation		X					X	X		X
• COPD: Smoking Cessation						X	X			
• HF: LVF Testing		X		X			X	X	X	X
• HF: Weight Measurement		X			X		X	X		X

Clinical Condition	NCQA	AMA/ PCCI	PVRP	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
• HF: Assessment of Clinical Symptoms of Volume		X								X
• HF: Blood Pressure Measurement		X					X	X		
• HF: Examination of the Heart		X					X	X		
• HF: Patient Education		X					X	X		
• HF: Ace Inhibitor Therapy		X	X		X		X	X	X	X
• HF: Warfarin for Atrial Fibrillation		X				X	X			X
• HF: Assessment of Activity Level		X								X
• HF: Lab tests		X					X	X		
Hypertension										
• Blood Pressure Measurement		X					X	X		
• Blood Pressure Control	X	X			X	X				X
• Patient Education						X	X	X		
• Plan of care		X								X
Prevention, Immunization, Screening										
• Tobacco Use		X			X	X	X	X	X	X
• Tobacco Counseling	X	X			X	X	X	X	X	X
• Screen for Alcohol Misuse		X			X	X	X	X		
• Influenza Vaccination	X	X		X	X	X	X	X	X	X
• Pneumonia Immunization	X	X			X	X	X	X	X	X
• Breast Cancer Screening	X	X		X	X	X	X		X	

Clinical Condition	NCQA	AMA/ PCCI	PVRP	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
• Colorectal Cancer Screening	X	X			X	X	X	X	X	
• Cervical Cancer Screening	X				X	X	X		X	
• Drugs to be Avoided in the Elderly	X							X		
• Assessment of Elderly Patients for Falls			X					X		
• Discussion of Urinary Incontinence	X							X		X
• Urinary Incontinence Treatment	X							X		X
Urinary Tract Infection										
• Urine Culture						X	X			
• Treatment						X	X			

Option 2: Outcomes of Care

Outcomes of care supply another assessment of the clinical quality of care provided. Although there is a strong desire to measure outcomes of care, measurement of outcomes involves significant challenges. In considering whether to include outcome measures, CMS will face specific challenges:

- **The need to adjust for differences in the patient mix treated by each provider.** The patient populations cared for by each physician vary, and these underlying characteristics affect patient outcomes in addition to the care provided (e.g., comorbidities, complexity of illness; differences in patient factors shown to affect compliance, such as education, socioeconomic status, race, ethnicity, primary language spoken (DiMatteo, 2004; Bosworth et al., 2006; Kaplan et al., 2004; Sloan et al., 2004)). While there is recognition of the need to adjust for differences in patient populations, frequently the necessary data or validated methods for adjusting across different outcomes are absent. Adjustment is also a critical consideration because of the possibility that P4P will incentivize physicians to avoid patients who are more difficult to treat and who are expected to have a negative impact on them financially (Wessert and Musliner, 1992; Ellis and McGuire, 1996; Newhouse, 1989; Sorbero et al., 2003; Shen, 2003).
- **The long time frame needed to observe some outcomes.** While intermediate outcomes can often be observed within three to 12 months from time of treatment, long-term outcomes may not be observed for many years, making them difficult and expensive to track. Another significant challenge with long-term outcomes is determining how to attribute the final outcome to a specific provider's actions.
- **Many outcome events are rare**. This makes it difficult to observe these events and to detect statistically significant differences across providers
- **Problems with data systems**. Most data systems that contain information on intermediate or long-term outcomes (e.g., laboratory data, death records) are not linked to existing administrative systems from which performance can be measured. Consequently, substantial investments in resources may be required to link information.

A more reasonable alternative for CMS than measuring long-term outcomes would be to track intermediate outcome information, such as whether a patient with diabetes is in glycemic control (HbA1c < 7) or whether LDL (low density lipoprotein) levels are less than 130, since these are important markers of future morbidity and mortality among

patients with diabetes and heart disease. To do so, CMS would need to create a process that enables it to secure and integrate laboratory values from the multitude of laboratory vendors across the country, or would have to modify its existing coding systems to require physicians to code laboratory results on the HCFA 1500. Steps toward developing Current Procedure Terminology (CPT) supplemental codes to capture some intermediate outcomes are under way.

Option 3: Assessment of Patient Experience

Assessments of patient experience, which are typically based on survey information, assess performance associated with such issues as patient access to routine and urgent care, coordination of care across providers, doctor-patient communication, chronic care management and support, and counseling on or provision of preventive care (e.g., smoking cessation, diet, exercise, flu shots). These measures capture aspects of care that are important to consumers and represent a different domain of performance. The national Consumer Assessment of Healthcare Providers and Systems (CAHPS) consortium has developed the Ambulatory CAHPS survey (A-CAHPS) for use at the physician level. The availability of a national survey tool provides an existing standardized survey method that could be readily applied by Medicare within a physician P4P program. Experience with the A-CAHPS tool shows that nearly all physicians could be included in an assessment of a patient's experience, given that reliable estimates of performance can be produced with 30 completes from an initial sample of 100 patients (Safran et al., 2006). While the inclusion of patient experience measures is advantageous in that all physicians could be measured and patient-centered care is a key dimension of care, these measures could not be produced without new data being collected at the direct expense of the physician or CMS. At an estimated cost of $200 per physician (which potentially could be borne directly by each physician, or Medicare could choose to underwrite costs for high performing physicians), the cost of surveying patients for the 500,000 physicians that provide Medicare Part B services would be $100 million.

Option 4: Measures of Cost-Efficiency or Resource Utilization

The key arguments for including these types of measures in P4P programs are the continued upward pressure on health care costs and variations across providers in amounts of services used to treat patients with similar conditions. CMS, within its PGP P4P demonstration, requires demonstrated cost-efficiency as a prerequisite to being

eligible for quality-based incentives; the savings generated provide a means for financing the incentive payments to medical groups.

There are several options for measuring resource utilization, ranging from simple utilization measures to more-complex episode-based total cost-of-care measures (i.e., cost-efficiency measures). Simple measures, such as use of generic drugs by Medicare beneficiaries, will be impossible for CMS to track, because the design of the Medicare Part D benefit does not enable CMS to capture pharmacy data for Medicare beneficiaries for whom Medicare is the secondary payer. Longitudinal cost-efficiency measures focus on creating episodes of care and comparing physicians to their peer group in terms of all resources—ambulatory (including pharmacy) and inpatient—used to treat similar patient episodes. A number of commercial vendors have developed proprietary longitudinal efficiency measures that are being used in some P4P programs; these tools would be expensive for CMS to run, and their proprietary nature would make it difficult for CMS to have a transparent tool.

Another challenge that CMS will face in using currently available cost-efficiency tools is that the scoring output (e.g., performance relative to peers) is not very actionable. Should CMS decide to include cost-efficiency measures, physicians would need assistance to understand what actions are necessary—i.e., what they need to change to improve their scores, or, more specifically, what drivers of cost distinguish a cost-efficient provider from a cost-inefficient provider.

Option 5: Structural Measures

Structural measures such as IT capabilities and degree of systemness (e.g., NCQA's Physician Practice Connections assessment tool) represent another potential area of measurement and reward (MedPAC, 2005b). Rewarding IT and other measures of system support may serve to stimulate faster improvement in quality and create a business case for investment in systems. However, because most performance measures that CMS might choose to use rely on administrative data sources, it can be argued that a de facto business case already exists for IT investment, since to do well on the P4P performance measures requires that physicians, practice sites, and medical groups have tools to ensure good electronic data capture. Inclusion of IT as a distinct performance measure could be viewed as duplicative. Another difficulty with using these measures is that they would be burdensome for doctors to self-report and a challenge for CMS to aggregate and analyze across 500,000 physicians.

Should CMS Use Existing Measures or Develop New Measures?

It will be faster, easier, and cheaper for CMS to use existing performance measures rather than spending time and resources to develop and test new measures. Using existing measures will also facilitate CMS's task of aligning its measurement efforts with those of other payers in the market. However, of the current, publicly available performance measures from which CMS could draw, only a small number apply to specialty care. In the near term, CMS will face limitations in the number of physicians it can score. This deficiency in the number of measures available for evaluating the performance of specialists is receiving greater attention, particularly from specialty societies and the AMA. In December 2005, the AMA reached an agreement with congressional leaders to develop performance measures for 34 clinical areas by the end of 2006 and performance measures to cover the majority of Medicare spending for physician services by the end of 2007 (Pear, 2006).

How Many Measures Should be Rewarded?

CMS will have the option of including many or including few measures when structuring a P4P program for Medicare physician services. The rationale for and advantages and disadvantages of these two options are as follows:

Option 1: Comprehensive Assessment

The desire to comprehensively measure the care provided by a physician and produce a better representation of that physician's overall quality of performance is the basis of arguments for the use of many measures. High performers in one clinical area frequently are not high performers in other clinical areas (Gandhi et al., 2002; Parkerton et al., 2003), so a narrow set of measures is unlikely to produce a clear signal of the overall quality of care delivered by a provider. A broad set of measures also represents the variety of patients and conditions in a practice. Using a large set of measures may avoid the narrow focus of "teaching to the test" and may reduce the likelihood of gaming by providers. A more comprehensive set of measures also focuses progress in performance improvement at the systems level (i.e., putting systems in place to routinely monitor all patients with chronic conditions) on many fronts, and has the potential to more quickly close the quality chasm than taking incremental steps of one measure at a time would. The counter argument is that with many measures, the effort on each is small, so progress is similarly small. The near-term challenge with pursuing a strategy of many measures is how to obtain the data to populate all of the measures, since current

administrative data systems do not capture the data needed to construct a large number of measures.

Option 2: Narrow Assessment

Physicians strongly support the use of a small number of measures because they prefer to focus their attention and quality improvement resources on only a few areas at a time. Physicians, particularly those in solo and small group practices, are faced with many challenges when required to implement wide-scale quality improvement (e.g., system re-engineering) approaches to address an array of performance deficiencies. They are more able to carrying out the traditional approach of addressing one problem at a time. Physicians cite problems with existing administrative data systems that are not designed to produce data capable of being used to broadly measure what a physician does. However, if CMS and other payers were to require a large, broad set of performance measures, physicians would be presented with an important stimulus for changing their data systems to support wide-scale quality management and improvement. Making the necessary IT system changes to support detailed performance measurement will take time and money, and will be especially challenging for physicians working in solo and small group practices who lack the financial and staff resources to upgrade their IT capabilities. A program that creates large burdens on physicians could prompt some physicians to retire or influence new physicians' choices of specialty.

Should CMS Align Its Measurement with Existing P4P Efforts?

Across the nation, many physicians are already exposed to performance measurement and P4P programs. When deciding among the performance measure options, CMS should consider whether and how it could align its physician measurement with existing measurement activities. The benefit of alignment is that it reduces not only the data collection and reporting burden, but also the confusion among providers about where they should focus their efforts. Alignment refers both to the choice of measure and to the specifications used; in the absence of alignment, physicians face the challenges associated with having to track and report data in multiple ways to satisfy different measuring entities.

DESIGN ISSUE: WHAT UNIT OF ACCOUNTABILITY SHOULD CMS MEASURE?

The decision about which unit of accountability—individual physician, practice site, or medical group—to measure and reward is not merely an element of the technical design. This decision will be influenced by the program goals CMS defines for it Medicare physician services P4P program. For example, if a key goal of the program is to increase coordination of care for Medicare beneficiaries, then focusing on a practice site or medical group rather than on individual physicians may help promote coordination across providers within a practice site or group. If the goal is to provide information to Medicare beneficiaries to enable them to choose high quality doctors, however, then measurement at the level of the individual physician might be more appropriate, since this is the level of choice for consumers.

Practice Sites

If CMS decides to focus measurement at the practice site level, rather than the individual physician level, CMS will face operational challenges given that currently there is no existing map of the 500,000 individual Medicare Part B physicians to practice sites. CMS would need to create the initial map and institute a process for updating it frequently. Private-sector P4P programs report that they have spent considerable resources identifying practice sites and mapping their physicians to practice sites, and note that this process requires considerable knowledge of the local health care market.

Choice of the unit of measurement will also be influenced by statistical considerations associated with the performance measures, primarily the problem of small numbers. We discuss this problem next, as well as the proposed options for dealing with it.

The Problem of Small Numbers

A typical primary care physician's panel size is 1,500 to 2,000 patients, and the panel comprises multiple payers. No single payer today—including Medicare—can use 100 percent of a physician's patient encounters in creating performance scores without pooling data with other payers. While Medicare may represent 40 percent of a physician's practice, even this level of penetration may not result in an adequate number of patient events for some quality indicators (if relying on Medicare data alone). The small numbers of certain types of clinical encounters within any given physician's practice raise concerns about the ability to reliably measure the performance of individual

physicians, particularly when evaluating individual indicators of clinical performance (as opposed to composite measures of performance).

To illustrate the small numbers problem, let's assume that the prevalence of an event is one percent in the population. A health plan with 500,000 enrollees would yield 5,000 patients eligible for scoring at the plan level. In contrast, a primary care physician with a patient panel of 2,000 would yield only 20 patients eligible for scoring on this measure, which is not enough to reach the denominator (oftentimes 30 minimum) that most clinical quality measures require to produce a stable estimate of performance. However, physicians who have a high concentration of Medicare beneficiaries, may be able to score at the physician level on frequently occurring events such as breast cancer screening, immunizations, and control for high blood pressure.

Three options that can be used to address the problem of small numbers are

1. Pooling data across multiple data sources (i.e., Medicare, Medicaid, commercial).
2. Pooling data over time.
3. Aggregating over multiple measures to create composite scoring.

Option 1: Pooling Across Multiple Data Sources

A number of conditions cut across both the commercial and Medicare population (e.g., the treatment of hypertension, heart failure, coronary artery disease [CAD], and diabetes). For these conditions, pooling of commercial and Medicare data may enable measurement at the physician level. The advantages of pooling data across organizations include

- A more representative view of a physician's performance, making scores more credible to physicians.
- The ability to rate more individual physicians because there are more patients to populate the numerator and denominator of the performance measure.
- The ability to score a provider on a larger number of performance measures because of adequate sample sizes, which in turn increases the face validity of the program.
- An increase in the reliability of performance estimates.
- The ability to provide individual physicians with a more integrated review of their performance across payers.

Option 2: Pooling Across Time

An option that is fairly easy to carry out is pooling data over time to yield more events that can be scored. Compared to pooling across multiple payers, this option has the advantage of rapidly amassing sufficient data to score a physician but may suffer from the problem of changes in clinical knowledge over time. Changes in clinical knowledge may cause measures to change from year to year, which could affect a multi-year measurement strategy. Pooling data over time also limits the ability to observe change, because old performance continues to be included in the score for a period of time. One method to address this limitation is to weight recent data more heavily when creating performance scores.

Option 3: Aggregating Over Multiple Measures

Composite scoring is another way to aggregate information in an effort to produce reliable performance scores at small units of measurement. Composites could be constructed by type of care (i.e., preventive, acute, and chronic) or across a particular condition (e.g., diabetes) where multiple performance measures may exist. There are a number of issues about composite construction that are not well understood and would require further testing, however. These include what measures to combine (i.e., are individual measures related to one another in a way that will yield a clear signal of quality?), whether to weight each individual measure included in the composite equally or differentially, and what the effects of different decisions around grouping and weighting of measures would be.

Should CMS choose to measure performance at the individual physician level, it will need to assess how many measures can be reliably produced at the individual physician level and whether there is a sufficient number of measures to evaluate a physician's performance. This effort will be aided by near-term work that is happening through a partnership between CMS, AHRQ, and the Ambulatory Healthcare Quality Alliance (AQA). On March 1, 2006, the partners announced that they will fund six pilot demonstration projects to test the pooling of commercial and Medicare data to produce physician level quality and efficiency performance scores (www.ahrq.gov, 2006). The pilot projects are expected to yield important information about whether pooling data to produce physician level performance scores increases the number of physicians that can be scored and the number of measures on which they can be scored reliably. They will also address the implementation issues associated with pooling data across public and private payers. Additionally, the AQA pilot projects will test various strategies for

composite construction and are expected to provide CMS and other sponsors with useful information that can serve to better guide design choices in this area in the next year.

DESIGN ISSUE: HOW SHOULD PATIENTS BE ATTRIBUTED TO PHYSICIANS FOR ACCOUNTABILITY

All P4P programs have defined a method for attributing patients (and their care) to the provider that is the unit of analysis (i.e., physician, practice site, or medical group). This will be especially challenging for CMS in a FFS delivery model, where no single physician is accountable for all care that a Medicare beneficiary receives. Instead, many physicians may "touch" a patient, and no single physician owns the patient and all that happens to that patient.

Attribution rules can be structured in some cases to help achieve the goals of the program, such as to create incentives to coordinate care, a sense of responsibility for the patient, and physician buy-in to the P4P program. Different attribution methods may lead to different results and have different implications for the P4P program. However, there is a lack of empirical evidence about the differential effect of specific attribution rules. We describe here two general options for attributing patients to physicians and highlight some of the likely effects.

Option 1: Attributing Patients to Physicians Based on Volume of Services or Costs Incurred

Within this option, there are four variations in how cost or volume rules could be constructed:

- Using highest volume of evaluation and management (E&M) services.
- Using greatest share of patient costs, either overall or on E&M services.
- Using highest volume of preventive care (in programs with a primary care focus).
- Assigning shared responsibility among physicians with at least a certain volume of services.

On the surface, attribution of care on the basis of dollars and attribution on the basis of volume of visits seem similar, but the results of these two differ. Attribution based on dollars results in more patients having primary accountability assigned to sub-specialists because costs are higher, whereas assignment based on volume of services shifts more patients to primary care specialty physicians who typically provide a greater number of services. The specific services and E&M codes included in the determination will also affect assignments of patients to providers. To attribute patients to providers

with whom they have an ongoing relationship, the CMS PGP demonstration excludes specific E&M codes for emergency department visits and consultations.

Option 2: Attributing Patients Based on Having Touched Them

Again, there are variations in how this option could be constructed:

- Based on all physicians of relevant specialty by measure who "touch" the patient.
- Based on all physicians who "touch" the patient regardless of specialty.

These two approaches differ from the previous method in that there is shared performance responsibility, since multiple providers are accountable for patient care. This may maximize the probability that a patient receives the recommended care, since all relevant physicians are responsible for ensuring a patient gets recommended care. However, depending on how it is constructed, this method runs the risk of double paying on any given performance measure—that is, paying a physician who did not provide the service but who was nonetheless assigned responsibility. The effects of this approach are not well understood; it is not known whether it increases or decreases responsibility for care delivery.

Another important consideration is the attribution method's face validity to providers. Attributing patients to physicians who believe they are not primarily responsible for their care may reduce provider acceptance of the program.

CMS has experimented with different attribution models in the development of its P4P demonstrations. A plurality of care rule, which is very similar to that used in the CMS PGP demonstration, results in the assignment of over 50 percent of patients who receive care from a particular group to that group practice. Interviews with providers indicated they felt that over 70 percent of the patients assigned to them were definitely their patients and that approximately 10 percent might be their patients; they felt no ownership for less than 20 percent of the patients assigned to them (Pope, 2002). Regardless of the approach adopted by CMS, assignment will not be correct 100 percent of the time.

The factors that should guide CMS in selecting an attribution methodology are that it (1) has face validity and is perceived as fair by physicians, (2) creates a sense of responsibility for the delivery of care, and (3) is operationally feasible to implement and manage (e.g., assignment can be made automatically, based on existing claims data).

CMS should work with the AQA pilot projects to model the effects of different attribution rules using existing Medicare Part B claims data and should assess the impact of different methodologies. As is occurring in the CMS PGP demonstration, a Medicare

physician P4P program should assign physician responsibility for patients at the start, prior to measurement, to allow physicians an opportunity to review the set of patients assigned to them and identify any critical errors in the assignment. For example, attribution could be performed based on data for the year prior to measurement, allowing corrections for patients who have changed physicians.

DESIGN ISSUE: HOW SHOULD THE FINANCIAL REWARD BE STRUCTURED?

The financial incentive can be constructed as

- A bonus.
- An adjustment to the fee schedule (for all services or selected services).
- Shared savings.

Option 1: Bonuses

Lump sum bonuses could be administered annually based on a physician's performance during the prior year. Bonuses have the advantage of being explicit and, if sufficiently large, may be effective in capturing a physician's attention and engagement in the program. The total amount allocated toward the bonus pool can also be set in advance—say through withholding a share of any increase in the physician fee schedule. Total payouts stay within the total amount set aside, providing certainty in budgeting. This approach, compared to the alternatives, is also relatively easy to administer. Because of the magnitude of data processing involved with running a P4P program nationally, anything more frequent than an annual calculation of scores and payouts would be administratively difficult for CMS to implement.

Option 2: Adjustments to Fee Schedules

Adjustments to fee schedules are prospective (meaning payouts occur moving forward, when services are rendered), based on performance in the prior year. A relatively small increase on E&M codes may be less likely to capture providers' attention than a lump-sum bonus, even if the magnitude of the incentive for a year is the same dollar amount. A disadvantage of this method is that in a FFS environment, it incentivizes increased use of services to maximize payment, as the increase in fee schedule typically applies to all services for which the provider bills. To minimize this problem, CMS could modify this approach and only adjust the fee schedule for specific services it wants performed and is measuring (e.g., flu shots, mammograms). A substantial drawback of

this approach for CMS is that CMS would not know in advance what the total liability would be and thus how to establish a "fixed" budget for incentive payments. The total amount paid in any year would vary based on physician compliance with the measures.

Option 3: Shared Savings

The sharing of savings generated by the P4P program between the program sponsor and providers and mitigates any P4P program's negative financial impact on physicians. For example, efforts to reduce misuse or overuse of services and the costs of improving information systems to improve quality of care may reduce net earnings for some providers to a greater extent than would be balanced by what they receive through the financial incentives. Shared savings reduces the sense providers have that P4P programs financially benefit others (health plans, employers, taxpayers) at their expense.

DESIGN ISSUE: WHAT SHOULD THE BASIS FOR THE REWARD BE?

There are four core options available to CMS to serve as the basis (used alone or in combination) for determining whether a physician is eligible to receive an incentive payment. The determination is made based on

1. Absolute threshold of performance or performance target (e.g., 92 percent of diabetic patients have an annual foot exam).
2. Relative threshold of performance (e.g., above the 75th percentile of all providers' or peer groups, scores).
3. Improvement in performance over time (e.g., 10 percent improvement in score from previous year).
4. Action or service provided (e.g., additional payment for every time a service, such as a mammogram, is provided to relevant patient population).

Option 1: Absolute Performance Threshold

This approach would require CMS to determine for each performance measure what level of performance would result in payment. The steps would entail (1) identifying a standard target level of performance—such as 100 percent mammography rates; (2) determining a reasonable degree of deviation from the standard that is acceptable (i.e., to account for variations in practice and allow for patient refusal or those not visiting the physician to mitigate incentives to unenroll non-adherent patients); 3) setting an appropriate target for incentive recipients to attempt to reach (e.g., 90 percent mammography rate).

A key strength of absolute performance thresholds is that they provide clear messages to physicians about what level of performance must be achieved to receive an incentive. They also help physicians understand the likelihood of their getting an incentive. A drawback of absolute thresholds is that in local areas where performance is low, no physicians will receive incentives, and physicians may remain disengaged if they view the target as unattainable. Another potential problem is that when incentives go toward reaching a common, fixed performance target, little benefit is achieved, since the incentive only motivates physicians to meet the target and not go beyond it (Dudley et al., 2004; Rosenthal et al., 2005). This could be addressed by having a sliding scale of absolute targets or by adjusting targets upward on a periodic basis.

Use of absolute performance targets can make budgeting for the incentive more challenging, because CMS will not know when it sets the budget how many providers will reach the threshold and thus have to be paid. This problem can be addressed by setting a fixed budget for each measure. Then the actual payout to each individual provider will depend on the number of providers who meet the threshold, such that the dollar amount paid for a measure will vary in relation to the number of providers meeting the threshold. This will, however, create uncertainty among physicians about the award amount they could expect to see if they hit the performance target.

Option 2: Relative Thresholds

An example of a relative threshold is pegging the reward to the 75th percentile performance rate of all providers nationally or regionally. Relative thresholds have two advantages: (1) the absolute level of the performance threshold continues to increase as overall performance by providers improves, and (2) it is easier to budget for this type of threshold than for an absolute threshold, since the number of providers receiving the incentive is known a priori (e.g., if the reward is pegged to the 75th percentile, then those in the 75th percentile or higher will be paid; so the number of physicians that will fall in the top quartile of performance can be calculated). The disadvantages with relative thresholds are that (1) providers will not know in advance how well they must perform to receive the incentive, so it increases a physician's uncertainty about the likelihood of receiving a reward, (2) the absolute value of the threshold will be arbitrary and may not reflect superior performance (e.g., if performance scores are uniformly low on a particular measure, the 90th percentile performance might be 40 percent), and (3) the differences in performance between providers who receive and do not receive the incentive may be very small.

Both absolute and relative targets result in rewarding the physicians who were already providing high quality of care prior to the start of the program (Dudley et al., 2004; Rosenthal et al., 2005). In consequence, poorer performers, not expecting to see a reward, may be discouraged from investing in quality improvement. Similar to what has been done by some private-sector P4P programs, CMS could work to mitigate this problem by providing bonuses for top performers and making quality improvement grants to poor performers to help move them into the competitive range.

Option 3: Improvement in Performance

This option would pay all physicians who achieve improvements year over year or relative to a baseline measure, such as occurs in the CMS Premier hospital P4P demonstration. Hospitals in the Premier demonstration must improve over their year-one performance or face the prospect of lower payments. This strategy benefits physicians whose performance is very low by allowing them to reap some of the rewards; as a consequence, it may engage a larger number of physicians in the program. This approach penalizes, or may at least discourage, providers who are already doing well and thus have less room to improve (Rosenthal et al., 2005). One challenge to implementing this option is that it is not unusual for performance measurement specifications to change year to year; for measures for which this occurs, CMS would be precluded from measuring improvements over time.

Option 4: Payment for Provision of a Service

This method is essentially enhanced FFS on a specific service basis. For example, rather than measuring mammography rates, the reimbursement for a mammogram would be increased. The advantages of this approach are that it allows providers to know what must be done to obtain the incentive and removes the necessity for providers to have adequate numbers of patients to create rates for individual measures or composite measures.

DESIGN ISSUE: HOW LARGE SHOULD THE INCENTIVE BE?

There is much debate, and very limited empirical data, on how much of a financial incentive is required to secure behavior change among physicians. There is evidence that small incentives were effective in situations where the cost of complying with the program requirements was low and/or there were few other income opportunities (Dudley, 2004; Greene et al., 2004). For situations in which the cost of complying with

program requirements is high for physicians (e.g., having to create behavior changes in patients to affect clinical outcomes, such as smoking cessation or weight loss), even quite sizable incentives may not be enough to induce behavior change.

While large incentives may in some cases create sufficient motivation for physicians to engage and comply, large incentives can also incentivize physicians to "game" data in order to win. There is also a potential concern that large incentives tied to a small number of performance areas may result in unintended consequences, such as physicians heavily weighting their efforts to the areas being measured and ignoring other areas of performance that are not being rewarded. Dudley (2004) also hypothesizes that the level of certainty associated with a physician's securing a bonus may impact whether a physician responds.

Among existing private-sector P4P programs that include physician level incentives, the average annual incentive payout at the physician level varies, currently ranging from $1,000 to $15,000. If we assume that a primary care physician earns approximately $160,000 annually, this amount represents approximately less than one percent to 9 percent of compensation. Whether this amount of money garners a physician's attention depends on the physician's specialty and average annual income within specialty, as well as what compliance with the program requirements entails. Incentive payments may get the attention of some physicians when the amount of additional income reflects an important percentage of their total income, but other physicians may view the amounts as insignificant in relation to existing earning opportunities and thus not worth the opportunity costs.

CMS announced in October 2005 (www.cms.gov, 2005), as part of its quality improvement efforts, the Physician Voluntary Reporting Program (PVRP). Under the PVRP, physicians can voluntarily report to CMS a set of 16 performance measures that will be used by the Quality Improvement Organizations (QIOs) in CMS's 8th Scope of Work, which guides the activities of the QIOs. Physicians will start reporting in January 2006 and will receive confidential reports on their performance. The PVRP program design is very similar to the starting point of the HQA (formerly NVHRI) initiative, and these initial efforts may serve as a stepping-stone to widespread reporting of data and transparency of performance results, and potentially to P4P. However, absent a financial inducement, the number of physicians who step forward to voluntarily report will be small. Introduction of a financial inducement, in the form of a 0.4 percent increase in hospital payments (MMA, 2003), caused voluntary reporting in the HQA program to go from approximately 400 hospitals submitting data to more than 4,000.

Current performance-incentive payout amounts in the United States stand in stark contrast to those of the new UK P4P program, where up to 25 percent of a general practitioner's (GP's) income is at risk for performance on 136 clinical, organizational, and patient experience indicators (Roland, 2005). The UK P4P program, which targets measurement at the practice level rather than at individual physicians, has been in operation since the start of 2005, so it is still too early to understand the program's effects. Early results from Scotland indicate that more than 50 percent of practices are receiving the full number of points available through the program, which has raised concerns that providers may be gaming the system through excessive use of reporting exceptions for the clinical indicators shown in Table 4 (Roland, 2005). A big difference between the UK P4P program and U.S. P4P programs is that the incentive payments largely represent new additional funding for GPs (on top of a 30 percent pay increase just prior to the start of the program), in contrast to a redistribution of current funding, and performance is based on physician self-report information.

Table 4
Reporting Exceptions in United Kingdom's Family Practice P4P Program

• Patient refused / not attended despite three reminders • Not appropriate, e.g. supervening clinical condition, extreme frailty, adverse reaction to medication, contraindication, etc. • Newly diagnosed or recently registered • Already on maximum tolerated doses of medication • Investigative service is unavailable

DESIGN ISSUE: SHOULD A NATIONAL OR A REGIONAL APPROACH BE PURSUED?

A key strategic decision for CMS to consider is whether to pursue a one-size-fits-all approach, meaning all physicians nationwide would be required to adhere to a uniform program with a common set of performance measures and payout structure, or to take a regional approach, tailored to the structure of the physician market and quality problems of a defined geographic area.

Option 1: National Approach

A uniform set of program requirements, including measures, would be easier for CMS to administer from a communications and management perspective (e.g., single

programming requirements, single analysis and reporting strategy). This approach also creates a level playing field for providers nationwide, which may be perceived as a fairness issue. A potential disadvantage of this approach is that physicians in some regions may already be high performers on the selected measures (whether as a result of prior exposure to P4P on similar measures or for other reasons) and, depending on the program structure, the incentive dollars would then be rewarding existing high performance rather than incentivizing improvements in regions with lower performance. Regional variation in performance may result in rewards being concentrated in certain areas.

Option 2: Regional Approach

A regional approach to a P4P program would allow tailoring of the program, meaning different geographic regions could structure measurement based on the organization of physicians in that region (i.e., they could vary the unit of analysis). Or, the different regions could focus on different performance measures and/or be subject to different thresholds for determining payout. CMS could identify a large set of candidate performance measures that, over an extended period of time, it wants all physicians to address. In the near term, however, each region could be allowed to pick measures from the predetermined menu of measures. This approach allows physicians some local decisionmaking in terms of the areas they see as priorities for measurement and improvement, and it might facilitate alignment with existing P4P across the country. Physicians may want to choose measures they are already dedicating resources to, creating synergies across measurement and improvement efforts under way at all levels of the system.

A potential downside of regional selection of measures is that physicians are likely to choose areas in which they are already performing well. In that case, the incentive dollars will be targeted not at improving major deficiencies in care delivery, but at rewarding existing high performance. This problem could be addressed in two ways: by setting thresholds for performance and then, if a region's performance were already above one of those thresholds, selecting another area on which it should focus; or by having Medicare select measures for a region based on areas of poor performance. A tailored approach would exponentially increase the CMS management burden, since CMS would need to keep track of different sets of measures for hundreds of thousands of physicians. This approach may also be perceived as unfair, because not all physicians would be adhering to an identical program.

DESIGN ISSUE: HOW SHOULD PROGRAM INFRASTRUCTURE BE ADDRESSED?

Operating a P4P program requires an ongoing administrative infrastructure to perform a range of functions on an annual basis (and sometimes more frequently). Examples of these functions are

- Communication with providers.
- Soliciting stakeholder input.
- Developing measures and/or refining measures specifications.
- Organizing public comment/feedback on measures specifications.
- Issuing measures specifications.
- Pilot testing measures.
- Measures maintenance.
- Programming measurement specifications.
- Data collection and data warehousing.
- Data analysis.
- Auditing.
- Feedback to physicians.
- Grievances/appeals process.
- Calculating and issuing performance payouts.
- Preparing internal or public scorecards.

These functions require dedicated staff and resources to sustain the operation of a P4P program and are an ongoing expense to any P4P program. The scale of operation of a Medicare physician P4P program is several orders of magnitude larger than any existing P4P program's operational infrastructure in the nation, particularly with respect to data collection and warehousing, data analysis, and performance feedback. Existing P4P programs in the United States focus on small communities (e.g., Buffalo, Louisville, Rochester) of providers or, in some cases, operate statewide (e.g., IHA program in California). In contrast, CMS would be operating a program in 50 states among the more than 484,000 physicians who provide services to Medicare beneficiaries.

Some of the functions outlined above can and should be conducted centrally (e.g., measures maintenance, preparing scorecards, soliciting input on measures). Others may be more effectively organized and conducted locally; for example,

- Communicating with and providing support to providers.
- Data collection and aggregation.
- Data warehousing.

- Analysis.
- Grievances and appeals.
- Auditing functions.

Thus, an important step in the design process is defining which functions need to occur centrally and which can be better performed locally. For example, given the varied organization of physician practices and the presence or absence of physician leaders across health care markets in the United States, having a locally knowledgeable party to engage, support, and communicate with physicians may be a critical design feature. The data processing task, when scaled up to 484,000 physicians, is an extremely large task and may be best done regionally. CMS currently contracts with Part B carriers to process Part B claims data, pay the bills, and interface with providers locally in each of the 50 states. Part B carriers, with their existing infrastructure and relationships with local providers, may have the potential to provide selected components of the operational infrastructure, including such "provider relations" functions as communication on measures specifications and timetables for data submission, and processing the data and making payouts to physicians. The CMS QIOs, which operate in most states, do not currently have existing infrastructure. They also have varying experience in working with Medicare ambulatory care data and, with the exception of the Iowa Foundation for Medical Care (IFMC), do not serve as data warehouses. QIOs may be best positioned to provide support for physician communication and education about the measures and program requirements, as well as to perform audit functions. Current regulations preclude the QIOs from sharing data with CMS, so without a change in the law, the QIOs could not serve as a local repository for the data.

Regardless of where the functions are carried out, CMS does not currently possess the capability to perform them and would thus be required to "build out" these functions to support a nationwide P4P program, unless a simple "pay for behavior" approach was taken. The administrative infrastructure of a P4P program is resource intensive and will take CMS time to build. CMS is operating demonstration projects that are starting to yield important information about the infrastructure requirements necessary to run a P4P program. However, these are small efforts compared with what would be required to support a nationwide rollout and maintenance of a P4P program.

Measures Specifications

CMS will need to define the exact specifications (i.e., numerator, denominator, and any exclusion criteria) for producing each measure included in the physician P4P

program. For some measures included in the program, specifications may already exist through the measure development activities described in Appendix C. However, some measures may have multiple definitions, and CMS will need to decide which specification it will adopt. For example, many existing measures sets include HbA1c control in diabetics (e.g., NCQA, AMA/PCCI, VHA, ICSI, RAND QA, ACOVE, AQA, NQF Approved measures), yet the specifications of the individual measures across these measure sources vary (Appendix D). Performance in very similar domains can vary based on the measures and specifications chosen (McGlynn et al., 1997).

The selection of commonly used measures specifications has the advantage of increasing the consistency across different programs in which providers may participate. Competing measure and specifications from different programs make it challenging for providers trying to understand how to improve their performance, since what helps them on one set of measures may not help them on another. For other measures, the specifications may need to be modified. For example, HEDIS measures, which were originally developed for use at the plan level, may need to be modified for use at the physician level. Also, co-existing chronic conditions are very common in the Medicare populations—approximately 62 percent of Medicare beneficiaries 65 years of age or older had at least two chronic conditions in 1999, and 20 percent had at least five chronic conditions (Anderson and Horvath, 2002)—but very few measures specifications explicitly address how to handle patients with comorbid conditions. Following existing measures may result in very complicated patient care regimens, adverse drug interactions if patients are not carefully monitored, and a lack of focus on the aspects of care most important to patients with many comorbidities (Boyd et al., 2005). One option to mitigate this is to allow doctors to “opt-out” patients with many comorbidities, but this may lead to excessive exclusion of patients and artificially high performance scores. Another alternative is to develop measures that focus on patients with multiple chronic conditions. Once the measures specifications have been defined, CMS will be at a good point to engage physicians and have them review the specifications to identify any problems before the measures go live.

Physician Engagement and Communication

CMS will experience fewer difficulties implementing a P4P program if it successfully engages physicians and establishes an ongoing communication channel with them. CMS will need to communicate with physicians about and solicit their feedback on

- Measures specification.

- Design of reports and feedback mechanisms.
- Processes for data corrections, grievances, and appeals.

Engaging physicians is challenging and resource intensive. However, it is an especially important function, since front-line physicians are busy and have difficulty on an individual level attending to the array of program requirements placed on them by public and private insurers. CMS will face considerable difficulty communicating with and engaging on a regular basis the approximately 484,000 physicians nationally that provide Medicare Part B services (MedPAC, 2006). This is especially the case when operating at the individual physician level and with the numerous physicians working in solo or very small practice settings, since they lack the infrastructure to monitor measures being rewarded, provide quality improvement support, and enable submission of data to demonstrate adherence to measures. Nationally, 69.8 percent of physicians practice in solo or two-person practices, and less than 10 percent are in groups with more than 20 physicians (Casalino et al., 2003; AMA, 2004).

More successful programs engaged physician intermediaries (e.g., medical group leaders) to carry the messages to physicians, or they used physician forums to explain the program requirements, solicit input, and share lessons being learned. CMS will need to consider using local intermediaries (such as the QIOs, local medical societies, Part B carriers, or local physicians who have been put under contract) to serve as program leaders and provide a conduit through which to communicate with, support, and engage physicians in a Medicare physician P4P program. At present, none of these entities provides to individual physicians the type of support that would be required. As a result, CMS would need to build the capability to carry out these physician communication and support functions that are essential to the functioning of a P4P program. One program sponsor suggested the use of regional quality coalitions in this role. Some regional quality coalitions, such as IHA and MHQP, have experience measuring physician performance in their region.

Pilot Testing All Aspects of Implementation

It will be critical for CMS to test all aspects of program operations in advance of going national, including data collection, construction and scoring of measures, reporting of scores on measures back to physicians for their reaction, testing of the reporting format for understandability, and computation of payouts. CMS is starting to test some aspects of an eventual P4P program design through the AQA pilot projects, the CMS P4P demonstrations, and the PVRP, but more testing will be required.

Data Collection and Management

CMS will need to engage in a host of data collection and management activities to support a P4P program, including

- Data collection.
- Data cleaning.
- Date aggregation and warehousing.
- Data processing to produce performance scores.
- Auditing.

Careful consideration needs to be given to (1) what entities would have responsibility for these various tasks as part of a Medicare physician P4P program, and (2) what infrastructure would need to be built, since the current infrastructure at CMS would not support the additional data collection and management tasks required to operate a P4P program.

The steps necessary for and the burden associated with data collection, warehousing, cleaning, and processing depend on the types of data that will be used to construct measures. If claims data are used as the primary source of data, the data collection burden will not be large, but there is substantial work required to pull the data to run the measures. Running the 100 percent Medicare claims data—which is an enormous amount of data—is a substantial task. CMS should explore whether the existing infrastructure of the Medicare Part B carriers or regional quality coalitions could be used to provide regional data warehousing and analysis functions. The Part B carriers' expertise is in claims payment, and they do not currently possess the skills to perform the analytic processing and scoring of the data. In addition, any measures would require the use of multiple data sources (e.g., Part A and Part D data), which may prove to be an additional challenge. Some regional quality coalitions have experience performing these tasks, but experience is variable across coalitions.

If data from review of medical charts were used in a P4P program, there would have to be a mechanism for facilitating the submission of these data, as well as warehousing, cleaning, processing, and scoring. Providers could submit information to organizations using structured abstraction tools, which would limit the amount of data that would require centralized processing (basically, just numerators and denominators). While in absolute terms the magnitude of the data is smaller using chart reviews, the incremental burden associated with the collection and management of these data is substantially greater than with the use of administrative data. While the QIOs may be a

natural structure to use for receiving data, there are barriers in the types of information that QIOs are able to share with CMS. These would need to be addressed with legislation.

Auditing

Ensuring the validity of the data used to produce performance scores and make payouts is an important function. For example, the PGP demonstration includes an audit of 30 medical records per condition out of the sample of patients selected for medical record review method: If the first eight records in the sample are verified, the remaining 22 records for the condition are not audited. Groups must achieve 90 percent agreement between data abstracted by the group practice and audit sample. While this is feasible with a small number of groups, this kind of audit process would be very expensive on a national scale.

The form of the audit will depend on several factors, including types of data used, amount of resources available, and audit objectives (e.g., understand level of data quality among all providers versus identify physicians who are gaming the system). CMS has well-established rules and penalties in place for filing false claims and misreporting information, and these potentially would serve as one form of deterrence against providers gaming their data. However, the federal government estimates that 10 percent of Medicare and Medicaid spending is related to fraud (CMS, 2006); so while the process described here would provide some level of deterrence, additional checks would be needed for data that appears odd (e.g., outlier providers) or random audits would have to be performed.

Two key issues will affect the likelihood of gaming by the provider: (1) the size of the incentive and (2) the likelihood of getting caught and the associated repercussions. As the size of the incentive dollars grows relative to total income, physicians will have greater incentives to over-report actual incentives. With respect to getting caught, CMS will need to think about constructing a process that puts all physicians at risk of being audited. Auditing is a time-consuming and expensive activity to perform, and it slows the process of making payouts since a payout cannot occur until the data are validated and final scores are run (after adjusting for data corrections). One strategy that has been applied in some public reporting programs that assess provider quality performance is to do targeted auditing, where the program sponsor does a focused review of all providers who are designated as “best” and “worst,” and to do some random auditing of providers who fall in the middle of the performance distribution. This guards against mislabeling

providers and, in the case of a P4P program, would guard against paying providers who do not actually perform well.

Providing Program Support to Physicians (e.g., Training, Registries)

Comparative performance information and financial incentives may not be enough to generate individual behavior change by physicians. CMS should not expect physicians to change their practice without additional support. In our discussions with P4P projects, most indicated the importance of providing physicians with information and tools (e.g., registries of patients with specific conditions who need services) that more easily allow them to take action against deficiencies. Provision of technical assistance, education, and facilitation of best practices sharing may be an appropriate role for the QIOs or local medical societies.

Reporting/Feedback Process

CMS will need to produce feedback reports for all physicians who are measured. A critical step in the process of constructing feedback reports is testing their design with end-users. Many P4P programs have gained experience in this area, and CMS should examine the lessons they have learned.

One of the key lessons reported by most programs is that while the content matters, the process of feedback is as important as, or more important than, the physical scorecard. This speaks to the importance of engaging physicians in a dialogue about their performance and what they can do to remedy deficiencies. The process also sets up a respectful and nonjudgmental exchange that allows physicians a way to voice their concerns and challenges and feel supported in the process.

Process for Review and Correction of Data

Regardless of how carefully a P4P program is planned, piloted, and implemented, some providers will still have questions and concerns about the accuracy of their scores. Many will pass through an agitated, often angry period as the paradigm shifts from autonomy to accountability (Beckman, 2006). Having an established process through which providers can raise concerns about their scores and express opinions will improve provider confidence in the program. Structuring the process so that providers can review their data and raise any concerns prior to incentive payouts will improve the program's administrative efficiency.

P4P programs vary in whether and how they provide opportunities for review and correction of data prior to the results being posted on a scorecard and eligible for payment. P4P sponsors tended to fall into two camps on data review and correction: (1) the “fix going forward” camp, which felt that seeing their results would cause physicians to be more complete in their data submissions in future measurement and reporting rounds; and (2) the “fix now” camp, who sought to work with physicians in advance of reporting and payout to make corrections. Those in the second camp typically provided performance scores to providers with express notification that this was an opportunity to review and make corrections prior to the reports being finalized. Some P4P programs provided a hotline that providers could call to make changes to their information.

Monitoring of Implementation and Impact

P4P programs are charting new territory, and at this stage in the collective knowledge base, their development is largely a learning-by-doing process. Therefore, building pilot testing and evaluation and monitoring functions into the program’s basic design is important for providing feedback on lessons learned in order to make program adjustments. The monitoring functions should optimally address process issues associated with implementation—such as challenges in communicating with and engaging physicians, difficulties in producing measures from selected data sources, and provider perceptions of how the program is being administered—as well as the impact of the incentive program on meeting the program’s specified goals.

Process Evaluation

Process evaluation tends to emphasize qualitative methods for assessing the implementation experience and can provide real-time feedback for adjusting program design. This is in contrast to impact evaluation, which involves tracking quantitative measures (i.e., changes in performance scores, ROI) over time. The impact evaluation would occur over a longer time horizon, because of the time needed to collect the data, and also because improvements in performance scores in the early years of most P4P programs are often largely the result of improvements in data capture rather than real gains in quality improvement. Therefore, it is important to look at the impact beyond the initial years of implementation to truly gauge what has happened as a function of the P4P program. For the impact evaluation, it is essential to establish the evaluation component at the front end of the project to capture baseline information prior to the start of the P4P intervention.

Learning organizations have ongoing feedback and monitoring activities. Because P4P is still in the "learning" stages, CMS should build ongoing evaluation and monitoring into program design, as this will prove to be a critical source of information for enabling programmatic adjustments to improve program functioning. Based on the experiences of currently operating P4P programs, ongoing adjustments to the programs include modifying goals (such as adding in cost measures), adding more providers (with the advent of more measures being developed), adding or retiring measures (once performance has topped out), and modifying the payment structure (such as the frequency of payouts, rewarding improvement over time).

Measures Maintenance

Performance measures are at the core of P4P programs, and because the measures themselves are not static and their use within a P4P program is not permanently fixed, all P4P programs must build in procedures for ongoing measures maintenance. Typically there is a need to add measures, remove measures, and update measures specifications to reflect changes in the clinical knowledge. National measures development organizations, such as NCQA, have existing processes for measures maintenance. Based on a study by Mattke and Damberg (2006), there are three key functions for a measures maintenance system: (1) ad hoc review to deal with unexpected problems, (2) annual maintenance to incorporate changes in coding conventions, and (3) regular re-evaluation to thoroughly review measures in pre-defined intervals. The evaluation review criteria reflect the criteria used to select the initial set of measures (see Appendix E)—i.e., the importance, scientific soundness, feasibility, and usability.

SUMMARY

This chapter describes a host of design components and options that CMS will need to consider when designing a P4P program for Medicare physician services. The current published literature reveals an absence of scientific evidence to suggest a single best strategy for or approach to designing a P4P program that is likely to yield maximum benefits—whether the goal is to improve quality, reduce costs, or a combination of the two. There is little empirical knowledge about the impact of design component options (e.g., basis for attributing care to a physician, or differential impact of using performance thresholds or year-to-year improvement, few versus many measures), or about the circumstances (i.e., local market characteristics, such as organization of physician practices and exposure to private P4P programs sponsored by individual organizations or

regional coalitions) under which the various design components are more or less likely to have an impact.

Much of what we know about P4P program design comes from the early experiences of P4P program sponsors, which are learning step by step, in trial-and-error fashion, and modifying their programs as they go. P4P program development largely remains a process of charting new territory without a well-specified roadmap. As Robinson (2006) observes, “P4P programs have ambitious goals, and the effectiveness of these programs will be enhanced to the extent that we are clear on goals, choice of measures, choice of structure, and on how choices among measures and structure reflect priorities among goals.”

5. CONCLUSIONS

In 2006, the federal government will spend $600 billion for Medicare and Medicaid, covering 87 million beneficiaries (McClellan, 2006). By 2030, expenditures for these two programs are expected to consume 50 percent of the federal budget, which means that spending on health care will jeopardize funding for nearly all other, discretionary programs. This increasingly greater spending on health care in the absence of health care quality improvements lacks support among policymakers and the public, who pay taxes to finance Medicare. Policy inaction in the current cost and quality environment is not an option. The CMS Administrator, Dr. Mark McClellan, has publicly stated that to sustain the ability to fund these programs, CMS will have to change existing policies and practices.

One policy change that CMS and Congress are seriously considering is that of altering how physicians under the Medicare Part B program are compensated, moving a portion of reimbursement to payment based on performance. Our review of the literature and our discussions with P4P programs about program design components provided important information for CMS and other potential sponsors of P4P programs, information that can help inform and guide policy discussions. To that end, we address the following questions.

- Is P4P possible for Medicare physician services?
- If it is possible, what steps could CMS take to prepare for P4P for Medicare physician services?

The answer to the first question is a qualified yes. It is possible to implement P4P for Medicare physician services, but CMS will face important challenges in designing and implementing the program, and these challenges must be addressed if implementation of a P4P program is to be successful. The most significant of the challenges are

- **The absence of an existing organizational infrastructure within CMS** to manage the myriad components of an operating P4P program, particularly a program of the size and scope necessary to measure and reward all or most physicians in the Medicare Part B program. To support a P4P program's operations, many systems will have to be designed, built, tested, and maintained. Doing this will require dedicated and sustained resources.

- **The size and scope of a P4P program for Medicare physician services.** No other P4P program of comparable size exists.
- **The absence of infrastructure (personnel and information systems) at the individual doctor level** to support a P4P program's requirements. For example, the majority of physicians' offices do not have electronic health records or sufficient staff to perform chart abstractions that may be required to provide the information needed to construct the performance measure.
- **The difficulty of communicating with and engaging individual physicians** in the program to achieve the desired behavior changes. Organized medical groups have the staff and structure to facilitate communication between a P4P program sponsor and front-line physicians. When P4P programs are working at the level of the individual physician, however, there is no "local physician leadership" or point person to help facilitate communication and engagement with physicians about the program and to assist with behavior change.
- **The rapid timetable for ramping-up a national operation**. Given Congress's mounting pressure for action, CMS is unlikely to have time to pilot the program at multiple sites. The agency will be under pressure to roll out a national program in a short time period.
- **Physician resistance to transparency (public reporting)** of performance data. Some people assert that public transparency and accountability are a valuable addition to any P4P program to drive behavior change among physicians. Physicians, however, have expressed concerns about public reporting of performance results, citing data inaccuracy problems and the lack of accounting for differences in the patient populations treated.

CRITICAL P4P LESSONS FOR CMS TO CONSIDER WHEN DESIGNING A MEDICARE PHYSICIAN SERVICES P4P PROGRAM

Our review of currently operating P4P programs yielded eight critical lessons for CMS to consider, should it decide to move forward with a Medicare physician P4P program:

1. **Physician involvement and engagement in all steps of the process are necessary for successful implementation of a P4P program.** This helps build trust and creates a respectful process. It is critically important to provide a process

by which physicians can raise questions and provide input and can feel like co-partners in the program.

2. **Pilot testing of all aspects of program design and operation is critical,** and programs need to be open to making revisions based on what is learned during pilot testing. Trial and error is common during creation and implementation of P4P programs.
3. **Starting small and demonstrating success helps to build trust among the stakeholders involved in the program.**
4. **There is a need for flexibility in design and for potential customization**. Health care remains "local," and the variation in the organization of physicians across geographic markets and in the P4P programs already in play suggests a need for flexibility in design and potential customization of the program.
5. **A commitment to building and maintaining the operational infrastructure for P4P program operation is necessary.** Such functions as data warehousing; data aggregation; programming and analysis; data auditing; processes for appeals and data correction; performance feedback; communication with and engagement and support of physicians; measures maintenance; and modification of data collection processes are necessary and will require both monetary and personnel resources.
6. **Alignment among various program sponsors on what physicians are being measured and reported on and what they are being incentivized to do is important.** Alignment will reduce not only confusion for physicians, but data collection and reporting burdens as well, creating economies of scale in the provision of data.
7. **Providers need support (technical, organizational, and professional) to successfully participate in P4P programs**. This may take the form of patient registries, technical support, education, etc. The most important aspect is that performance feedback information be actionable.
8. **Continuous evaluation of program impact and operations is essential.** It provides critical information for adjusting the program.

TAKING THE FIRST STEPS TO IMPLEMENT A P4P PROGRAM FOR MEDICARE PHYSICIAN SERVICES

There are several steps that CMS could choose to take immediately and in the near term, as well as in the longer term, to prepare itself for designing and implementing a P4P program for Medicare physician services. These actions, if taken, would provide information to guide program planning, to help generate awareness and engagement among physicians, and to begin building the program infrastructure needed to support P4P.

Near-Term Steps (6 to 18 months)

Model critical design components using existing data

CMS could start laying the groundwork for structuring a P4P program by modeling options for various program design components using existing Medicare claims data. Some of the critical design issues to be addressed in modeling the components are (1) the implications of different attribution rules, (2) the number of measures that can be scored today using claims data, (3) the number of physicians whose performance can be reliably scored using measures based on administrative data, and (4) the increase in the number of physicians that CMS could score if scores were based on composite measures versus individual indicators of performance. CMS also will be able to look to the newly started AQA pilots for information about some of these design issues.

Monitor the experiences of the Physician Voluntary Reporting Program and consider how to address emerging lessons in the design of P4P for Medicare physician services.

Implementation of the PVRP, a program started in January 2006 that will provide internal comparative performance feedback to providers on a starter set of 16 measures, offers CMS a potential foundation on which to build a P4P program. The lessons being learned in the PVRP will provide CMS with valuable information; in particular, the monitoring of physician participation and growth in participation over time will provide indications about the readiness of physicians nationally to provide information on the selected measures. Interviews with physicians could give CMS valuable insights about why physicians agreed or did not agree to participate. Participating providers could describe the challenges they experience with the data collection and reporting process, as well as their reactions to performance feedback reports. Non-participating providers could help to identify barriers to participation and actions needed to address them. Information gained from physician interviews could be useful in determining how to

modify the program going forward as a stepping-stone to full P4P. The interviews also would allow CMS to build communication channels with physicians before a P4P program is implemented and would constitute an important step in soliciting physician input about program design.

Mid-Term Steps (18 to 36 months):

Create incentives for participation in the PVRP as a way to help physicians move toward understanding performance measurement, to build systems to support measurement, and to work towards performance transparency.

Low participation in PVRP may suggest the need to provide inducements for participation, such as pay-for-reporting. Increasing the number of physicians participating in the PVRP would give physicians an opportunity to gain experience with submitting data and receiving performance feedback, well in advance of P4P. Allowing physicians time to see performance scores in a confidential manner gives them the opportunity to improve systems for data capture and to identify and correct quality problems in advance of public reporting. This is an important step for CMS to take on the path to public transparency.

Expand the PVRP measurement set and administrative collection of measures.

CMS could also continue to expand the PVRP 16-measure set so that it is consistent with P4P program design decisions about which measures to reward to drive improvements. In addition, to support the administrative reporting of data to produce performance measures, particular attention should be paid to modifying the HCFA 1500, the form physicians use to submit claims to Medicare, to capture administratively the data elements needed to support performance measurement (e.g., working with the AMA to develop CPT supplemental codes).

Plan for program evaluation and collect baseline data.

It is also very important for CMS to build into its P4P design the continuous evaluation of program implementation and effects. Ongoing evaluation will give CMS critical information so that the program can be adjusted in real time. Assessment of program effects would require that CMS collect baseline information about performance. If CMS expects to compute the ROI, it will need to track program costs.

Longer-Term Step (36 months and beyond)

Scale up incrementally and continue to build infrastructure capacity

As the PVRP matures, CMS could scale up the program incrementally by adding measures and physician specialties and continuing to build infrastructure to accommodate the program's increasing size. By building gradually on successes, CMS will help to build trust within the provider community and will gain experience along the way.

MUCH REMAINS UNKNOWN ABOUT P4P

As a result of lack of evidence and limited program experiences, there are still many unanswered questions concerning P4P. Little is known about the impact of P4P on either the cost or the quality of care delivered. While it is likely that IT facilitates improved performance, the extent to which P4P programs in and of themselves motivate investments in IT is unknown. Furthermore, the effect of P4P on vulnerable patients is unknown. It may reduce disparities through overall improvements in quality of care delivered. Alternatively, physicians may refuse to take on or continue to treat patients they view as being difficult to treat or less compliant because of the possibility of reduced performance scores.

Also unknown is the best way to structure rewards to motivate physician behavior change. This includes the impact of negative versus positive financial incentives and the use of bonuses versus enhanced fee schedules, as well as the magnitude of the incentive necessary to motivate change. In addition, the extent to which performance measurement and public reporting are key drivers of behavior change, relative to the effects of incentive payments, is not known. Work to answer these questions will provide valuable information on best design attributes and the implications of P4P programs for CMS and other P4P sponsors.

FINAL NOTE

There are two prerequisites for a successful P4P program: good implementation and good program design. Much of the information about the first of these, good implementation, has been learned from existing P4P programs. The second prerequisite, good program design, is an area that much less is known about. There is limited literature to inform the selection of one design approach over another, and few programs have carefully evaluated the effects of their designs. There are, however, AQA pilots currently

under way that are designed to shed light on some of the unknowns around P4P. These pilots should provide CMS with valuable information.

Appendix

A. PAY-FOR-PERFORMANCE DESIGN PRINCIPLES AND RECOMMENDATIONS SET FORTH BY NATIONAL ORGANIZATIONS

							PHYSICIAN GROUP ORGANIZATIONS					
Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
Medicare Specific												
Medicare should fund the program by setting aside a small share of payments in a budget neutral approach		X										
A Medicare P4P program must not be budget neutral or subject to artificial Medicare payment volume controls											X	X

[5] Joint Commission on Accreditation of Healthcare Organizations, 2005
[6] Miller, 2005
[7] National Business Group on Health, 2005
[8] eHealth Initiative and Foundation, 2005
[9] American Healthways and Johns Hopkins, 2004
[10] Pacific Business Group on Health, 2005
[11] American Medical Association, 2005
[12] American Academy of Family Practice Physicians, 2005
[13] American College of Physicians, 2005
[14] ACCF Quality Strategic Oversight Committee, 2005
[15] Medical Group Management Association, 2005
[16] Society for Vascular Surgery, 2005

Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	PHYSICIAN GROUP ORGANIZATIONS AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
Medicare should distribute all payments that are set aside to providers achieving quality criteria		X										
Medicare should establish a process for continual evolution of measures		X										
A Medicare P4P program should be phased in gradually starting with reporting on structural measures and moving to enhanced payment based on evidence based clinical measures									X			
Medicare should consider expanding the proportion of payment based on reimbursement data over time			X									
A Medicare P4P program must be pilot tested across settings and specialties and phased-in over an appropriated period of time												X
Measures												
Programs should utilize accepted, evidence-based measures	X	X	X		X	X	X	X	X	X	X	X
Measures adopted should be developed by nationally recognized measurement organizations and recommended by consensus building organizations			X				X					X
Structure, process and outcome measures should be utilized			X		X				X	X		
Measures should be selected from the following domains: patient centeredness, effectiveness, safety, and efficiency				X								

Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	PHYSICIAN GROUP ORGANIZATIONS					
							AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
Measures should include efficiency measures			X							X		
Efficiency measures should only be used when both the cost and quality of a particular treatment is considered									X			
Mechanisms must be established to allow performance awards for physician behaviors in hospital settings that produce cost savings												X
When measuring quality, focus on misuse and overuse as well as underuse			X							X		
Measures should be high volume, high gravity, strong evidence base, gap between current and ideal practice, good prospects for quality improvement, measurement reliability, measurement feasibility				X								
Program designers should include a sufficient number of measures across a spectrum of health promotion activities to provide a balanced view of performance				X		X						
The development, validation, selection and refinement of measures should be a transparent process that has broad consensus among stakeholders.									X			
Measures should be stable over time							X		X			
Measures should be kept current to reflect changes in clinical practice												X
Local measures should closely follow national measures as long as they are reportable from electronic data sets						X						

Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	**PHYSICIAN GROUP ORGANIZATIONS** AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
To prevent physician de-selection of patients, programs should use risk adjustment methods	X	X			X		X	X	X	X	X	X
Outcome measures must be subject to the best-available risk-adjustment for patient demographics, severity of illness and co-morbidities							X			X		
To ensure fairness, performance data must be fully adjusted for sample size and case mix composition including age/sex distribution, severity of illness, number of comorbid conditions, patient compliance and other features of the practice or patient population that may influence the results					X				X		X	X
Data Collection, Reporting, Feedback												
Data should be collected without undue burden on providers	X	X					X	X	X			
Programs must reimburse physicians for any administrative burden for collecting and reporting data							X	X			X	X
Allow physicians to review, comment and appeal results prior to payment or reporting							X		X		X	X
Programs should have a mix of financial and non-financial incentives i.e. public reporting	X		X		X							
Physician performance data must remain confidential and not subject to discovery in legal proceedings												X

Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	PHYSICIAN GROUP ORGANIZATIONS					
							AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
Public reporting to consumers is essential						X						
Performance data feedback should provide comparisons to peers and benchmarks								X				
Educational feedback should be provided to physicians							X		X			
Programs should favor the use of clinical data over claims-based data										X		
Programs should use administrative data and data from medical records							X					
Performance data should be audited	X							X		X		X
Metric assessments and payments should be made as frequently as possible to better align rewards to performance					X			X				
Data reporting must not violate patient privacy							X		X			
Incentives												
Align reimbursement with the practice of high quality, safe health care	X	X			X		X	X	X	X	X	X
Incentives should be based on rewards, not penalties					X		X	X	X	X		X
Programs should reward providers based on improving care and exceeding benchmarks		X			X	X	X	X	X			X
Programs must not reward physicians based on ranking compared with other physicians in the program							X					

Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	PHYSICIAN GROUP ORGANIZATIONS					
							AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
Incentives must be significant enough to drive desired behaviors and support continuous quality improvement									X			
Provide positive physician incentives for adoption and utilization of information technology	X		X	X	X	X	X	X		X	X	X
Programs implemented by either the public or private sector involving HIT should incentivize only those application and systems that are standards-based to enable interoperability and connectivity and should address the transmission of data to the point of care				X								
General Program Design												
Programs should offer voluntary physician participation							X	X			X	X
Physicians should be involved in the program design							X	X	X		X	X
Most providers should be able to demonstrate improved performance-focus on areas needing improvement		X			X		X					
When selecting areas of clinical focus, programs should strongly consider consistency with national and regional efforts	X				X	X						
P4P assessments should be done with sample sizes (denominators) large enough to produce statistically significant results					X		X	X	X			

Pay-for-performance Design Principles/Recommendations	JCAHO[5]	MedPAC[6]	Natl.Bus.Group on Health[7]	eHealth Initiative Found[8]	Health-ways/ Johns Hopkins[9]	Pacific Bus Group on Health[10]	PHYSICIAN GROUP ORGANIZATIONS					
							AMA[11]	AAFP[12]	Am College of Physicians[13]	Am College of Cardiol-ogy Found[14]	Medical Group Mgmt Assoc[15]	Surgical Specialty Orgs*[16]
Programs should be consolidated across employers and health plans to make the bonuses meaningful and the program more manageable for physicians.								X				
Programs should be designed to include practices of all sizes and levels of information technology capabilities							X	X				
Physician organizations should be the accountable entity in P4P programs rather than individual physicians					X		X			X		
Payments should recognize systemic drivers of quality in units broader that individual provider organizations and practitioner groups	X											
Programs should be designed to acknowledge the united approach i.e. team approaches, integration of services, continuity of care	X						X			X		
The results of P4P programs should not be used against physicians in health plan credentialing, licensure or certification							X	X	X			
The data or the program should be adjusted for patient noncompliance					X		X	X	X			
Programs should incorporate periodic objective evaluations of impacts and make adjustments									X	X		

*Surgical Specialty Organizations:
American Academy of Ophthalmology
American Academy of Otolaryngology

American Association of Neurological Surgeons
American Association of Orthopedic Surgeons
American College of Surgeons
American Society of Cataract and Refractive Surgery
American Society of Plastic Surgeons
American Urological Association
Congress of Neurological Surgeons
Society for Vascular Surgery
Society of American Gastrointestinal and Endoscopic Surgeons
Society of Gynecologic Oncologists
Society of Surgical Oncology
The Society of Thoracic Surgeons

B. CMS REPORTING ACTIVITIES AND PAY-FOR-PERFORMANCE DEMONSTRATIONS

CMS has begun to chart the path of P4P with a series of demonstrations and programs, of which 6 are currently ongoing and 5 have yet to be implemented. The structure of the incentives used in the CMS programs differs from many P4P programs nationally. The CMS programs are much more likely to involve shared savings, where providers receive the financial incentive only if they meet the savings threshold set forth by CMS. In addition, in some of the demonstrations, the providers are not being paid differentially based on performance, but lose their portion of the shared savings if certain quality performance thresholds are not met. Many of the demonstrations focus on specific clinical conditions, the most common being diabetes, congestive heart failure and coronary artery disease. In January 2006, the PVRP started and is likely to lay the groundwork for a physician P4P program. Here we briefly describe 11 ongoing and planned P4P activities sponsored by CMS as well as the PVRP.

Ongoing CMS Physician Reporting Activity

Physician Voluntary Reporting Program (PVRP). Voluntary reporting of quality of care data through the PVRP began in January 2006. The initial 36 measures to be included in the program were scaled back to a 16 measure starter set to reduce reporting burden and create better alignment with other physician quality measurement activities (the list of measures is included in Appendix C). These measures can be reported using existing claims based systems and a set of G-codes with the physician claim form. The G-codes are viewed as an interim step until electronic submission with EHRs is widespread. They will supplement the usually submitted data and are submitted with a zero charge. The QIOs are working with physicians in the creation of systems to facilitate reporting the measures. Providers participating in PVRP will receive feedback on their performance (anticipated start of feedback is summer 2006). While providers may report on all 36 measures, feedback will be limited to the 16 measure starter set. Participating providers may also provide input on how to reduce the burden and improve the process of quality reporting. It is anticipated that the scope of the PVRP will expand as additional measures are developed through consensus activities.

Ongoing CMS Pay-for-Performance Activities

Physician Group Practice Demonstration. The three-year Physician Group Practice Demonstration, involving 10 medical groups with 200+ physicians each initiated in April 2005. A set of 32 measures for six clinical conditions is being phased in over the three-year period. Diabetes is the focus in year 1; congestive heart failure and coronary artery disease are added in year 2 and hypertension and screening for colorectal and breast cancer are included in year 3. Bonus payments are contingent on the medical group generating savings and the proportion devoted to cost and quality varies by program year, with 70 percent eligible as cost savings and 30 percent as quality bonus in program year one, 60/40 in year 2 and 50/50 in year 3. Each measure carries a weight – one point for measures that are medical record-based, four points for measures that are primarily claims-based. The total number of available "quality points" in a given measurement year is the sum of all of the weights for the measures that are designated for that year. Quality points for each measure are earned if the group meets any one of three targets: 1) the higher of 75 percent compliance or the Medicare HEDIS mean (if applicable); 2) demonstrate a 10 percent or greater reduction in the gap between the group's baseline and 100 percent compliance; 3) or achieve at least the 70th percentile Medicare HEDIS level. The ratio of quality points earned to total available quality points determines the proportion of the bonus set aside for quality performance that the group will earn.

Premier Hospital Quality Incentive Demonstration. CMS' first venture in P4P was the voluntary Premier Hospital Quality Incentive Demonstration program to reward Premier hospitals for superior performance on a set of clinical quality measures.14 The three-year demonstration, which started in October 2003, includes five clinical areas: acute myocardial infarction, heart failure, pneumonia, coronary artery bypass graft and hip and knee replacement. Individual measures are rolled into composite scores for each clinical condition, on which incentives are determined. Hospitals performing in the top 10 percent for a specific condition will receive a 2 percent bonus on Medicare payments for that condition, while hospitals in the second 10 percent receive a 1 percent bonus. All hospitals in the top 50 percent receive recognition on the CMS website. In the third year of the program, negative financial incentives also exist. Any hospitals in the third year performing below the level of 10th percentile from the first year of the program will have a 2 percent reduction in Medicare payments for the relevant clinical condition, while any hospitals performing below the 20th percentile from the first year will receive a 1 percent reduction. Early program results released by CMS suggest improved performance in the

last quarter of the first year of the demonstration compared to the first quarter of the first year of the program. Results from the full evaluation, however, are not yet available.

Hospital Quality Initiative. This initiative is part of HHS's broader National Quality Initiative that aims to increase hospitals' collection and reporting of performance data. Mandated by Section 501(b) of the MMA, the Hospital Quality Initiative implements financial incentives for hospitals to report data on 10 specific quality measures. Hospitals that voluntarily report their quality scores through the CMS Hospital Quality Incentive (HQI) Data Initiative receive a 0.4 percent higher Medicare payment update for fiscal years 2005, 2006, and 2007. The 10 measures are among a set of 20 measures developed by CMS and JCAHO in collaboration with the Hospital Quality Alliance and various researchers, and endorsed by the NQF. The measures address acute myocardial infarction/heart attack, heart failure, pneumonia, and surgical infection prevention. CMS reports that virtually all hospitals eligible to participate are reporting the required data (98.3 percent) (CMS, 2005).

Health Support Program (previously named Chronic Care Improvement Program). Authorized by Section 721 of the MMA, this program will test a population-based model of disease management for chronically ill Medicare beneficiaries with advanced CHF and/or complex diabetes. Participating organizations must implement individualized care management plans that include a designated point of contact for the beneficiary; steps for coordinating care among different providers; patient and caregiver education; information about hospice, palliative, and end-of-life care; and the use of monitoring technologies that enable patients to record and transmit to their physicians information on vital signs and other aspects of their condition (Super, 2004).

CMS is phasing in implementation of the program, and as of late-2005 had chosen nine sites to participate, all of which are insurers or disease management companies. These participants are Humana in South and Central Florida, XLHealth in Tennessee, Aetna in Illinois, LifeMasters in Oklahoma, McKesson in Mississippi, CIGNA in Georgia, Health Dialog in Pennsylvania, American Healthways in Washington, D.C. and Maryland, and Visiting Nurse Service of New York and United Healthcare in Queens and Brooklyn, New York. CMS has entered into three-year contracts with each of these organizations (CMS, 2005).

The program will utilize a risk-based financial model to encourage participants to control the costs of treating this population. Each participating organization will receive a monthly fee for each beneficiary. CMS will expect each organization to achieve at least 5 percent cost savings, net of fee payments, compared to a control group of beneficiaries. If an organization fails to do so, it must refund to CMS some or all of its program fees. The organization may also have to refund fees if care for its population of chronically ill beneficiaries does not meet certain quality standards, or if beneficiaries or providers report a certain level of dissatisfaction (CMS, 2005; Super, 2004).

CMS is required to have the program independently evaluated. Within two years of beginning implementation, CMS has the option of expanding the program to more sites so long as evaluators find the program to be successful in improving quality of care and controlling costs (CMS, 2005; Super, 2004).

Disease Management Demonstration for Severely Chronically Ill Medicare Beneficiaries. Authorized by the Medicare, Medicaid, and SCHIP Benefits Improvement and Protection Act of 2000 (BIPA), this three-year program was launched in February 2004. The program tests the capability of disease management models, combined with the provision of comprehensive prescription drug coverage, to improve care for Medicare FFS beneficiaries with advanced-stage CHF, diabetes, or coronary artery disease. The program also seeks to lower the cost of treating these beneficiaries (CMS, 2005; Guterman, 2003).

Three sites are participating in the program: XLHealth in Texas, CorSolutions in Louisiana, and HeartPartners in California and Arizona. Participants receive a monthly payment for each enrolled beneficiary to offer disease management services and a comprehensive prescription drug benefit (CMS, 2005). Up to 30,000 Medicare beneficiaries are enrolled in the program (Guterman, 2003). Eligible beneficiaries, whose participation is voluntary, were randomly assigned to treatment and control groups. Those in the treatment group receive disease management services plus prescription drug benefits, while those in the control group receive "usual care" (typical Medicare benefits with no disease management or prescription drug coverage) (CMS, 2003).

Disease Management Demonstration for Chronically Ill Dual Eligible Beneficiaries. This demonstration program tests the provision of disease management services to Medicare beneficiaries who are also eligible for Medicaid and suffer from advanced-stage CHF, diabetes, or coronary artery disease. There is only one organization

participating in the program, LifeMasters, which is delivering these services to beneficiaries in Florida. LifeMasters receives a fixed monthly fee for each enrolled beneficiary, and has assumed full financial risk for these fees based on certain performance targets. Under the program, beneficiaries receive disease management services from LifeMasters and pharmacy benefits from the state Medicaid program. The goal is to achieve better, more coordinated care for dually eligible beneficiaries with chronic diseases while also reducing total program costs. Cost savings are to be shared between CMS and LifeMasters (CMS, 2005).

Future CMS Pay-for-performance Activities

Medicare Care Management Performance Demonstration. Authorized by Section 649 of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003 (MMA), this demonstration program aims to enhance the quality and coordination of care for chronically ill Medicare FFS beneficiaries through the use of health information technology (HIT). The program is modeled on the "Bridges to Excellence" program. Eligible demonstration sites are small and medium-sized physician practices. Sites will have some discretion over the kinds of HIT they implement through the demonstration. These include electronic medical records, patient registries, physician alerts, and clinical decision support systems. Physician practices will receive bonus payments for adopting and using HIT, and for meeting certain performance targets (Magno, 2005). The chronic diseases targeted by the program are diabetes, heart failure, and coronary artery disease (Magno, 2005; CMS, 2005).

The program is scheduled to begin in 2006 and continue for three years. CMS has not yet selected demonstration sites but plans to select up to 4 sites from the following states, which are also involved in CMS's DOQ-IT program: Arkansas, California, Massachusetts, and Utah. Section 649 of the MMA specifies that CMS must select no more than 4 demonstration sites (P.L. 108-173, 2003). CMS has indicated that up to 2,800 physicians will participate (Federal Register, 2005). As is the case with the DOQ-IT program, state Quality Improvement Organizations (QIOs) are to provide technical assistance to demonstration sites under the program (Magno, 2005).

Medicare Health Care Quality Demonstration. Mandated by Section 646 of the MMA, this is a five-year demonstration program that has a strong—though not exclusive—focus on HIT. The program is intended to further the six dimensions of high-

quality care identified by the IOM (patient safety, effectiveness, efficiency, timeliness, patient-centeredness, and equity). The specific goals of the program are to use financial incentives to encourage adoption of evidence-based best practices and use of decision support tools to reduce variations in quality of care and patient safety. Additional goals include increased efficiency and improved cultural competence. Demonstration sites have not yet been selected. Eligible applicants include physician group practices, integrated health care delivery systems, and regional coalitions of physician groups or integrated delivery systems. CMS plans to select a total of between 8 and 12 sites. The program targets Medicare FFS beneficiaries and those with Medicare Advantage plans (CMS, 2005; Magno, 2005).

Specific financial incentives have not yet been defined. Bonus payments will be tied to cost savings and improvements in as of yet unspecified process and outcome measures for the target population of beneficiaries compared to a similar population. Applicants can propose payment methodologies (e.g., shared savings, capitation, per member per month fee, restructured FFS, regional global budgets) (Magno, 2005).

CMS expects to select and fund sites in two rounds. The first round of applications is due January 30, 2006, and the second round is due September 29, 2006 (Federal Register, 2005).

ESRD Disease Management Demonstration. This four-year demonstration program aims to "increase the opportunity for Medicare beneficiaries with end stage renal disease (ESRD) to join managed care plans," and “has been designed to test the effectiveness of disease management models to increase quality of care for ESRD patients while ensuring that this care is provided more effectively and efficiently” (CMS, 2005).

The basic approach of this program is to have dialysis providers partner with insurance companies offering Medicare Advantage plans to provide managed care plans for ESRD patients, with a strong disease management focus. In October 2005, CMS chose two provider-insurance company partnerships to participate and the program is projected to become operation in 2006. One partnership consists of DaVita, a dialysis provider, and SCAN Health Plan. The second partnership consists of Fresenius Medical Care North America (along with its own Fresenius Health Plan) and Sterling Life Insurance Company, as well as American Progressive Life and Health Insurance Company. These partnerships serve Medicare beneficiaries in four states: California,

Texas, Pennsylvania, and Massachusetts. The program targets Medicare beneficiaries with Medicare Advantage plans (CMS, 2005).

These organizations will receive capitated payments for managing the care of ESRD patients, and will be eligible for bonus payments if they improve upon past performance and perform "above the National averages for quality measures related to dialysis". CMS plans to reserve 5 percent of the capitation payment rates for these incentive payments (CMS, 2005).

Care Management for High Cost Beneficiaries Demonstration. This is the first CMS demonstration program to specifically target high-cost and high-risk FFS Medicare beneficiaries. Under this three-year program, CMS is testing several care management approaches to providing lower-cost but higher-quality care to these beneficiaries. Such approaches include "intensive case management, increased provider availability, structured chronic care programs, restructured physician practices, and expanded flexibility in care settings" (CMS, 2005).

In the fall of 2005, CMS selected six organizations to participate. These include a consortium of physician clinics, a physician home visiting program, a hospital-physician group collaborative, an integrated delivery system, a renal disease management organization, and a consortium consisting of a hospital and physician practices. These organizations are located in Oregon, Washington, California, Texas, Florida, Massachusetts, and New York.

The payment methodology will be similar to that being implemented for the Chronic Care Improvement Program (see below). Each participating organization will receive a monthly fee for each beneficiary participating in the program to cover administrative and care management costs. But "organizations will be required to assume financial risk if they do not meet established performance standards for clinical quality of care, beneficiary and provider satisfaction, and savings to Medicare." Participants' performance will be judged on the basis of certain quality of care standards as well as achieving cost savings (CMS, 2005).

Nursing Home Pay-for-performance Demonstration. CMS is currently planning a P4P demonstration program that will target Medicare FFS beneficiaries in nursing homes. CMS has contracted with Abt Associates to design the program. Under the program, CMS intends to offer financial rewards to nursing that meet certain quality standards of care. Because the program is still in the design phase, CMS has not solicited

proposals or selected sites (CMS, 2005). CMS plans to start the program in late 2006 or early 2007, with a few hundred facilities in 3 or 4 states (Abt Associates, 2005).

C. PHYSICIAN PERFORMANCE MEASUREMENT

Numerous measures of physician performance have been developed by quality and research organizations, physician groups and consortiums for purposes of quality improvement, benchmarking, and accountability. The indicators typically measure clinical processes and outcomes; efficiency or appropriateness of care; use of information technology or clinical decision support; administrative processes; patient experience; and patient safety. In the past year, there has been increasing focus on the development of measure for individual specialties. Many of the measurement sets utilize common measures or measures with only slight variations. We identified the physician performance measures described below primarily through web-based sources including National Committee for Quality Assurance (NCQA), Center for Medicare & Medicaid Services (CMS) , the Agency for Healthcare Research and Quality (AHRQ), the National Quality Forum (NQF) and specialty society websites. We have included all the measures applicable to physicians in each set we describe even though some are not relevant to the Medicare population.

As stated previously, clinical measures are the most commonly utilized indicators of physician performance, and 94 percent of P4P programs utilize some HEDIS clinical measures (Med-Vantage 2005). HEDIS is a set of standardized performance measures developed by the National Committee for Quality Assurance (NCQA) to measure and report health plan quality, however a number of the indicators are being used, with some adjustments, to measure physician groups as well as individual physicians. HEDIS measures are heavily used because they are well specified, they are not new to doctors so they are well-accepted, and health plans are accountable for the measures also allowing alignment. The measures cover prevention (immunizations and screening) as well as management of cancer, diabetes, heart disease, asthma, mental health and smoking. Several new measures addressing appropriateness of care have been added for 2006 reporting. HEDIS also includes measures of access to care as well as the CAHPS 3.0H survey measuring members' experience with their care. Data for the construction of measures are collected from a combination of administrative data, medical records and patient surveys. Table 5 lists the HEDIS measures.

Table 5
HEDIS 2006 Effectiveness of Care Measures

MEASURE	Status
Childhood Immunization Status	
Adolescent Immunization Status	
Appropriate Treatment for Children With Upper Respiratory Infection	
Appropriate Testing for Children With Pharyngitis	
Inappropriate Antibiotic Treatment for Adults With Acute Bronchitis	New 2006
Colorectal Cancer Screening	
Breast Cancer Screening	
Cervical Cancer Screening	
Chlamydia Screening in Women	
Osteoporosis Management in Women Who Had a Fracture	
Controlling High Blood Pressure	
Beta-Blocker Treatment After a Heart Attack	
Persistence of Beta Blocker Treatment After a Heart Attack	
Cholesterol Management for Patients with Cardiovascular Conditions	
Comprehensive Diabetes Care: Eye Exams HbA1c Testing HbA1c Control LDL-C Screening LDL-C Control Nephropathy	
Use of Appropriate Medications for People with Asthma	
Use of Spirometry Testing in the Assessment and Diagnosis of COPD	New 2006
Follow-up After Hospitalization for Mental Illness	
Antidepressant Medication Management: Optimal Practitioner Contact/Follow-up Acute Phase Continuation Phase	
Follow-up Care for Children Prescribed ADHD Medication	New 2006
Glaucoma Screening in Older Adults	
Use of Imaging Studies for Lower Back Pain	
Disease Modifying Anti-Rheumatic Drug Therapy in Rheumatoid Arthritis	New 2006
Annual Monitoring for Patients on Persistent Medications	New 2006
Drugs to be Avoided in the Elderly	New 2006
Medical Assistance with Smoking Cessation	
Flu Shots for Adults Age 50-64	
Pneumonia Vaccination for Older Adults	
Medicare Health Outcomes Survey	New 2006
Management of Urinary Incontinence in Older Adults	

Discussing Urinary Incontinence Treatment	
Physical Activity in Older Adults Discussing Physical Activity Advising Physical Activity	

Additional clinical measures for primary care as well as specialty care are in the development, testing and approval stages across an array of organizations. CMS has worked collaboratively with the American Medical Association's Physician Consortium for Performance Improvement (PCPI) and NCQA to develop a set of ambulatory care measures. The PCPI includes representatives from more than 70 medical specialties, state medical societies, CMS and AHRQ. Together they submitted 99 measures to the National Quality Forum (NQF) for expedited review as part of Phase 2 of the group's Ambulatory Care Quality Measurement and Reporting project. The measures address asthma, depression/behavioral health, bone conditions, diabetes, heart disease, hypertension, pre-natal care, and prevention/screening. The measures submitted by PCPI were developed for internal quality improvement purposes, while the NCQA measures are recommended for public reporting and accountability. It is recommended that PCPI measures be constructed from the medical record. A complete list of the measures submitted; those approved on August 3, 2005, and those moving to a second round of voting are located in Table 6.

Table 6
Ambulatory Measures Submitted to NQF

MEASURES	**SOURCE**	**STATUS** 11/05*
Asthma/Respiratory Illness		
Asthma: Use of Appropriate Medications	NCQA	X
Asthma Assessment	PCPI	X
Asthma: Pharmacologic Therapy	PCPI	X
Asthma: Pharmacologic Therapy (distribution of control therapy by medication, severity classification, age range)	PCPI	
Appropriate Treatment for Children with Upper Respiratory Infection	NCQA	X
Appropriate Testing for Children with Pharyngitis	NCQA	X
Depression/Behavioral Health		
Screening for Depression and Follow-up (screening)	VHA	
Screening for Depression and Follow-up (follow-up assessment or referral)	VHA	

MEASURES	SOURCE	STATUS 11/05*
Continuation of Antidepressant Medication	PCPI	
Optimal Practitioner Contacts for Medication Management	NCQA	X
Effective Acute Phase Treatment	NCQA	X
Effective Continuation Phase Treatment	NCQA	X
Follow-up After Hospitalization for Mental Illness	NCQA	
Diagnostic Evaluation (depression)	PCPI	
Suicide Risk Assessment	PCPI	
Severity Classification (initial visit)	PCPI	
Treatment: Psychotherapy, Medication management and/or Electroconvulsive Therapy (appropriate therapy)	PCPI	
Bone Conditions		
Osteoporosis Management in Women who had a Fracture	NCQA	Did not pass
Osteoarthritis: Assessment for use of OTC Anti-inflammatory or Analgesic	AAOS/PCPI	X
Osteoarthritis: Gastrointestinal Prophylaxis	AAOS/PCPI	
Osteoarthritis: Functional and Pain Assessment	AAOS/PCPI	X
Osteoarthritis: Non-Steroidal Anti-Inflammatory Drug Risk Assessment	AAOS/PCPI	
Osteoarthritis: Physical Examination of the Involved Joint	AAOS/PCPI	
Osteoarthritis: Anti-Inflammatory /Analgesic Therapy	AAOS/PCPI	
Osteoarthritis: Therapeutic Exercise	AAOS/PCPI	
Diabetes		
A1c Management –Screen (one or more tests)	NCQA	X
A1c Management –Screen (one or more tests)	AMA	2nd round
A1c Management –Screen (distribution of number of tests)	AMA	2nd round
A1c Management –Control (poor control)	NCQA	X
A1c Management –Control (distribution of A1c values)	AMA	2nd round
Blood Pressure Management	NCQA	X
Blood Pressure Management (distribution of values)	AMA	2nd round
Lipid Management (at least one LDL)	NCQA	X
Lipid Management (at least one lipid profile or all component tests)	AMA	2nd round
Lipid Management (at least one LDL)	NCQA	X
LDL Cholesterol level <130mg/dL	NCQA	X
LDL Cholesterol level (distribution)	AMA	2nd round
LDL Cholesterol level <100mg/dL	NCQA	X
Urine Protein Screening (at least one test for microalbumin)	NCQA	X
Urine Protein Screening (any test for microalbumin)	AMA	2nd round
Urine Protein Screening (patients with no urinalysis or urinalysis with negative or trace protein receiving a test for microalbumin	AMA	2nd round

MEASURES	SOURCE	STATUS 11/05*
Eye Examination (dilated eye exam or seven standard field stereoscopic photos with interpretation and comparison to prior year)	NCQA	X
Eye Examination (dilated exam by ophthalmologist or optometrist)	AMA	2nd round
Eye Examination (seven standard field stereoscopic photos with interpretation	AMA	2nd round
Foot Examination (at least one)	NCQA	X
Foot Examination (at least one-visual, sensory exam, pulse exam)	AMA	2nd round
Smoking Cessation (status ascertained and documented annually)	NCQA	X (a)
Smoking Cessation (assessed for smoking status)	AMA	2nd round
Smoking Cessation (% smokers)	AMA	2nd round
Smoking Cessation (Patients recommended or offered counseling or pharmacologic intervention	AMA	2nd round
Aspirin use	AMA	
Influenza Immunization (received during recommended calendar period)	AMA	2nd round
Influenza Immunization (received or refused)	AMA	2nd round
Pregnancy Counseling (pre-pregnancy counseling)	AMA	
Pregnancy Counseling (family planning, contraception)	AMA	
Heart Disease		
Coronary Artery Disease (CAD): Antiplatelet therapy	PCPI/ACC/AHA	X
CAD: Drug Therapy for Lowering LDL Cholesterol	CMS/PCPI/ ACC/AHA	X
CAD: Beta Blocker Therapy After a Heart Attack	NCQA	X
CAD: Beta Blocker Therapy-Prior MI	PCPI/ACC/AHA	X
CAD: Blood Pressure Measurement (during last office visit)	PCPI/ACC/ AHA	
CAD: Blood Pressure Measurement (<140/90 mm Hg)	PCPI/ACC/AHA	
CAD: Blood Pressure Measurement (distribution)	PCPI/ACC/AHA	
CAD: Lipid Profile (at least one)	NCQA	X
CAD: Lipid Profile (at least one or all component tests)	PCPI/ACC/ AHA	X
CAD: Lipid Profile (distribution)	PCPI/ACC/ AHA	
CAD: LDL Cholesterol Level (<130mg/dL and <100mg/dL)	NCQA	X
CAD: LDL Cholesterol Level (<130mg/dL)	CMS	X
CAD: Ace Inhibitor/ARB Therapy (diabetes or LVSD)	PCPI/ACC/AHA	X
CAD: Symptoms and Activity Assessment	PCPI/ACC/AHA	X
CAD: Smoking Cessation (queried one or more times)	PCPI/ACC/ AHA	X (b)
CAD: Smoking Cessation (intervention)	PCPI/ACC/AHA	X (b)
Heart Failure (HF): LVF Assessment	PCPI/ACC/AHA	X

MEASURES	SOURCE	STATUS 11/05*
HF: Left Ventricular Ejection Fraction Testing	CMS	
HF: Weight Measurement	PCPI/ACC/AHA	X
HF: Blood Pressure Measurement	PCPI/ACC/AHA	
HF: Blood Pressure Measurement (distribution)	PCPI/ACC/AHA	
HF: Patient Education	PCPI/ACC/AHA	Did not pass
HF: Beta Blocker Therapy	PCPI/ACC/AHA	X
HF: Ace Inhibitor Therapy	PCPI/ACC/AHA	X
HF: Warfarin Therapy for patients with Atrial Fibrillation	PCPI/ACC/AHA	X
HF: Assessment of Clinical Symptoms of Volume Overload	PCPI/ACC/AHA	X
HF: Assessment of Activity Level	PCPI/ACC/AHA	X
HF: Assessment of Clinical Signs of Volume Overload	PCPI/ACC/AHA	
HF: Examination of the Heart	PCPI/ACC/AHA	
HF: Laboratory Tests	PCPI/ACC/AHA	
Hypertension		
Blood Pressure Management	PCPI/ACC/AHA	
Blood Pressure Measurement	PCPI/ACC/AHA	
Blood Pressure Control (BP≤140/90 mm Hg)	NCQA	
Blood Pressure Control (BP<140/90 mm Hg	CMS/NCQA	X
Plan of Care	PCPI/ACC/AHA	X
Prenatal Care		
Prenatal Flow Sheet	PCPI	
Blood Groups and Antibody Testing	PCPI	
Anti-D Immune Globulin	PCPI	X
Screening for Congenital Anomalies --patients ≥ 35	PCPI	
Screening for Congenital Anomalies --patients ≥ 35 (amniocentesis or CVS)	PCPI	
Screening for Gestational Diabetes	PCPI	
Cervical Cytology	PCPI	
Screening for HIV	PCPI	X
Screening for Asymptomatic Bacteriuria	PCPI	
Prevention, Immunization and Screening		
Tobacco Use	PCPI	X (b)
Tobacco Cessation	PCPI	X (b)
Advising Smokers to Quit Discussing Smoking Cessation Medication Discussion Smoking Cessation Strategies	NCQA NCQA NCQA	X (c) X (c) X (c)
Problem Drinking	PCPI	
Discuss Urinary Incontinence Receiving Urinary Incontinence Treatment	NCQA NCQA	X (b) X (b)
Influenza Vaccination for Older Adults (65+) Influenza Vaccination for Adults (50-64)	CMS/NCQA NCQA	X (b) X (b)

MEASURES	SOURCE	STATUS 11/05*
Influenza Vaccination	PCPI	X
Pneumonia Vaccination	CMS/NCQA	X
Pneumonia Vaccination	PCPI	
Childhood Immunization Status	NCQA	X
Adolescent Immunization Status	NCQA	
Breast Cancer Screening	CMS/NCQA	X
Breast Cancer Screening	PCPI	
Colorectal Cancer Screening	NCQA	X
Colorectal Cancer Screening	PCPI	
Cervical Cancer Screening	NCQA	X
Chlamydia Screening in Women	NCQA	

* X indicates approved
a. Approved by members but deferred by Board of Directors
b. Measure pair
c. Measure triad

In May 2005, The Ambulatory Care Quality Alliance (AQA), a broad-based coalition of health care agencies and organizations including the American Academy of Family Physicians, the American College of Physicians, America's Health Insurance Plans, CMS and AHRQ, announced a "starter set" of 26 clinical performance measures for ambulatory care with the goal of eventually establishing uniform performance measurement standards addressing both quality and patient safety issues. The initial set is a subset of the clinical measures submitted to NQF for approval and includes the two new HEDIS efficiency measures introduced for 2006. The measure set was recommended as an ambulatory care starter set by the IOM in their report, *Performance Measurement: Accelerating Improvement* (IOM 2005). The AQA views this as an initial step in a process that will include the introduction of additional efficiency measures, sub-specialty measures, cross-cutting measures and patient experience measures. These measures are derived from either a combination of administrative and medical chart data or medical chart data only. The AQA Measures are listed in Table 7.

Additionally, The Accreditation Institute for Ambulatory Health Care Institute for Quality Improvement has developed two measures of quality for colonoscopy including the intra-procedure complication rate and patient education.

Table 7
AQA Clinical Performance Measures

Prevention Measures	**SOURCE**
Breast Cancer Screening	NCQA
Colorectal Cancer Screening	NCQA
Cervical Cancer Screening	NCQA
Tobacco Use	PCPI
Advising Smokers to Quit	NCQA
Influenza Vaccination (ages 50-64)	CMS/NCQA
Pneumonia Vaccination	PCPI
Coronary Artery Disease	
Drug Therapy for Lowering LDL Cholesterol	CMS/PCPI
Beta-Blocker Treatment After Heart Attack (w/in 7 days of discharge)	NCQA
Beta Blocker Therapy-Post MI (6 months post discharge) Note: Measure not reviewed by NQF and therefore not approved	PCPI
Heart Failure	
ACE Inhibitor/ARB Therapy	PCPI/CMS
LVF Assessment	PCPI/CMS
Diabetes	
HbA1C Management	NCQA
HbA1C Management Control	NCQA
Blood Pressure Management	NCQA
Lipid Measurement	NCQA/PCPI
LDL Cholesterol Level (<130mg/dL)	NCQA
Eye Exam	NCQA
Asthma	
Use of Appropriate Medications	NCQA
Pharmacologic Therapy	PCPI
Depression	
Antidepressant Medication Management-Acute Phase	NCQA
Antidepressant Medication Management-Continuation Phase	NCQA
Prenatal Care	
Screening for HIV	PCPI
Anti-D Immune Globulin	PCPI
Quality Measures Addressing Overuse or Misuse	
Appropriate Treatment for Children with Upper Respiratory Infection (URI)	NCQA
Appropriate Testing for Children with Pharyngitis	NCQA

An additional measurement set, the National Diabetes Quality Improvement Alliance Performance Measurement Set, is a subset of twenty of the diabetes measures submitted to NQF. This group's mission is to identify, maintain and promote a single set

of performance measures for diabetes care. The Alliance includes CMS, AHRQ, JCAHO, NCQA, AMA, CDC, American Diabetes Association, American College of Physicians, American Academy of Family Physicians, The Endocrine Society, the U.S. Department of Veteran Affairs, National Institute of Diabetes, Digestive and Kidney Diseases, and American Association of Clinical Endocrinologists. The measure set was approved and released in 2003. Measures are listed in Table 8.

Table 8
National Diabetes Quality Improvement Alliance Performance Set for Adult Diabetes

Measure	**HEDIS Reporting**
A1c Management –Screen (one or more tests)	X
A1c Management –Control (poor control)	X
Blood Pressure Management (<140/90mmHg)	X
Lipid Management (at least one LDL)	X
Lipid Management (at least one lipid profile or all component tests)	
LDL Cholesterol level <130mg/dL	X
Urine Protein Screening (at least one test for microalbumin)	X
Any test for microalbumin	
Urine Protein Screening (no urinalysis or urinalysis with negative or trace urine protein--any test for microalbumin)	
Eye Examination (fundoscopic photo with interpretation by ophthalmologist or optometrist)	
Eye Examination (dilated exam by ophthalmologist or optometrist)	
Eye Examination (dilated exam or evaluation of retinal photographs if patient is at low risk for retinopathy)	X
Foot Examination (at least one)	X
Foot Examination (at least one-visual, sensory exam, pulse exam)	
Smoking Cessation (status ascertained and documented annually)	X
Smoking Cessation (assessed for smoking status)	
Smoking Cessation (Patients recommended or offered counseling or pharmacologic intervention)	
Aspirin use	
Influenza Immunization (received during recommended calendar period)	
Influenza Immunization (received or refused)	

RAND Quality Measurement Instruments

RAND has developed a number of measurement sets to assess the quality of care for various conditions and populations. The RAND QA Tools is a comprehensive system for assessing care delivered in 46 clinical areas using over 400 measures. The indicators are primarily process measures and cover care for women; children and adolescents; general medicine; oncology and HIV; and cardiopulmonary conditions. They were developed through a review of the literature and reviewed by expert panels utilizing a modified Delphi method. The RAND QA Tools were pilot-tested by Humana, Inc to develop quality of care scores for physicians in ten specialties using a software scan to analyze claims data.

In order to meet the need for quality of care information for older adults, RAND, in collaboration with Pfizer, Inc, created a quality of care assessment system entitled Assessing Care of Vulnerable Elders (ACOVE). Vulnerable elders are defined as those community dwelling individuals, 65 years of age or older, who are at moderate to high risk for functional decline or death over the next two years. To develop the system, a national panel of geriatrics experts identified the medical conditions prevalent among older adults that contribute most to morbidity, mortality and functional decline; that could be measured; and for which effective methods of treatment are available. The advisory committee selected 22 topics including diseases, syndromes, physiological impairments, and clinical situations relevant to this population. The RAND team then developed a quality indicator set containing 236 explicit process measures covering the 22 topics. The indicators and supporting literature were reviewed by independent panels of experts as well as the American College of Physicians-American Society of Internal Medicine Aging Task Force prior to adoption. The resulting quality indicators address four domains of care: prevention (26 percent), diagnosis (21 percent), treatment (36 percent) and follow-up (17 percent). They address issues of communication between the provider and the patient or the patient's proxy as well as detection and treatment of conditions that are under detected in the elderly such as dementia, depression and functional impairments. The QA and the ACOVE indicators were developed for use at the health plan level, but many can be applied to physicians or physician groups.

CMS Physician Performance Measurement Sets

CMS has developed several initiatives addressing the quality of care across various health care settings. The End Stage Renal Disease (ESRD) Clinical Performance Measures project, The Physician Group Practice Initiative, and, most recently, the

Physician Voluntary Reporting program are each using different measure sets to gauge the quality of physician care being delivered to Medicare patients.

The Physician Voluntary Reporting Program, announced October 28, 2005, initially invited physicians to report on a set of 36 quality measures addressing heart disease, diabetes, end stage renal disease, osteoporosis, depression, surgery, and prevention/screening. Following consultation with a number of physician organizations, CMS announced in December that the number of measures would be reduced to a starter set of 16. This is intended to reduce the reporting burden for physicians and better align the measures with other quality measurement programs. The primary care measures are NQF-endorsed and part of the AQA starter set with the exception of the ACOVE measures of falls for elderly patients. Further, they are measures that will be used by the Quality Improvement Organizations in CMS' 8th scope of work. Measures were also collected from HEDIS, ACOVE, the National Voluntary Consensus Standards for Cardiac Surgery and the CMS Premier program. Reporting begins in January 2006, and data will be gathered though claims supplemented with a set of HCPCS codes ("G-codes") as an interim step until physicians are able to submit data electronically through electronic medical records. Physicians who participate will receive feedback on their performance and will provide input to CMS on the measures and the process. Several of these measures, such as "aspirin at arrival" and "beta blocker at arrival" for heart attack patients are hospital-based care measures that will be reported by individual physicians for the first time. Confidential reports will be provided to physicians on the 16 measures, while CMS further develops the additional 20 measures as well as measures suggested by physician groups. Initial measures are located in Table 9.

Table 9
CMS Physician Voluntary Reporting Program Measures

MEASURE	SOURCE
1. Aspirin at arrival for AMI	CMS
2. Beta blocker at time of arrival for AMI	CMS
3. Hemoglobin A1c control in patient with Type I or II diabetes, age 18-75	NCQA
4. Low-density lipoprotein control in patient with Type I or II diabetes age 18-75	NCQA
5. High blood pressure control in patient with Type I or II diabetes, age 18-75	NCQA
6. Angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker therapy for left ventricular systolic dysfunction	PCPI
7. Beta-blocker therapy for patients with prior myocardial infarction	PCPI
8. Assessment of elderly patients for falls	ACOVE
9. Dialysis dose in end stage renal disease patients	ESRD

10. Hemocrit level in end stage renal disease patients	ESRD
11. Receipt of autogenous aterovenous fistula in end-stage renal disease patient requiring hemodialysis	ESRD
12. Antidepressant medication during acute phase for patient diagnosed with new episode of major depression	NCQA
13. Antibiotic prophylaxis in surgical patient	CMS
14. Thromboembolism prophylaxis in surgical patient	CMS
15. Use of internal mammary artery in CABG	CMS
16. Pre-operative beta blocker for patient with isolated CABG	STS

The ESRD measure set was developed in response to a mandate in the Balanced Budget Act to develop a method to measure and report the quality of renal dialysis services received by Medicare beneficiaries by January 1, 2000. The clinical performance measures (CPMs) are based on the National Kidney Foundation's (NKF) Kidney Disease Outcomes Quality Initiative (KDOQI) Clinical Practice Guidelines (formerly known as Dialysis Outcome Quality Initiative). The sixteen measures are divided between hemodialysis (HD) adequacy, peritoneal dialysis (PD) adequacy, anemia management and vascular access. Additional measures for kidney transplant referral and ESRD bone disease are under development. While the measures are currently used at the facility level, they could be applied at the physician level as well. A list of the current ESRD measures is located in Table 10.

Table 10
CMS ESRD Performance Measure Set

Adequacy of Hemodialysis (HD*)*
Monthly measurement of delivered HD dose
Method of measurement of delivered HD
Minimum delivered HD dose
Method of post-dialysis blood urea nitrogen sampling
Baseline total cell volume measurement of dialysis intended for reuse
Adequacy of Peritoneal Dialysis (PD)
Measurement of total solute clearance at regular intervals
Calculate weekly Kt/V urea and Creatinine Clearance in a standard way
Delivered dose of PD
Vascular Access
Maximizing placement of arterial venous fistulae
Minimizing use of catheters as chronic dialysis access
Preferred/non-preferred location of HD catheters located above the waist
Monitoring arterial venous grafts for stenosis

Anemia Management
Target hemotocrit/hemoglobin for Epoetin therapy
Assessment of iron stores among anemic patients or patients prescribed Epoetin
Maintenance of iron stores - target
Administration of supplemental iron

Veteran's Health Administration (VHA) Performance Measurement System

The VHA measures system performance in the areas of cancer screening, cardiovascular care, diabetes, infectious disease, mental health, tobacco counseling and treatment, and clinic waiting times. Some of the measures in the set are applicable to hospital or other VA facility performance, particularly those related to ischemic heart disease and mental health. Additionally, some of the measures are applied to sub-groups such as spinal cord injured or homeless veterans. We have listed only those measures that could potentially be attributed to physicians or physician groups in Table 11.

Table 11
VHA Performance Measurement System (Physician/Clinic Measures)

Cancer
Breast cancer screening
Cervical cancer screening
Colon cancer screening
Cardiovascular
Weight monitoring prior to admission for heart failure
ACEI or ARB prior to admission for heart failure
Poor blood pressure control for patients with hypertension
Blood pressure control for patients with hypertension
Smoking cessation counseling during hospital stay for patients with AMI
Full lipid panel and LDL-C in last two years for patients with previous AMI
Diabetes
Retinal exam by eye care specialist
Foot exam using monofilament
Full lipid panel in prior 2 years and LDL-C <120mg/dl
HgbA1c > 9 or not done
Blood pressure $\geq$ 160/100 or not done in prior year
Blood pressure $\leq$ 140/90
Infectious
Patients admitted to the hospital for community acquired pneumonia who had a pneumococcal immunization prior to admission
Patients admitted to the hospital for community acquired pneumonia who had an

influenza immunization in preceding period
Influenza immunization in prior year
Pneumococcal immunization
Mental Health
Medication for depression during acute phase
Follow-up after prescription for medication during acute phase
High risk mental health patients screened for intensive case management
Screening for alcohol misuse
Tobacco Use
Patients using tobacco in prior year
Tobacco use counseling
Waiting Times
Appointments when requested
Waiting time for provider < 20 minutes

Institute for Clinical Systems Improvement (ICSI) Physician Measures

ICSI, an independent non-profit collaborative, is comprised of 54 health care organizations primarily located in Minnesota. The organization includes medical groups, hospitals and health plans representing 7500 physicians. ICSI has developed quality measures from clinical guidelines addressing prevention and treatment of a broad range of conditions for primary and specialty care. Those measures applicable to physicians are located in Table 12.

Table 12
ICSI Physician Performance Measures

Acute Pharyngitis
Strep Screen
Strep Screen with Viral Upper Respiratory Infection
Acute Sinusitis in Adults
Sinus X ray
First line antibiotic prescribed at office visit
Adult Low Back Pain
% Patients receiving X rays
Ankle Sprain
Documentation of patient education
X ray within 3 days
Assessment and Management of Acute Pain
% Patients who rate pain >4 (on 10 Point scale) 48 hours after admission or procedure (inpatient)

Atrial Fibrillation
Warfarin for patients at risk for thromboembolism
Breast Cancer Treatment
Clinical trial offered
Cervical Cancer Screening
At least 1 Pap smear in past 3 years
Up to date for cervical cancer screening
Screening within 6 months of reminder
Chronic Obstructive Pulmonary Disease
Smoking cessation inquiry
Colorectal Cancer Screening
Counseling on screening
Up-to-date on screening
Community Acquired Pneumonia (Adults)
Chest X-ray to confirm diagnosis
Diagnosis and Management of ADHD in Primary Care (Children and Adolescents)
Follow-up visits for patients on medication
Discussion of school based supports and educational service options
DSM IV or DSM-PC criteria discussed
Diagnosis and Management of Basic Infertility
Both partners assessed
Recommended tests prior to laparoscopies or tubal surgery
Diagnosis and Outpatient Management of Asthma
Adults on inhaled corticosteroids medication
Children on inhaled corticosteroids medication
Patient Education documented
Spirometry or peak flow meter reading at last visit
Diagnosis and Treatment of Adult Degenerative Joint Disease of the Knee
Knee X ray including standing view of knee
Patient education
Diagnosis and Treatment of Headache
Migraines-patient education
Migraines-treatment plan
Diagnosis and Treatment of Otis Media in Children
Follow-up visit for children < 5
Medication
Patient/Caretaker education
Diagnosis of Breast Disease
Time (<14days) from discovery of abnormality to biopsy
Time (<14 days) from abnormal mammogram to biopsy
Hypertension Diagnosis and Treatment:
Patient education
Blood pressure control

Immunizations Measures
Adolescents up-to-date with recommended immunizations
Two year olds up-to-date –primary series
Young adults up-to-date –Hepatitis B
Lipid Management in Adults
Patients on lipid-lowering medication who have fasting lipid panel every 3-12 months
Diet evaluation for patients with CHD
Lipid Screening in Adults:
Cholesterol test in last 5 years
Exercise and nutritional assessment
Lipid Screening in Children and Adolescents
At-risk children receiving serum cholesterol level
Exercise and nutritional assessment for children with family history of heart disease
Major Depression in Adults for Mental Health Care:
Assessment (PHQ <5, Hamilton <7) within 6 months of treatment
Results of assessment show 50% decrease within 6 months of treatment
Documentation of DSM IV criteria within 3 months of diagnosis
Major Depression in Adults in Primary Care
Follow-up after treatment initiation
Assessment (PHQ <5, Hamilton <7) within 6 months of treatment
Results of assessment show 50% decrease within 6 months of treatment
Reassessment within 3 months of beginning treatment
Screening patients with fatigue
Documentation of DSM IV criteria within 3 months of diagnosis
Management of Initial Abnormal Pap Smear
Clinical follow-up within 6 months
Management of Type 2 Diabetes Mellitus
Frequency of LDL values
A1C <7%
Screen for A1C in last 6 months, annual LDL test, A1C<7%, blood pressure control, no tobacco, regular aspirin use
A1C measured in last 6 months
Eye exam
Microalbumin
Menopause and Hormone Therapy
Bone mineral testing after hormone therapy
Preoperative Evaluation
Electrocardiograms for patients 40-54
Preoperative health assessment
Preterm Birth Prevention
Interventions for risk factors
Patient education
Screening for risk factors
Prevention and Management of Obesity

Patient counseling/education
Preventive Counseling and Education (tobacco, nutrition, etc)
Preventive Services for Adults
Up-to-date for 10 key preventive services
Preventive Services for Children and Adolescents
Up-to-date for 10 key preventive services
Rhinitis
Prophylactic medication
Routine Prenatal Care
Interventions for risk factors
Counseling/Education
Education for VBAC eligible women
Stable Coronary Artery Disease
Aspirin use
Lipid profile
Tobacco Use Prevention and Cessation-Adults and Mature Adolescents
Counseling
Documentation of use/nonuse in chart
Tobacco Use Prevention and Cessation-Infants, Children, Adolescents
Counseling
Documentation of use/nonuse in chart
Uncomplicated Urinary Tract Infection in Women
Patient satisfaction
Recommended short course therapy
Urine culture at initial encounter
Vaginal Birth After Cesarean
Patient education
Venous Thromboembolism
Low molecular weight heparin (LMWH) eligible patients treated in outpatient setting
LMWH use
Viral Upper Respiratory Infection
Patient education
Antibiotic use
Office visits (for patients with symptoms < 7 days)

A summary table of common measures for physicians in ambulatory settings and the sources of the measures follows (Table 13). While the overlapping measures are addressing the same issues, the specifications may differ.

Table 13
Common Physician Measure Types

Measure Type	NCQA	AMA/ PCCI	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
Asthma/Upper Respiratory									
• Asthma Assessment		X							X
• Pharmacologic Therapy	X	X			X	X		X	X
• Spirometry/peak flow meter					X	X			
▪ Pharyngitis Screening	X				X				
▪ Appropriate Treatment for Asthma	X							X	
Depression/Behavioral Health									
• Follow-up after diagnosis and/or treatment	X			X	X	X	X		
• Medication during acute phase	X			X		X	X	X	X
• Medication during continuation phase	X								X
• Suicide risk assessment		X				X	X		
• Screening for alcohol misuse				X		X	X		
Bone Conditions									
• Osteoarthritis—OTC medications		X				X	X		X
• Osteoarthritis-Exercise recommended		X				X	X		
• Osteoarthritis-Functional and pain assessment		X				X	X		X

Measure Type	NCQA	AMA/ PCCI	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
Diabetes									
• A1C Screen	X	X		X	X	X	X	X	X
• A1C Control	X	X		X	X	X	X	X	X
• Blood Pressure Control	X	X		X	X		X	X	X
• Lipid Screen	X	X		X	X	X	X	X	
• LDL Control	X	X		X	X		X	X	X
• Urine Protein Screening	X	X			X	X	X		X
• Eye Exam	X	X		X	X	X	X	X	X
• Foot Exam	X	X		X		X	X		X
• Smoking Status	X	X			X				X
• Smoking Cessation		X			X				
• Aspirin Use		X			X		X		
Heart Disease									
• CAD: Antiplatelet therapy		X			X	X	X		X
• CAD: Drug Therapy for Lowering LDL Cholesterol		X	X					X	X
• CAD: Beta Blocker for prior AMI patients	X	X				X		X	
• CAD: Lipid Profile	X	X			X				X
• CAD: LDL Cholesterol Level	X		X				X		X
• CAD: Smoking Cessation		X				X	X		X
• COPD: Smoking Cessation					X	X			

Measure Type	NCQA	AMA/ PCCI	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
• HF: LVF Testing		X	X			X	X	X	X
• HF: Weight Measurement		X		X		X	X		X
• HF: Blood Pressure Measurement		X				X	X		
• HF: Examination of the Heart		X				X	X		
• HF: Patient Education		X				X	X		
• HF: Ace Inhibitor Therapy		X		X		X	X	X	X
• HF: Warfarin for Atrial Fibrillation		X			X	X			X
• HF: Lab tests		X				X	X		
Hypertension									
• Blood Pressure Measurement		X				X	X		
• Blood Pressure Control	X	X		X	X				X
• Patient Education					X	X	X		
Prenatal Care									
• Screening for HIV		X				X		X	X
• Anti-D Immune Globulin		X						X	X
• Blood Groups and Antibody Screen		X				X			
• Blood Pressure Measurement		X				X			
• Counseling and Education					X	X			
• Screening for Gestational Diabetes		X				X			

Measure Type	NCQA	AMA/ PCCI	CMS	VHA	ICSI	RAND QA	ACOVE	AQA	NQF Approved
• Cervical Cytology		X				X			
Prevention, Immunization, Screening									
• Tobacco Use		X		X	X	X	X	X	X
• Tobacco Counseling	X	X		X	X	X	X	X	X
• Screen for Alcohol Misuse		X		X	X	X	X		
• Influenza Vaccination	X	X	X	X	X	X	X	X	X
• Pneumonia Immunization	X	X		X	X	X	X	X	X
• Childhood Immunization Status	X				X		X		
• Adolescent Immunization Status	X				X				
• Breast Cancer Screening	X	X	X	X	X	X		X	
• Colorectal Cancer Screening	X	X		X	X	X	X	X	
• Cervical Cancer Screening	X			X	X	X		X	
• Chlamydia Screening in Women	X				X				
• Drugs to be Avoided in the Elderly	X						X		
• Discussion of Urinary Incontinence	X						X		X
Urinary Tract Infection									
• Urine Culture					X	X			
• Treatment					X	X			

Commercial Insurance Company Measurement Sets

Pacificare is one of many health plans that has developed its own set of quality measures for physicians, some of which are derived from the HEDIS measure set and some of which are developed internally. Pacificare's Quality Index utilizes 50 measures of clinical and service quality for physicians in the following categories: Staying Healthy, Appropriate Care, Patient Safety, Patient Satisfaction, Complaints/Transfers, Affordability and Administrative.

State Data Organization Measures

Several state organizations collect and make public measures of physician performance. For example, The Pennsylvania Healthcare Cost Containment Council compiles data on hospital and physician performance for total hip and knee replacements utilizing measures of deep joint infection or device problems, blood clot rate, wound infection rate, readmission rate and average post operative length of stay. The organization also collects and publicizes readmission and mortality rates by hospital and physician for CABG. New York State publicly reports outcome measures for CABG and coronary angioplasty at the hospital and physician levels.

Medical Specialty Measures

Numerous specialty societies have been involved in the development of physician performance measures either alone or in conjunction with other groups through a consensus process. However, the number of measures and consensus around these measures still lags that of primary care. Specialty measures are often derived from established guidelines within the specialty. For example, the American Academy of Ophthalmology worked with NCQA to develop a performance measure for glaucoma screening for HEDIS 2006 and contributed to the development of the HEDIS diabetes eye exam measure. In another specialty area, NQF established the Cancer Project to address the multiple and sometimes conflicting measures of cancer care. The group identified the priority areas of breast, colorectal and prostate cancers and the cross cutting priority areas of access and competence, communication/coordination/IT, prevention and screening, symptom management and end-of-life care. Technical panels were convened to review candidate measures and make recommendations; and a review of the proposed measures is underway. In the case of surgery, measures of infection prevention included in the Surgical Care Improvement Project (SCIP) are applicable across many surgical

specialties. A summary of the status of measure development or approval for various specialties follows in Table 14.

Table 14
Specialty Performance Measure Status

Specialty Society	**Type of Measures**	**Status of Measures**
Anesthesiology	Some SCIP measures, appropriate evaluation of the patient	SCIP measures approved, In Use
Emergency Medicine	Hospital JCAHO/CMS measures ie. Aspirin and beta blocker treatment at arrival for AMI	NQF Approved, In Use
Gastroenterology (AGA)	Quality/access to care, customer satisfaction	In development
Internal Medicine-Cardiology (ACC)	HF, AMI and CAD measures (JCAHO/CMS) Additional measures in development	Many submitted and approved by NQF, In Use
Internal Medicine-Neurology (American Academy of Neurology)	Appropriate treatment of stroke, stroke rehabilitation, diagnosis of dementia	Developed 1995-2005
Internal Medicine-Hema-Oncology	Patient experience of care (pain, nausea, fatigue)	Clinical measures in development
Internal Medicine-Nephrology	Renal Physicians Association Clinical Performance Measures on Appropriate Patient Preparation for Renal Replacement Therapy (Table 11)	Developed 2002
Internal Medicine-Pulmonology	COPD measures	In development
Internal Medicine-Rheumatology	Osteoarthritis	Several approved by NQF
Radiology (American College of Radiology)	Appropriateness of tests, communication of results	In development
Ophthalmology (American Academy of Ophthalmology)	Glaucoma screening, diabetes eye exam (HEDIS)	Eye Exam approved by NQF
Surgery (ACS)	National Surgical Quality Improvement Program (NSQIP), Surgical Care Improvement Project (SCIP)	Some NQF approved, In use
Surgery-Orthopedic (AAOS, PCCI)	Osteoarthritis of the Knee (Bone Conditions) Additional measures in development	Submitted to NQF-2 approved

Specialty Society	Type of Measures	Status of Measures
Surgery-Thoracic/Cardiac (STS)	Society of Thoracic Surgeons Cardiac Surgery Measures	NQF endorsed
Psychiatry	Depression measures	NQF approved several
Pathology	Appropriateness of tests, communication of results	In development
Obstetrics/Gynecology (ACOG)	Mammography, Cervical Cancer screening, etc	Submitted to NQF, some approved
Oncology (Quality Oncology Practice Initiative)	Indicators of quality care within an office or practice setting	National roll-out end 2005

Table 15
Renal Physicians Association Clinical Performance Measures on Appropriate Patient Preparation for Renal Replacement Therapy 2002

1. Counseling for increased physical activity
2. Patient Education
3. Referrals to vocational rehabilitation center
4. Discussion of renal replacement therapy (RRT) modalities
5. Referrals for surgery for construction of AV fistula on index date
6. Screening for dyslipidemia within 1 year
7. Lipid lowering treatment
8. LDL less than 100 mg/dL
9. Serum bicarbonate measured within last 3 months
10. Serum calcium and phosphorous measured within the last 3 months
11. Measurement of iPTH
12. iPTH measured within last 3 monts for patients on a phosphate binder
13. iPTH greater than 100 pg/mL
14. Phosphorus greater than 4.5 mg/dL
15. 25(OH) vit D levels measured
16. Elemental calcium prescribed
17. Blood pressure checked within last 3 months
18. Phosphorus greater than 4.5 mg/dL after a low phosphorus diet for one month , now on a phosphate binder
19. Antihypertensive therapy intensified
20. Blood pressure less than 130/80 mmHg on index date

Measurement of Patient Experience with Care

The most frequently utilized measures of patient experience are drawn from the Consumer Assessment of Health Care Providers and Systems (CAHPS) instrument. While the initial focus of CAHPS was on health plans, variations of the instrument are used to measure patients' experiences with physician groups and individual physicians. To meet the demand to measure patient experience with physicians at a local and national level, a new Clinicians and Group Survey was developed by the CAHPS Consortium with AHRQ funding and will be available in the spring of 2006. This new Ambulatory version of CAHPS, A-CAHPS, will go through NQF review in the Spring of 2006 and has the potential to become the new national standard. The topics addressed in the new survey are similar to those in the health plan survey: access to care, coordination of care, doctor's communication and thoroughness, shared decision-making, health promotion and education, follow-up on test results, medical office staff, patient concerns about cost of care, and a global rating of one's doctor. The IOM recommended the use of this tool for the ambulatory care setting (IOM 2005).

Measures of Physician Resource Use and Cost-Efficiency

Efficiency is a relatively new realm of measurement for physicians and physician organizations with few standardized measures. Most cost-efficiency measures use the physician as the unit of analysis to compare the resources utilized by the physician delivering care relative to peers (IOM 2005). Data sources for measurement typically include encounter and claims data. Comprehensive efficiency measures require complex methods to adjust for differences in patient demographics and morbidity, service demand and prices to accurately measure resource use and make comparisons across physicians or hospitals. Episode Treatment Groups (ETGs) are often used for this purpose and various commercial vendors offer thee tools including Symmetry (ETGs), Medstat (MEGs), Health Dialog, and the Care Marketbasket System. Individual measures of resource use commonly utilized for physicians are inpatient admissions and inpatient days per 1000.

D. COMPARISON OF DIABETES MEASURES' DESCRIPTIONS FROM SELECTED MEASURE SETS

Measure	Source	Specifications
HbA1c Management (Screen)	NCQA	Percentage of patients 18-75 receiving one or more A1c test(s) in the measurement year[1]
	AMA	Percentage of patients receiving one or more A1c test(s) in the measurement year
	AMA	Distribution of number of tests done (0,1,2,3, or more)
	ICSI	Percentage of patients who had a screen for A1c in the past 6 months
	RAND	Patients with diabetes should have a glycosylated hemoglobin or fructosamine every 6 months
HbA1c Management (Control)	NCQA	Percentage of patients 18-75 with most recent A1c level in the measurement year > 9% (poor control)[2]
	AMA	Distribution of most recent A1c value by range: < 6.0, 6.1-7.0, 7.1-8.0, 8.1-9.0, 9.1-10.0, >10.0, undocumented
	ICSI	Percentage of patients with A1c value < 7.0%
	VHA	HGBA1c > 9 or not done
Blood Pressure Management	NCQA	Percentage of patients 18-75 with most recent blood pressure < 140/80 mm Hg[3]
	AMA	Percentage of patients who had their blood pressure documented in the past year < 140/90 mm Hg[4]
	AMA	Distribution of most recent blood pressure values by range: Systolic: <120, 120-129, 130-139, 140-149, 150-159, 160-169, 170-179, >180, undocumented Diastolic: <75, 75-79, 80-89, 90-99, 100-109, >110, undocumented
	ICSI	Percentage of patients with blood pressure < 130/80 mm Hg
	VHA	Percent of patients with blood pressure $\geq$ 160/100mm Hg or not done in prior year
	VHA	Percent of patients with blood pressure $\leq$ 140/90mmHg
Lipid Management	NCQA	Percentage of patients 18-75 with at least one low density lipoprotein cholesterol (LDL-C) test in the measurement year[5]
	AMA	Percentage of patients who received at least one lipid profile (or all component tests)

	ICSI	Percentage of patients who had an annual low-density lipoprotein (LDL) test
	VHA	Percent of patients with full lipid profile in prior 2 years
	RAND	Patients with diabetes should have total cholesterol tested annually
LDL Cholesterol Level	NCQA	Percentage of patients 18-75 with most recent LDL-C in the measurement year < 130mg/dL[6]
	NCQA	Percentage of patients 18-75 with most recent LDL-C in the measurement year < 100mg/dL[7]
	AMA	Distribution of most recent test values by range: Total cholesterol: >240, 200-239, <200, undocumented; LDL-C: > 160, 130-159, 100-129, <100, undocumented; IF non-HDL cholesterol is reported, record the test values in the following ranges: ≥190, 160-189, 130-159, <130, undocumented; Triglycerides: >400, 200-399, <200, 150-199, <150, undocumented
	ICSI	Percentage of patients with LDL < 100
	VHA	Percent of patients with LDL-C < 120mg/dL
Urine Protein Screening	NCQA	Percentage of patients 18-75 with at least one test for microalbumin during the measurement year or who had evidence of medical attention for existing nephropathy (diagnosis of nephropathy or documentation of microalbumin or albumuria)[8]
	AMA	Percentage of patients with no urinalysis or urinalysis with negative or trace urine protein who received a test for microalbumin
	ICSI	Percentage of patients with microalbumin tested within the last 12 months
Eye Examination	NCQA	Percentage of patients 18-75 who received a dilated eye exam or seven standard field stereoscopic photos with interpretation by an ophthalmologist or optometrist or imaging validated to match diagnosis from these photos in the reporting year or during the prior year if the patient is at low risk for retinopathy[9]
	AMA	Percentage of patients receiving a dilated retinal eye exam by an ophthalmologist or optometrist
	ICSI	Percentage of patients with an eye exam documented within last 12 months
	VHA	Percent of patients who have had a retinal eye exam by an eye care specialist in a specified time period
Foot Examination	NCQA	Percentage of patients 18-75 receiving at least one foot exam, defined in any manner in the

		measurement year[10]
	AMA	Percentage of eligible patients who received at least one complete foot exam (visual inspection, sensory exam with monofilament and pulse exam)
	VHA	Percent of patients having a foot exam using a monofilament within the past year
	RAND	Patients with a diagnosis of diabetes should have examination of feet at least twice a year

Measures also used in the following programs:
[1] ACQA, IHA
[2] ACQA, Bridges to Excellence, PVRP, IHA
[3] Bridges to Excellence, PVRP
[4] ACQA
[5] ACQA, IHA
[6] ACQA, Bridges to Excellence, IHA
[7] ACQA, Bridges to Excellence, PVRP
[8] Bridges to Excellence, IHA
[9] ACQA, Bridges to Excellence
[10] Bridges to Excellence

E. PERFORMANCE MEASURE SELECTION CRITERIA

A number of national organizations, including the Institute of Medicine (IOM), the National Quality Forum (NQF), the Agency for Healthcare Quality and Research (AHRQ) and the National Committee for Quality Assurance (NCQA) have established criteria to guide their own processes for selecting performance indicators (McGlynn, 200X). The criteria defined by these national entities overlap to a great extent. In particular, areas they collective emphasize are in choosing measures that will are important (i.e., low performance and/or wide variation), are methodologically sound and evidence-based, and feasible to implement, as is highlighted in Table 16. The other criteria used include impact (extent of burden of the disease), inclusiveness (relevant to a broad range of different types of people), and usability (the extent to which the intended end user can understand the results and make use of them).

Table 16
National Organizations' Criteria for Selecting Performance Measures

	Impact	Inclusiveness	Importance/ Relevance/ Improvability	Scientific Acceptance	Usability	Feasibility
AHRQ	X	X	X			
IOM			X	X	X	X
NQF			X	X	X	X
NCQA			X	X		X

Below we describe in more detail the criteria used by the organizations listed in Table 16.

Impact refers to the extent of the burden, (e.g.,, disability, mortality, and economic costs) imposed by a condition. In addition to the effect on patients, this include the effects a condition has on families, communities, and societies.

Improvability refers to the extent of the gap between current practice and evidence-based best practice, as well as the likelihood that the gap can be closed and conditions improved through change in an area. The areas of focus for change are the six national quality aims identified in the IOM's Quality Chasm report (safety, effectiveness, patient-centeredness, timeliness, efficiency and equity).

Inclusiveness encompasses the relevance of an area to a broad range of individuals with regard to age, gender, socioeconomic status, and ethnicity/ race (equity); the generalizability of associated quality improvement strategies to many types of conditions and illnesses across the spectrum of health care (representativeness); and the breadth of change effected through such strategies across a range of health care settings and providers (reach).

Importance of a measure is determined by the extent to which performance in the area (e.g., setting, procedure, condition) is poor and considerable variation in quality of care exists, both of which indicat opportunities for improvement in care .

Scientific acceptability refers to a measure having well defined specifications for the numerator and denominator that accurately represents the concept being evaluations (valid), produces the same results consistently when applied in the same population (reliable), and is precise enough to distinguish between the performance of different providers. Furthermore, the measure is adaptable to patient preferences, risk-adjustment strategy exists if applicable, and evidence exists linking structure or process measure to patient outcomes.

Usability reflects the extent to which the intended audiences (e.g., consumers, purchasers, providers) can understand the results of the measure and are likely to find them useful for decision making. This includes differences in performance levels being statistically, practically and clinically meaningful.

Feasibility considers the benefits versus the financial and administrative burden of implementing a measure. This includes consideration of how the data is collected, the auditing strategy required and confidentiality concerns.

REFERENCES

Abt Associates, "Nursing Home Pay-for-Performance Demonstration Draft Design," Open Door Forum presentation, September 20, 2005. Available at http://www.cms.hhs.gov/DemoProjectsEvalRpts/downloads/NHP4P_ Handout.pdf.

ACCF Quality Strategic Oversight Committee, "ACCF Principles to Guide Physician Pay-for Performance Programs for the Delivery of Cardiovascular Care," approved by Executive Committee September 14, 2005.

Agency for Healthcare Research and Quality, National Quality Measures Clearinghouse, http://www.qualitymeasures.ahrq.gov.

Agency for Healthcare Research and Quality, *National Healthcare Quality Report: Summary*, AHRQ Publication No. 04-RG003, Rockville, MD: AHRQ, 2003.

American Academy of Family Physicians, "Pay-for-Performance," n.d. Available at http://www.aafp.org.

American Academy of Family Physicians, "Per Family Physicians: Medicare Value Purchasing Act Misses the Mark," July 1, 2005. Available at http://www.aafp.org/x36087.xml.

American College of Physicians, "Linking Physician Payments to Quality Care," Position paper, 2005. Available at http://www.acponline.org/ hpp/thinkpay.pdf.

American Healthways and Johns Hopkins, "Outcomes-Based Compensation: Pay-for-Performance Design Principles," presented at 4th Annual Disease Management Outcomes Summit, November 11–14, 2004.

American Medical Association, "Principles for Pay-for Performance Programs." Available at http://www.ama-assn.org.

Amundson, G., L. I. Solberg, M. Reed, E. M. Martini, and R. Carlson. "Paying for Quality Improvement: Compliance with Tobacco Cessation Guidelines," *Joint Commission J. Quality and Safety*, 29:59–65, 2003.

Anderson, G., and J. Horvath, *Chronic Conditions: Making the Case for Ongoing Care*, Princeton, NJ: Robert Wood Johnson Foundation's Partnership for Solutions, 2002.

Armour, B. S., C. Friedman, M. M. Pitts, J. Wike, L. Alley, and J. Etchason, "The Influence of Year-End Bonuses on Colorectal Cancer Screening," *American J. Managed Care*, 10:617–624, 2004.

Asch, S. M., E. A. Kerr, J. Keesey, J. L. Adams, C. M. Setodji, S. Malik, and E. A. McGlynn, "Who Is at Greatest Risk for Receiving Poor-Quality Health Care?" *New England J. Medicine*, 354(11):1147–1156, 2006.

Association of Health Insurance Plans, "Rewarding Quality: Health Insurance Plan Examples," Appendix to AHIP testimony before House Ways and Means Subcommittee, March 15, 2005.

Baker, G., and B. Carter, *Provider Pay-for Performance Incentive Programs: 2004 National Study Results*, San Francisco, CA: Med-Vantage Inc, 2005.

Beckman H., A.L. Suchman, K. Curtin, R.A. Greene. "Physician reactions to quantitative individual performance reports." *American Journal of Medical Quality*, 21(3):192-9, 2006.

Berwick, D. M., "Quality of Health Care. Part 5: Payment by Capitation and the Quality of Care," *New England J. Medicine*, 335:1227–1231, 1996.

Bosworth, H. B., T. Dudley, M. K. Olsen, C. I. Voils, B. Powers, M. K. Goldstein, and E. Z. Oddone, "Racial Differences in Blood Pressure Control: Potential Explanatory Factors," *American J. Medicine*, 119(1):9–15, 2006.

Boyd, C. M., J. Darer, C. Boult, L. P. Fried, L. Boult, and A. W. Wu, "Clinical Practice Guidelines and Quality of Care for Older Patients with Multiple Comorbid Diseases: Implications for Pay for Performance," *JAMA*, 294:716–724, 2005.

Bridges to Excellence, http://www.bridgestoexcellence.org/bte.

Burt C. W., and E. Hing, "Use of Computerized Clinical Support Systems in Medical Settings: United States, 2001–03," Advance Data from Vital and Health Statistics, No. 353, Hyattsville, MD: National Center for Health Statistics, 2005.

Casalino, L. P., K. J. Devers, T. K. Lake, M. Reed, and J. J. Stoddard, "Benefits of and Barriers to Large Medical Group Practice in the United States," *Archives of Internal Medicine*, 163(16):1958–1964, 2003.

Centers for Medicare & Medicaid Services (CMS), CMS Claims and Billing Information, http://www.noridianmedicare.com/bene/claims/fraud.html (as of March 10, 2006).

Centers for Medicare & Medicaid Services (CMS), "Medicare Demonstrations," 2005. Available at http://www.cms.hhs.gov/Demo ProjectsEvalRpts/MD/list.asp#TopOfPage.

Centers for Medicare & Medicaid Services (CMS), 2004. "Hospital Quality Initiative Overview" Available at http://www.cms.hhs.gov/Quality/Hospital/Premierfactsheet.pdf.

Centers for Medicare & Medicaid Services (CMS), "Physician Voluntary Reporting Program Core Starter Set Background and General Information as of December 27, 2005." Available at http://new.cms.hhs.gov/PhysicianFocusedQualInits/.

Center for Medicare & Medicaid Services (CMS), Department of Health and Human Services (H&HS), "Medicare Program; Revisions to Payment Policies Under the Physician Fee Schedule for Calendar Year 2006 (CMS-1502-P)," *Federal Register*, August 8, 2005.

Damberg, C. L., K. Raube, and T. McGinnis, "Physician Organizations' Responses to Pay-for-Performance," Manuscript submitted for review, 2006.

DiMatteo, M. R., "Variations in Patients' Adherence to Medical Recommendations: A Quantitative Review of 50 Years of Research," *Medical Care*, 42(3):200–209, 2004.

Dudley, R. A., R. H. Miller, T. Y. Korenbrot, and H. S. Luft, "The Impact of Financial Incentives on Quality of Health Care, *Milbank Quarterly*, 76(4):649–686, 511, 1998.

Dudley, R. A., A. Frolich, D. L. Robinowitz, J. A. Talavera, P. Broadhead, and H. S. Luft, *Strategies to Support Quality-Based Purchasing: A Review of the Evidence*, Technical Review 10, AHRQ Publication No. 04-0057, Rockville, MD: Agency for Healthcare Research and Quality, July 2004.

eHealth Initiative and Foundation, "Parallel Pathways for Quality Healthcare, A Framework for Aligning Incentives with Quality and Health Information Technology," Recommendations of Working Group for Financing and Incentives, eHealth Initiative and Foundation, May 25, 2005.

Ellis, R. P., and T. G. McGuire, "Hospital Response to Prospective Payment: Moral Hazard, Selection, and Practice-Style Effects," *J. Health Economics*, 15(3):257–277, 1996.

Endocrine Insider, "More Details Emerge on AMA Deal with Congress Regarding P4P,", Endocrine Society, March 23, 2006. Available at http://www.endo-society.org/publicpolicy/insider/More_Details_ Emerge_on_AMA_Deal_Art2.cfm.

Epstein, A. M., T. H. Lee, and M. B. Hamel, "Paying Physicians for High-Quality Care," *New England J. Medicine*, 350(4):406–410, 2004.

Fairbrother, G., S. Friedman, K. L. Hanson, and G. C. Butts, "Effect of the Vaccines for Children Program on Inner-City Neighborhood Physicians," *Archives Pediatric Adolescent Medicine*, 151:1229–1235, 1997.

Fairbrother, G., K. L. Hanson, S. Friedman, and G. C. Butts, "The Impact of Physician Bonuses, Enhanced Fees, and Feedback on Childhood Immunization Coverage Rates, *American J. Public Health*, 89:171–175, 1999.

Fairbrother, G., M. J. Siegel, S. Friedman, P. D. Kory, and G. C. Butts, "Impact of Financial Incentives on Documented Immunization Rates in the Inner City: Results of a Randomized Controlled Trial," *Ambulatory Pediatrics*, 1:206–212, 2001.

Federal Register, "Medicare Care Management Performance Demonstration, Notice of a New System of Records," 70(193):58442–58446, October 6, 2005.

Federal Register, "Medicare Program; Medicare Health Care Quality (MHCQ) Demonstration Programs," 70(179):54751–54752, September 16, 2005.

Francis, D. O., H. Beckman, J. Chamberlain, G. Partridge, and R. Greene, "Introducing a Multifaceted Intervention to Improve the Management of Otitis Media: How do Pediatricians, Internists and Family Physicians Respond?" *American J. Medical Quality*, 21:134–143, 2006.

Gandhi, T. K., E. F. Cook, A. L. Puopolo, H. R. Burstin, J. S. Haas, and T. A. Brennan, "Inconsistent Report Cards: Assessing the Comparability of Various Measures of the Quality of Ambulatory Care," *Medical Care*, 40(2):155–165, 2002.

Grady, K. E., J. P. Lemkau, N. R. Lee, and C. Caddell, "Enhancing Mammography Referral in Primary Care," *Preventive Medicine*, 26:791–800, 1997.

Green, D. C., J. P. Koplan, and C. M. Cutler, "Prenatal Care in the First Trimester: Misleading Findings from HEDIS," *International J. Quality in Health Care*, 11(6):465–473, 1999.

Greene, R. A., H. Beckman, et al., "Increasing Adherence to a Community-Based Guideline for Acute Sinusitis Through Education, Physician Profiling, and Financial Incentives," *American J. Managed Care*, 10:670–678, 2004.

Guterman, S., "Medicare and Disease Management," Presentation at Third National Disease Management Summit, May 12, 2003. Available at http://www.ehcca.com/presentations/dmconference1/1_01.pdf.

Gutman J., ed., *Case Studies in Health Plan Pay-for-Performance Programs*, Washington, D.C.: Atlantice Information Services, Inc., 2004.

Heffler S., S. Smith, S. Keehan, C. Borger, M. K. Clemens, and C. Truffer, "Trends: U.S. Health Spending Projections for 2004–2014," *Health Affairs*, Web Exclusive W5-74-85, 2005.

Hibbard, J. H., J. Stockard, and M. Tusler, "Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts?" *Health Affairs*, 22(2):84–94, 2003.

Hillman, A. L., K. Ripley, N. Goldfarb, J. Weiner, I. Nuamah, and E. Lusk, "The Use of Physician Financial Incentives and Feedback to Improve Pediatric Preventive Care in Medicaid Managed Care," *Pediatrics*, 104:931–935, 1999.

Hillman, A. L., K. Ripley, N. Goldfarb, I. Nuamah, J. Weiner, and E. Lusk, "Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care," *American J. Public Health*, 88:1699–1701, 1998.

Hillman, A. L., M. V. Pauly, K. Kerman, and C. R. Martinek, "HMO Managers' Views on Financial Incentives and Quality," *Health Affairs*, 10:207–219, 1991.

Hoffer, T. P., R. A. Hayward, S. Greenfield, E. H. Wagner, S. H. Kaplan, and W. G. Manning, "The Unreliability of Individual Physician 'Report Cards' for Assessing the Costs and Quality of Care of Chronic Disease," *JAMA*, 281:2098–2105, 1999.

Institute of Medicine, *Crossing the Quality Chasm: A New Health System for the 21st Century*, Washington, D.C.: National Academy of Sciences, 2001.

Institute of Medicine, *Performance Measurement: Accelerating Improvement*, Washington, D.C.: National Academy of Sciences, 2005.

Institute of Medicine, "Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs," January 30, 2006. Available at http://www.iom.edu/?id=31317.

Joint Commission on Accreditation of Healthcare Organizations, *Joint Commission Benchmark*, 7(1), January/February 2005.

Kaplan, R. C., N. C. Bhalodkar, E. J. Brown, Jr., J. White, and D. L. Brown, "Race, Ethnicity, and Sociocultural Characteristics Predict Noncompliance with Lipid-Lowering Medications," *Preventive Medicine*, 39(6):1249–1255, 2004.

Kouides, R. W., N. M. Bennett, B. Lewis, J. D. Cappuccio, W. H. Barker, and F. M. LaForce, "Performance-Based Physician Reimbursement and Influenza Immunization Rates in the Elderly: The Primary-Care Physicians of Monroe County," *American J. Preventive Medicine*, 14:89–95, 1998.

Leapfrog Incentive and Reward Compendium, http://ir.leapfroggroup.org/ compendium (as of May 12, 2006).

Levin-Scherz, J., N. DeVita, and J. Timbie, "Impact of Pay-for-Performance Contracts and Network Registry on Diabetes and Asthma HEDIS Measures in an Integrated Delivery Network," *Medical Care Research and Review*, 63(1, Supplement):14S–28S, 2006.

Magno, L., "Medicare Demonstrations: Support for Health IT," Briefing for Capitol Hill Steering Group on Telehealth and Healthcare Informatics, May 11, 2005. Available at http://www.ehealthinitiative .org/initiatives/policy/briefings.mspx.

Mattke, S., and C. L. Damberg, "Keeping It Real: A Survey of Current Practices in Maintenance Policies for Quality of Care Measures," *International J. Healthcare Quality*, .

McClellan, M. B., Presentation given at National Pay for Performance Summit, Los Angeles, CA, February 7, 2006.

McClellan, Mark B., "Medicare Physician Payments," Testimony before House Committee on Energy and Commerce, Subcommittee on Health, November 17, 2005a. Available at http://www.cms.hhs.gov/apps/media/ ?media=testm.

McClellan, Mark B., "Value-Based Purchasing For Physicians Under Medicare," Testimony before House Ways and Means Subcommittee on Health, July 21, 2005b. Available at http://www.cms.hhs.gov/apps/ media/?media=testm.

McClellan, Mark B., "Value-Based Purchasing for Physicians Under Medicare," Testimony before House Subcommittee on Health, September 29, 2005. Available at http://www.cms.hhs.gov/apps/media/?media= testm.

McCormick, D., D. U. Himmelstein, S. Woolhandler, S. M. Wolfe, and D. H. Bor, "Relationship Between Low Quality-of-Care Scores and HMOs' Subsequent Public Disclosure of Quality of Care Scores," *JAMA*, 288(12):1484–1490, 2002.

McGlynn, E. A., S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeChristofaro, and E. A. Kerr, "The Quality of Health Care Delivered to Adults in the United States," *New England J. Medicine*, 348(26):2635–2645, 2003.

McGlynn, E. A., "Six Challenges in Measuring the Quality of Health Care," *Health Affairs*, 16:7–21, 1997.

Med-Vantage, "Pay for Performance Programs for Providers Continue to Increase, Diversify in 2005, According to Med-Vantage," http://biz.yahoo.com/bw/051117/20051117005785.html?.v=1. Accessed on 11/30/2006.

Medicare Payment Advisory Commission (MedPAC), "Report to the Congress: Medicare Payment Policy," Washington, D.C.: MedPAC, March 2005.

Medicare Payment Advisory Commission (MedPAC), "Report to the Congress: Medicare Payment Policy," Washington, D.C.: MedPAC, March 2006.

Medicare Payment Advisory Commission (MedPAC), "Medicate Value-Based Purchasing for Physicians' Services Act of 2005," HR. 3617, and "Medicare Value Purchasing Act of 2005," S. 1356, U.S. Congress, 2005.

Medicare Payment Advisory Commission (MedPAC). Public meeting transcript, September 8, 2005.

Medicare Prescription Drug, Improvement, and Modernization Act, PL 108-178. See Public Law PL 108-178, below.

Medical Group Management Association, "Principles for Pay-for Performance Programs and Recommendations for Medical Group Practices," Position paper approved by MGMA Board of Directors, February 2005. Available at http://www.mgma.com/about/MGMSppospayfor performance.cfm.

Miller, M. E., "Pay-for-Performance in Medicare," Statement of Executive Director of Medicare Payment Advisory Commission (MedPAC) before Committee on Finance, U.S. Senate, July 25, 2005.

Miller, R. H., and H. S. Luft, "Managed Care Plan Performance Since 1980: A Literature Analysis," *JAMA*, 271:1512–1519, 1994.

Morrow, R. W., A. D. Gooding, and C. Clark, "Improving Physicians' Preventive Health Care Behavior Through Peer Review and Financial Incentives," *Archives of Family Medicine*, 4:165–169, 1995.

National Business Group on Health Statement, "Congress Should Implement Medicare Pay-For-Performance Now," November 23, 2005. Available at http://nbgh.com/healthcarepolicy/payperformance.cfm.

National Committee for Quality Assurance (NCQA), http://www.ncqa.org.

National Committee for Quality Assurance (NCQA) Health Plan Report Card, "NCQA Lauds Medicare Value Purchasing Act Legislation," June 29, 2005. Available at http://www.ncqa.org/communications/news/ Medicare_legislation_statement_05.htm.

New York State Department of Health, "Adult Cardiac Surgery in New York State 1998–2000," February 2004.

Newhouse, J. P., "Do Unprofitable Patients Face Access Problems?" *Health Care Financing Review*, 11(2):33–42, 1989.

Pacific Business Group on Health, "Pacific Business Group on Health Lauds Introduction of Medicare Value Purchasing Act of 2005," June 30, 2005.

Pacific Business Group on Health, National Committee on Quality Assurance, California HealthCare Foundation in partnership with the California Medical Association, Centers for Medicare and Medicaid Services, and Lumetra, "Aligning Physician Incentives: Lessons and Perspectives from California," Report from conference of September 2005.

Parkerton, P. H., D. G. Smith, T. R. Belin, and G. A. Feldbau, "Physician Performance Assessment: Nonequivalence of Primary Care Measures," *Medical Care*, 41(9):1034–1047, 2003.

Pear, R., "AMA to Develop Measure of Quality of Medical Care; Pact with Congress Has Medicare Link," *New York Times*, February 21, 2006, p. A12.

"Physician Voluntary Reporting Program: Medicare Fact Sheet," Centers for Medicare and Medicaid Services (CMS), Department of Health and Human Services, Washington, D.C., October 28, 2005. Available at http://www.cms.gov.

Pope, G. C., M. Trisolini, J. Kautter, and W. Adamache, "Physician Group Practice (PGP) Demonstration Design Report," Health Economics Research, Inc., for Centers for Medicare & Medicaid Services, October 2, 2002.

Public Law 108-173, Medicare Prescription Drug, Improvement, and Modernization Act of 2003, December 8, 2003. Available at http://web.lexis-nexis.com/congcomp/document?_m=22cf794b25b96c61ad9 b26835801a6b0&_docnum=3&wchp=dGLbVzbzSkSA&_md5=295a95e5a9db741ab 27ae95d95e270cb.

Rewarding Results, http://www.leapfroggroup.org.

Robinson, J. C., "Strategic Choices in Pay-for-Performance Programs," presented at IHA Pay-for-Performance Summit, Los Angeles, CA, February 6, 2006.

Roland, M., "Linking 25% of UK FP's Pay to Quality of Care: A Major Experiment in Quality Improvement," presented at AcademyHealth 2005 Annual Meeting, Boston, MA, June 27, 2005.

Rosenthal, M. B., and R. G. Frank, "What Is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review*, 63(2):135–157, 2006.

Rosenthal, M. B., R. G. Frank, Z. Li, and A. M. Epstein, "Early Experience with Pay-for-Performance: From Concept to Practice, *JAMA*, 294:1788–1793, 2005.

Rosenthal, M. B., R. Fernandopulle, H. R. Song, and B. Landon, "Paying for Quality: Providers' Incentives for Quality Improvement," *Health Affairs*, 23(2):127–141, 2004.

Roski, J., R. Jeddeloh, et al., "The Impact of Financial Incentives and a Patient Registry on Preventive Care Quality: Increasing Provider Adherence to Evidence-Based Smoking Cessation Practice Guidelines," *Preventive Medicine*, 36:291–299, 2003.

Safran, D. G., M. Karp, et al., Measuring Patients' Experiences with Individual Primary Care Physicians: Results of a Statewide Demonstration Project, *J. General Internal Medicine*, 21:13–21, 2006.

Schuster, M. A., E. A. McGlynn, and R. H. Brook, "How Good Is the Quality of Health Care in the United States?" *Milbank Quarterly*, 76(4):517–563, 1998.

Shen, Y., Selection Incentives in a Performance-Based Contracting System, *Health Services Research*, 38(2):535–552, 2003.

Sloan, F. A., D. S. Brown, E. S. Carlisle, G. A. Picone, and P. P. Lee, "Monitoring Visual Status: Why Patients Do or Do Not Comply with Practice Guidelines," *Health Services Research*, 39(5):1429–1448, 2004.

Smart, D. R., *Medical Group Practices in the U.S.*, Chicago, IL: American Medical Association, 2004.

Society for Vascular Surgery, "Developing a Quality Improvement Framework for Surgical Care," Letter to Ways and Means Committee, U.S. House of Representatives, July 20, 2005. Available at http://www.vascularweb.org.

Sorbero, M. E., A. W. Dick, J. Zwanziger, N. Wyle, and D. Mukamel, "The Effect of Capitation on Switching Primary Care Physicians," *Health Services Research*, 38(1):191–209, 2003.

Super, N., "Medicare's Chronic Care Improvement Pilot Program: What Is Its Potential?" National Health Policy Forum Issue Brief No. 797, May 10, 2004. Available at http://www.nhpf.org/pdfs_ib/IB797_ ChronicCare.pdf.

Thames, Thomas "Byron," "Improving Quality in Medicare: The Role of Value-Based Purchasing," Testimony before U.S. Senate Committee on Finance, July 27, 2005.

Thompson, J. W., S. D. Pinidiya, K. W. Ryan, E. D. McKinley, S. Alston, J. E. Bost, J. B. French, and P. Simpson, "Health Plan Quality-of-Care Information Is Undermined by Voluntary Reporting," *American J. Preventive Medicine*, 24(1):62–70, 2003.

Wang, S. J., B. Middleton, et al., "A Cost-Benefit Analysis of Electronic Medical Records in Primary Care," *American J. Medicine*, 114:397–403, 2003.

Wenger N.S. D.H. Solomon, C.P. Roth, C.H. MacLean, D. Saliba, et al. "The Quality of Medical Care Provided to Vulnerable Community-Dwelling Older Patients." *Annals of Internal Medicine*, 139(9) 740-747, 2003.

Werner, R. M., and D. A. Asch, "The Unintended Consequences of Publicly Reporting Quality Information," *JAMA*, 293(10):1239–1244, 2005.

Wessert, W. G., and M. C. Musliner, "Case Mix Adjusted Nursing-Home Reimbursement: A Critical Review of the Evidence," *Milbank Quarterly*, 70(3):455–490, 1992.