Contract No.: 500-00-0033/T.O. 05
MPR Reference No.: 6138-121

MATHEMATICA
Policy Research, Inc.

Evaluation of the Medicare Care Management Performance Demonstration: Design Report

Final Report

May 25, 2007

Lorenzo Moreno
Stacy Dale
Suzanne Felt-Lisk
Leslie Foster
Julita Milliner-Waddell
Eric Grau
Rachel Shapiro
Anne Bloomenthal
Amy Zambrowski

Submitted to:

U.S. Department of Health and Human Services
Centers for Medicare & Medicaid Services
Office of Research, Development, and Information
C3-23-04 Central Bldg.
Mail Stop C3-19-07
7500 Security Blvd.
Baltimore, MD 21244-1850

Project Officer:
Lorraine Johnson

Submitted by:

Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005

Project Director:
Lorenzo Moreno

CONTENTS

CONTENTS (continued)

Chapter **Page**

CONTENTS (continued)

TABLES

TABLES *(continued)*

FIGURES

Figure **Page**

EXECUTIVE SUMMARY

PURPOSE OF THIS EVALUATION DESIGN REPORT

This report describes the evaluation design for the Medicare Care Management Performance (MCMP) Demonstration. In it, we discuss our approach to the impact analysis, including (1) identification of a valid, nonexperimental comparison group; (2) statistical methods; (3) data sources; and (4) outcome measures. We also describe the goals and framework to be used in the implementation analysis. Finally, we discuss a framework for synthesizing our quantitative and qualitative findings to assess the scalability and generalizability of the demonstration.

RATIONALE FOR THE DEMONSTRATION

Section 649 of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003 (MMA) requires the Secretary of the U.S. Department of Health and Human Services to establish a *pay-for-performance* (P4P) demonstration program with physicians to meet the needs of eligible beneficiaries through the adoption and use of health information technology (HIT) and evidence-based outcome measures. The goals of the three-year demonstration are to improve quality of care to eligible fee-for-service Medicare beneficiaries and encourage the implementation and use of HIT. The specific objectives are to promote continuity of care, help stabilize medical conditions, prevent or minimize acute exacerbations of chronic conditions, and reduce adverse health outcomes. The Centers for Medicare & Medicaid Services (CMS) is responsible for designing and operating the MCMP demonstration.

Under the demonstration, physician practices that meet or exceed performance standards established by CMS in clinical performance process and outcome measures will receive a bonus payment for managing the care of eligible Medicare beneficiaries. Practices that submit performance data electronically using a certified electronic health record (EHR) system to CMS will also be eligible for an increase in the incentive payment. The bonuses will be in addition to the normal fee-for-service Medicare payment they receive for services delivered. In a predemonstration (baseline) year, the demonstration will be a *pay-for-reporting* (P4R) initiative to help physicians become familiar with the process of reporting quality measures. The demonstration builds on P4P models used in the private sector, most notably *Bridges to Excellence*™.

DEMONSTRATION DESIGN

The MCMP demonstration will target practices serving at least 50 traditional fee-for-service Medicare beneficiaries with selected chronic conditions for whom they are providing primary care. Under this demonstration, physicians practicing primary care[1] in solo or small- to medium-

[1] The following physician specialties will be eligible to participate in the MCMP demonstration *if they provide primary care*: general practice, allergy/immunology, cardiology, family practice, gastroenterology, internal

size group practices (practices with 10 or fewer physicians, although there may be exceptions) will be eligible to earn incentive payments for (1) reporting quality measures for congestive heart failure, coronary artery disease, diabetes, and the provision of preventive health services during a baseline (predemonstration) period; (2) achieving specified standards on clinical performance measures during the three-year demonstration period; and (3) submitting clinical quality measures to CMS electronically using an EHR that meets industry standards specified by the Certification Commission for Healthcare Information Technology (CCHIT).

The MMA authorizes a total of four sites in both urban and rural areas.[2] The demonstration sites are in Arkansas, California, Massachusetts, and Utah. The Quality Improvement Organizations (QIOs) in these four states will recruit the practices on relationships built through CMS's Doctor's Office Quality - Information Technology (DOQ-IT) project. Only practices participating in DOQ-IT will be eligible to participate in the demonstration. It is expected that the demonstration will enroll 250 practices per state in California and Massachusetts and 150 practices per state in Arkansas and Utah, with an estimated 2,800 physicians participating in MCMP. These practices will represent many organizational structures, and, to participate, they must have at least 50 Medicare beneficiaries. Recruitment of demonstration practices started in January 2007. The demonstration will begin operations on July 1, 2007, and will end in June 2010.

Demonstration practices will be defined by one or more tax identification numbers (TINs). Physicians will be linked to each practice using individual Medicare provider identification numbers (PINs). Medicare beneficiaries who are treated by primary care providers, or those medical subspecialties likely to provide primary care, for the targeted conditions and who are covered under traditional fee-for-service Medicare for both Part A and Part B coverage will be linked to the practices.

Demonstration practices will submit performance data to CMS on up to 26 clinical measures covering treatment related to congestive heart failure, coronary artery disease, diabetes, and the provision of specific preventive and screening services for all beneficiaries assigned with a chronic condition. Through several contractors, CMS will collect data on all the clinical measures for the baseline period and all three years of the demonstration.

The demonstration practices will be eligible to receive up to three incentive payments. First, demonstration practices will receive an incentive of $20 per beneficiary per category (up to $1,000 per physician to a maximum of $5,000 per practice) for reporting baseline clinical quality measures. The payment will not be contingent on the practice's score on any of these measures. Second, for each of the three demonstration years, based on the clinical measures data that the

---

(continued)

medicine, pulmonary disease, geriatric medicine, osteopathic medicine, nephrology, infectious disease, endocrinology, multispecialty clinic or group practice, hematology, hematology/oncology, preventive medicine, rheumatology, and medical oncology.

[2] Appendix A contains a copy of the law.

practices report, CMS will calculate a composite score for each chronic condition (as well as the preventive measures) and compare it against performance thresholds.

Physicians will be eligible for payments of up to $70 per beneficiary for meeting standards related to a specific chronic condition. Beneficiaries who have more than one condition will be counted in each of the relevant groups. For preventive services, physicians will be eligible for a payment of up to $25 per beneficiary with any chronic condition. Physicians will be eligible to earn up to $10,000 per year for performance on all clinical measures. The maximum annual payment to any single practice will be $50,000, regardless of the number of physicians in the practice. Third, practices with a CCHIT certified EHR system that can extract and submit performance data to CMS electronically will be eligible to increase the incentive payment by up to 25 percent, or $2,500 per physician (up to $12,500 per practice) per year during the demonstration period for electronic submission. Thus, practices could receive up to $192,500 over the three years of the demonstration (including the baseline period).

Finally, Congress also mandated an independent evaluation of the MCMP demonstration. The evaluation must include an assessment of P4P's impacts on improving quality of care, care coordination, and continuity of care; reducing Medicare expenditures; and improving health outcomes. The legislation specified that a final evaluation report must be submitted to Congress within 12 months of the demonstration's conclusion. CMS, with funding from the Agency for Healthcare Research and Quality (AHRQ), has contracted with Mathematica Policy Research, Inc. (MPR) to conduct this evaluation.

GOALS OF THE EVALUATION

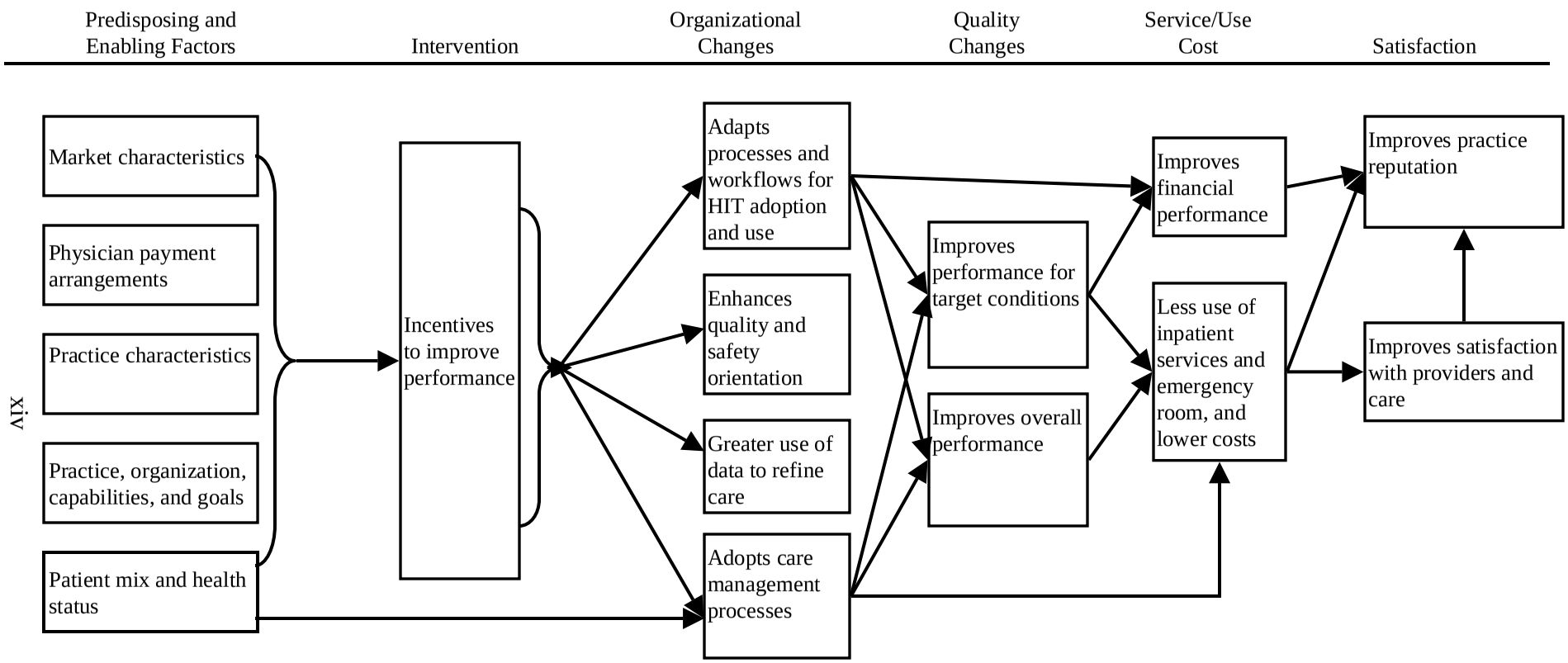
The main goal of the evaluation is to provide CMS with valid estimates of the incremental effect, or *impact*, of providing performance-based financial incentives on the quality of care, use of Medicare-covered services, adoption and use of HIT, and Medicare costs of the chronically ill Medicare beneficiaries served by the demonstration practices. To provide this information, the evaluation must generate rigorous quantitative estimates of the intervention's impacts. In addition, the evaluation will examine the dynamics of practice response to the incentives and supports provided by the demonstration. Figure 1 depicts a logic model for the evaluation, which we will discuss in more detail in this report.

The impact analysis will test the hypotheses that the financial incentives (1) improve quality of care, (2) lower Medicare costs for services by enough to offset the costs of the incentives, (3) influence the adoption and use of HIT, (4) improve continuity of care and care coordination, (5) improve patient satisfaction with care, and (6) improve physician satisfaction. The quality-of-care analysis will assess the care delivery process and the clinical outcomes of Medicare beneficiaries. The cost analysis will include impacts on costs to the Medicare program and Medicare service use.[3] The analysis of HIT use will assess whether practices adopted or increased their HIT use in various office procedures. The continuity-of-care analysis will assess

[3] In addition, as required by OMB, we will monitor budget neutrality during the first 18 months of the demonstration.

FIGURE 1

LOGIC MODEL FOR THE EXPECTED EFFECTS OF THE MCMP DEMONSTRATION



HIT = health information technology.

whether the adoption of P4P reduces care fragmentation. In the satisfaction analysis, patient satisfaction with care and physician satisfaction with the demonstration and its effects on their practices also will be analyzed. Subgroup analyses will test whether the intervention is more effective for certain types of beneficiaries and practices than for others.

The implementation analysis will study the planned interventions as envisioned by a representative set of practices, practices' actual experience with the adoption and use of performance measurement technology (for example, EHRs or disease registries) and care management processes, and the factors that helped or hindered the practices' efforts. The detailed description of the practices' plans will cover the background information on the range of HIT used before the demonstration and how the practices implemented the intervention (the specific changes made to improve patient adherence, refine care processes, lessen fragmentation of care, or avoid adverse drug interactions).

Finally, the synthesis will combine the practice-specific analyses, using impact estimates and implementation analysis findings, to draw inferences about the types of practices that appear to be most successful. It will also examine the generalizability and scalability of the demonstration. As required by CMS, the synthesis will be the basis for the report to Congress, and it will be included in the final evaluation report.

CHALLENGES FOR THE EVALUATION

Several technical challenges must be overcome to achieve the evaluation's objectives. The main challenges are to (1) obtain valid, comparable estimates of impacts for each state; (2) measure some qualitative outcomes; (3) link specific changes in HIT use to specific improvements; and (4) assess the scalability and generalizability of the demonstration.

Estimating Impacts

Three factors may complicate estimation and interpretation of impacts: (1) the feasibility of credible comparison strategies, (2) the period during which the demonstration will be operational, and (3) data completeness for linking physicians to practices.

Although random assignment is generally the strongest study design, several factors made it infeasible for the MCMP demonstration. Therefore, the impact analysis will use a *comparison group* (or *quasi-experimental*) design. To identify the comparison group, the evaluation will use DOQ-IT practices in selected nondemonstration states that match most closely to those in the demonstration states. To be considered a valid comparison practice, the practice must have predemonstration service use and cost patterns similar to those in practices in demonstrations states. It should also have comparable baseline characteristics. We will use statistical tests to ensure that the demonstration and comparison practices do not differ on preenrollment characteristics. However, comparison and demonstration practices may still differ on other observed or unobserved factors (such as interest in, and ability to adopt and use, an EHR system) that are difficult to control for in the impact analysis. Furthermore, the operation of a national Medicare P4R program (Physician Quality Reporting Initiative [PQRI]) for physicians beginning July 1, 2007, or any future Medicare P4P programs for this type of provider, will make it even

more difficult to understand which factors are responsible for the estimated impacts (even after controlling for those practices in the comparison group that decide to participate in the P4R program, since demonstration practices will be exempt from reporting quality measures to PQRI to obtain the bonus). Thus, because of the expected fluidity of the Medicare physician reimbursement environment during the demonstration period, it will be difficult to measure a credible counterfactual (that is, what would have been the experience of the physician practices in the absence of the demonstration) to the demonstration.

The period during which the demonstration will be operational is likely to constrain our ability to identify valid comparison groups. Demonstration practices will be recruited in winter/spring 2007, and many of these practices will have already received technical assistance from the QIOs on implementing EHRs for several years, have actually implemented and adopted this type of system, and may be using it to report quality measures. In contrast, DOQ-IT practices in nondemonstration states that are selected to be part of the comparison group will only be starting to decide whether to adopt an EHR system or receiving technical assistance to reorganize themselves. Controlling for length of the interval since practices first began receiving technical assistance from the QIOs or signed a contract for purchasing an EHR system may not be enough to ensure comparability of demonstration and comparison practices on this dimension. It also may complicate the interpretation of the impact estimates because of the lag between demonstration and comparison practices in receiving technical assistance and implementing EHRs.

Data accuracy and completeness in identifying comparison practices may also present difficulties. The algorithm to be used for linking physicians to potential comparison practices will be the same as the one used for linking participating physicians to demonstration practices. However, the accuracy and completeness of the identifiers available in Medicare Part B claims data for this process will differ and, therefore, may make it difficult to identify comparable DOQ-IT practices in nondemonstration states. Likewise, while demonstration practices will be required to correctly enter the practice and physician identifiers in the claims they submit in order to receive the incentive payments, comparison practices will not be required to do so. Thus, variations in the accuracy and completeness of the identifiers in claims data may result in discrepancies between the two groups in how beneficiaries are assigned to practices during the demonstration (even if the definition of the practice will remain unchanged from baseline in both demonstration and comparison groups) and confound the interpretation of the impact estimates.

Measuring Some Qualitative Outcomes

Several organizational changes that practices may undertake, such as enhancing the practice's quality and safety orientation, probably will affect important qualitative outcomes. For example, as previously discussed, it is expected that the financial incentives will improve continuity of care, which in turn could reduce *fragmentation of care* through better care coordination. Thus, it will be difficult to measure these qualitative outcomes because of the inherent complications of identifying all the providers involved in the care of beneficiaries assigned to specific practices and collecting the data needed to characterize whether the interactions among providers were appropriate, clinically meaningful, and timely.

Linking Changes in Specific HIT Functionalities to Specific Improvements

Demonstration practices will have considerable latitude in deciding how to use HIT, including EHRs, for measuring performance and submitting data electronically to CMS. Furthermore, the demonstration will not assess the incremental effect of specific EHR functionalities. Thus, it will be difficult to link changes in use of specific HIT functionalities to specific outcome improvements.

Assessing the Scalability and Generalizability of the Demonstration

The number of practice characteristics potentially related to improvement in clinical outcomes is large, so linking these characteristics to the success or failure of the practice may be difficult. Moreover, pooling of observations across states may be inappropriate, given such factors as the differences in state regulations; physician licensing arrangements; and P4R, P4P, and EHR penetration. Thus, it will be difficult to generalize the evaluation findings to the Medicare program or to other P4P programs, and to assess the scalability of the intervention, given the demonstration's focus on small- to medium-size practices in only four states.

OVERVIEW OF EVALUATION COMPONENTS

Implementation Analysis

The implementation analysis will examine the dynamics of practice response to the incentives and supports provided by the demonstration. It will also identify barriers and facilitators to (1) successful adoption of HIT to better manage care and improve patient outcomes, (2) greater use of data by the practice to refine the care process, and (3) an enhanced practice orientation to quality and safety. In addition, we will examine how the practice has adapted its patient flows and documentation processes as HIT is implemented, which has implications for efficiency and quality outcomes. Across all these topics, we will include a specific focus on the role of financial incentives and technical assistance.

Data Sources. The analysis of implementation of the demonstration will rely on several data sources: (1) the Office Systems Survey, (2) site visits, (3) telephone discussions with highly successful practices, and (4) telephone discussions with unsuccessful practices, including those that withdrew from the demonstration, if any (Table 1). A literature review will ensure that the site visit discussion guides are consistent with recent research and that all site visitors are knowledgeable about the latest research on P4P as we enter the field.

Office Systems Survey. The Maine Health Information Center will administer the DOQ-IT Office Systems Survey to all demonstration and comparison practices twice during the demonstration for both demonstration and comparison practices. These data will allow us to identify whether demonstration practices have changed how they use electronic tools to improve quality differently from the comparison practices. We will also draw practice characteristics so that we can examine results by practice size and location (by state and whether the practice is located in an urban or rural area).

TABLE 1

DATA SOURCES FOR IMPLEMENTATION, IMPACT, AND SYNTHESIS ANALYSES

Data Source	Time Frame	Expected Sample Size for Demonstration Group[a]			
		Arkansas	California	Massachusetts	Utah
Practice-Specific Data for All Analyses					
Office Systems Survey	Collected during 2007 and 2010	150	250	250	150
Implementation Analysis					
Site Visits	Conducted February through May 2008 and October 2009 through January 2010	8 practices in each state per wave			
Telephone Discussions with Highly Successful Practices	Conducted in years 2 and 3 of demonstration operations	Up to 12 practices (across all states) randomly selected in each wave			
Telephone Discussions with Unsuccessful Practices	Conducted in years 2 and 3 of demonstration operations	Up to 6 practices (across all states) in each wave			
Impact Analysis					
Physician Survey	Conducted in July 2009 (25 months after demonstration begins)	200	200	200	200
Beneficiary Survey	Conducted in January 2009 (19 months after demonstration begins)	600	600	600	600
Medicare Claims Data[b]	Acquired in September 2007 (for baseline measures) and in January of each year from 2009 – 2011 (for follow-up measures)	46,000	50,000	56,000	34,000
Synthesis Analysis					
Financial Payment Data	Acquired in December of each demonstration year	150	250	250	150

[a] For each data source, the comparison group sample will be approximately equal in size; however, financial payment data are only relevant for the demonstration group.

[b] Expected number of Medicare beneficiaries per participating practice estimated based on the actual number of Medicare beneficiaries per practice in each state.

Site Visits. The site visits will provide us with an understanding of how the demonstration is being implemented by practices and, thus, how the practice changes under way may be influencing quality of care, practice efficiency, and patient satisfaction. We plan to conduct one wave of site visits in year 1 (February through May 2008) and a second wave in year 3 (October 2009 through January 2010). Our recommendation to visit the same practices in both time periods would allow us to (1) relate what the practices expected to do and to learn early in the demonstration to what actually they did, and (2) examine how their goals changed. This recommendation implies a reduction in the total number of practices visited on-site from the 76 we originally proposed to 40.

We plan to visit eight demonstration practices and two comparison practices in each state. It is important to visit comparison group practices, because the congressionally mandated link between provider reporting of quality data and financial incentives makes it likely that comparison and demonstration sites both will be improving on the dimensions of interest during the period. Demonstration sites to be visited in each state will be selected judgmentally from among geographically feasible choices, while ensuring that we achieve a mix of practice sizes, urban versus rural location and HIT sophistication. A semistructured interview protocol (which will generally be similar for the demonstration and comparison practices) will be the central data collection instrument.

Telephone Discussions with Highly Successful Practices. After the first and second payouts, we plan to conduct telephone discussions with 12 practices that benefited substantially from the program (for example, practices that received the maximum payment).

Telephone Discussions with Unsuccessful Practices. On a continuing basis beginning in year 2 of the demonstration, we will conduct telephone discussions with unsuccessful practices, including practices that withdrew from the demonstration, if any (up to six practices in demonstration year 2 and six in year 3) to investigate what factors led the practice to suboptimal performance or to withdraw from the demonstration, and what changes might have retained the practice. If no practices withdraw, we will discuss with CMS alternative allocations of these interviews to practices that remained in the demonstration but were unsuccessful (for example, they did not receive performance payments).

Analysis. The implementation analysis will be conducted demonstration-wide—overall and by practice characteristics, including state location. We will identify themes and illustrative examples from among the many site visits and telephone interviews.

Our analysis of the site visit data should answer questions regarding physicians' perspectives on the demonstration; practices' direct responses to the incentives; the influence of other incentive and reporting programs on response; practices' adoption of care management processes; the extent to which data are being used to refine the care process; the extent to which practices enhanced quality and safety; and the adaptation of patient flow and documentation processes as HIT is implemented. We will look for patterns in demonstration experience by state as we analyze each major topic. We also will explore the relationship between practice classifications and outcomes. Practice classifications will be developed from site visit data and other data sources, and will be defined by characteristics such as size, fraction of Medicare patients with targeted conditions, urban versus rural practices, state location, level of HIT use

before the demonstration, practices in higher- versus lower-income areas, and practices that were aggressively developing care management processes for one or more conditions targeted by the demonstration versus those that were not.

Impact Analysis

Estimating impacts of the MCMP demonstration will require a rigorous research design, data from several sources on the outcomes the P4P intervention is expected to influence, and strong statistical models to provide unbiased and efficient estimates of program impacts. Several factors make this task challenging, including (1) the need to rely on a quasi-experimental design; (2) the need for separate impact estimates for each state; and (3) the considerable variation of many factors across states, including the timing and intensity of technical assistance for implementing EHRs, P4P and EHR penetration, physician licensure regulations, and accuracy and completeness of key identifiers in claims data.

Research Design. The two key features for ensuring that valid estimates of impacts are obtained are (1) the comparison group strategy (how we select the group to estimate what would have occurred to demonstration practices and beneficiaries without the P4P incentives), and (2) the sample size. We will estimate impacts of the demonstration through a *difference-in-differences* approach. With this approach, we will compare changes in quality measures and other outcomes of *practices* in the demonstration states and comparison states before and after the start of the demonstration. We will carefully select the comparison practices to minimize the possibility of bias in our impact estimates. Having adequate sample sizes will ensure that impacts of policy-relevant size do not go undetected.

Quasi-Experimental Design for Demonstration Impacts. The impact analysis will use a *comparison group* (or *quasi-experimental*) design. To identify the comparison group, the evaluation will select DOQ-IT physician practices in selected nondemonstration states that match most closely those in demonstration states. The collection of matched comparison practices in the nondemonstration states will form the comparison group for this evaluation.

Selection of Comparison States. Using a reproducible process, we selected nondemonstration states using criteria that aimed to identify states with environments similar to those of the demonstration states in that they at least had EHR and P4P programs.[4] Based on this selection process, we proposed the following states be used as comparison states for the MCMP demonstration states: *Arkansas:* Nebraska, with Texas as alternate; *California:* for comparison to southern California only, Arizona; for comparison to California overall, Oregon, with Washington as alternate; *Massachusetts:* New York, with Connecticut as alternate; and *Utah*: Idaho. Although the comparison states chosen have face validity and meet the criteria used for selection, they are only an approximate match to the demonstration states.

Selection of Comparison Group Practices. To be considered a valid comparison practice, the practice's patients must have predemonstration service use and cost patterns similar to those

[4] Appendix C describes in detail the process for selecting comparison states.

of practices in demonstration states. The practice also should have comparable baseline characteristics. We will use statistical matching methods to ensure that the demonstration and comparison practices do not differ on predemonstration service use measures, costs, and baseline characteristics. To do this, we will first stratify the sample by constructing cells defined by practice size, and experience with HIT (for example, whether practices have a disease registry or an EHR system). Within each cell, we will use statistical matching methods to identify the comparison practice that best matches each demonstration practice in terms of predemonstration service use measures, costs, and baseline characteristics. Ideally, we will have several suitable comparison practices for each demonstration practice. If so, we will select the comparison practice that provides the closest predemonstration match to a demonstration practice.

The measures we plan to use to match practices (in addition to practice size, and experience with HIT) include the number of Medicare fee-for-service beneficiaries served by the practice, number of evaluation and management (E&M) visits, number of hospital admissions, and Medicare expenditures per beneficiary. However, the final list of baseline characteristics will depend on the availability of specific data elements in the Office Systems Survey database. For the measures of service use and expenditures, we will use claims data to be supplied by the financial support contractor.

Identification of Comparison Group Beneficiaries. To link beneficiaries to comparison group practices, the demonstration's financial support contractor will use PINs available in claims data and the practices' demonstration application form or Office Systems Survey (for both demonstration and comparison practices), and an algorithm for allocating beneficiaries to only one practice. For demonstration practices, this procedure avoids double-counting of beneficiaries on whom the incentive payments will be based. For both demonstration and comparison practices, the algorithm assigns each beneficiary represented in the claims files to the physician who provided the *plurality* of E&M services during a given period. As a tiebreaker for beneficiaries with more than one such physician, the algorithm assigns the physician with the most recent E&M visit, the practice with the highest Medicare expenditures for that patient, and whether it is a demonstration practice (only for practices in demonstration states).

Estimation of Demonstration Impacts. We will use the difference-in-differences methods to estimate impacts on claims-based outcomes, such as use of Medicare-covered services and costs. This regression-based method implicitly accounts for all factors—both measured and unmeasured—that do not change over time when estimating impacts, and thus is likely to yield unbiased estimates. In addition, we will explore the use of statistical methods to control for *selection bias* in our estimates from survey data (such as measures of satisfaction with care). This approach may account for systematic differences between practices in the comparison and treatment groups on preenrollment characteristics that are difficult to measure, such as the motivation to provide higher or lower quality of care.

The practice will be the *unit of analysis*, because it is the *unit of intervention*. That is, the practice—not the beneficiary—will receive the financial incentives. Furthermore, the physician will *not* be the unit of analysis. As a result, our analytic sample for estimating impacts from the demonstration will consist of all beneficiaries in the demonstration and comparison group practices. In addition, we will measure relevant subgroups of beneficiaries defined, for example, by the demonstration chronic conditions.

Because the analysis is multilevel (beneficiaries will be nested within practices), we plan to use *hierarchical* (or multilevel) linear models to estimate impacts, which will include a range of individual and practice characteristics and their interactions. For example, this type of modeling will allow us to examine the interactions between beneficiary and practice characteristics, while accounting for clustering of beneficiaries within practices.

Data Sources. The impact analysis will use data from four data sources: (1) a beneficiary survey, (2) a physician survey, (3) Medicare claims and eligibility data, and (4) practice-specific data. We will administer a mail survey (with telephone followup) to 4,800 eligible beneficiaries (600 from the demonstration group and 600 from the comparison group in each state) 19 months after the beginning of the demonstration's operations (in or around January 2009). This survey will measure demonstration and comparison group members' well-being, access to care, adherence to self-care management principles, continuity of care, satisfaction with care, and awareness of the demonstration (or the DOQ-IT program, in nondemonstration states). We will also administer a mail survey (with telephone followup) to 1,600 physicians (200 in the demonstration group and 200 in the comparison group in each state) 25 months after the start of the demonstration (in or around July 2009). This survey will measure demonstration and comparison group barriers to transforming the practices' clinical encounters with beneficiaries and other office procedures, barriers and facilitators to adoption of HIT, experience implementing this type of system, experience with P4R and P4P (in the demonstration sites only), and satisfaction with EHRs. In addition, there will be a separate module for demonstration physicians, focusing on how participation in the demonstration influenced the practice, their perceptions of the effects of the financial incentives on their practices, and their satisfaction with the demonstration. We will obtain data on use of Medicare-covered services and expenditures, as well as scores for selected clinical measures, from Medicare claims data, and demographic and eligibility data from the Medicare Enrollment Database (EDB). We will use these Medicare data both to construct outcome measures and to construct regression control variables based on the baseline period. Finally, we will draw practice-specific measures for the impact analysis from the Office Systems Survey and from financial incentive payment data.

Sample Sizes. The demonstration's budget and the number of practices likely to enroll in the DOQ-IT program determined the minimum number of physicians and beneficiaries required for detecting demonstration impacts.

Minimum Detectable Differences for the Full Sample. Assuming that there are 4,800 responding beneficiaries (600 in demonstration practices in each state and 600 in comparison group practices in each state), we will be able to detect a difference in a binary outcome (with mean equal to 50 percent) of 8 percentage points for within-state analyses and 4 percentage points for analyses that are pooled across states (Table 2) (assuming 80 percent power and 5 percent level for a two-sided test). Moreover, when using claims data, we will be able to detect even smaller differences in outcomes between the treatment and comparison groups (less than one percentage point in within-state analyses and less than one-third of a percentage point for the analyses that pool all states together). With the full sample of 1,600 physicians (200 in the demonstration practices in each state and 200 in the comparison group practices in each state), detectable differences are large (about 16 to 20 percentage points for within-state analyses and about 9 percentage points for analyses that pool all states) but still adequate for identifying major

impacts. Finally, for continuous expenditure variables derived from Medicare claims data, we will detect differences of about two percent in within-state analyses and about one percent for the analyses that pool all states together (assuming a coefficient of variation of 2.5 and 80 percent power).

Precision for Descriptive Estimates of Clinical Outcomes Among Demonstration Practices. To examine changes in practice performance over time, and the correlation of these trends with practice characteristics (including incentive payments for a preceding year), we will consider the practice as the unit of analysis. In such analyses, for binary outcomes, the half-width for the 95 percent confidence interval is less than one-half of a percentage point for within-state analyses, and is about one-fifth of a percentage point for the sample that is pooled across states (Table 2). For continuous variables (assuming a coefficient of variation of 1.75), the half-width for the 95 percent confidence interval is less than 1.5 percent for within-state analyses, and is less than 1 percent for the sample that is pooled across states. Subgroup analyses reduce the precision, especially for continuous variables. For example, in within-state analyses, the half-width of a 95 percent confidence interval for a binary variable for a 50 percent subgroup is less than 1 percentage point, while the corresponding half-width for a continuous variable ranges from 1.9 to 2.4 percent of the comparison group mean.

Synthesis Analysis

The ultimate goal of the evaluation will be to provide guidance to CMS on whether P4P incentives for improving quality of care and for adopting HIT in solo or small- to medium-size group physician practices serving Medicare beneficiaries with chronic illnesses should be implemented on a larger scale and, if so, how this intervention might best be structured. Whether P4P should be implemented depends on whether the demonstration leads to improved quality of care and is at least budget neutral. Structuring of the intervention requires assessing the answers to three questions: (1) For which types of practices were the incentives most effective? (2) How did clinical outcomes vary with the incentives? and (3) How did quality of care, Medicare costs, and the financial incentives vary with HIT use?

To address these goals, we will synthesize our findings for the report to Congress (and for the final evaluation report). In the synthesis, we will pull together our findings from practices in all four states and outcome measures from both the implementation and impact analyses. We will use this information to draw inferences about the role that financial incentives play in improving care for Medicare beneficiaries with chronic illnesses and on the adoption and use of HIT, and about the most successful ways to implement the incentives (and the technology for performance reporting). The synthesis will entail determining how the intervention's effectiveness varies with practice characteristics.

TABLE 2

MINIMUM DETECTABLE DIFFERENCES AND HALF-WIDTHS FOR
95 PERCENT CONFIDENCE INTERVALS

	Arkansas	California	Massachusetts	Utah	All States Pooled
	Number of Demonstration Practices				
	150	250	250	150	800
Minimum Detectable Differences, 80 Percent Power, Two-Tailed Tests					
Binary Variable with Mean of .5					
Physician Survey	20.4	15.6	15.6	20.1	8.8
Beneficiary Survey	8.1	8.1	8.1	8.1	4.2
Claims Data	0.7	0.6	0.6	0.7	0.3
Continuous Variable with Coefficient of Variation of 2.5					
Claims Data	2.3	1.8	1.8	2.3	1.0
Half-Widths for 95 Percent Confidence Level in Practice-Level Descriptive Analysis					
Binary Variable with Mean of .5					
Full Sample	0.4	0.3	0.3	0.4	0.2
Subgroup Size:					
80 percent	0.5	0.4	0.4	0.5	0.2
50 percent	0.7	0.5	0.5	0.7	0.3
15 percent	1.4	1.1	1.1	1.4	0.6
Continuous Variable with Coefficient of Variation of 1.75					
Full Sample	1.4	1.1	1.1	1.4	0.6
Subgroup Size:					
80 percent	1.7	1.3	1.3	1.7	0.7
50 percent	2.4	1.9	1.9	2.4	1.0
15 percent	4.9	3.8	3.8	4.9	2.1

To accomplish the evaluation's basic goals, we will draw on the state-specific implementation and impact analyses to describe physician practices' experiences adopting and using HIT for performance reporting and the care management strategies they use for chronically ill fee-for-service beneficiaries to improve quality of care. Likewise, we will describe how impacts varied with many of the practice characteristics that could potentially influence the efficiency of P4P programs. Our approach to the synthesis will involve three components, all of which feed into the recommendations. In the first component, we will use *exploratory* and *confirmatory* analyses to assess which practice characteristics seem to successfully improve clinical outcomes and reduce costs. In the second component, we will assess how clinical outcomes vary with the incentives the practices will receive for attaining predetermined performance standards. Finally, in the third component, we will examine the association between changes in clinical outcomes, costs, and the incentives, and the practice's level of HIT use.

Reporting of Demonstration Findings

The demonstration evaluation will produce several reports, including an implementation report, a report on site visits, and a cost neutrality monitoring report, as well as interim and final reports that synthesize findings across states and analytic components. The interim reports will be adapted to develop a report to Congress. Table 3 summarizes the schedule for the deliverables.

TABLE 3

SCHEDULE OF DRAFT REPORT DUE DATES

	Draft Due	
Report	Project Month[a]	Calendar Month
Design report	n.a.	February 2007
Implementation report	13	July 2008
First interim synthesis	16	October 2008
Cost neutrality monitoring report	24	June 2009
Second interim synthesis	28	October 2009
Report to Congress (Third interim synthesis)	40	October 2010
Site visits report	46	April 2011
Final synthesis	51	September 2011

[a] Refers to months after the start of the demonstration (July 1, 2007).

n.a.= not applicable.

I. INTRODUCTION

This report describes the evaluation design for the Medicare Care Management Performance (MCMP) Demonstration. In it, we discuss our approach to the impact analysis, including (1) identification of a valid, nonexperimental comparison group; (2) statistical methods; (3) data sources; and (4) outcome measures. We also describe the goals and framework to be used in the implementation analysis. Finally, we discuss a framework for synthesizing our quantitative and qualitative findings to assess the scalability and generalizability of the demonstration.

A. RATIONALE FOR THE DEMONSTRATION

Section 649 of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003 (MMA) requires the Secretary of the U.S. Department of Health and Human Services to establish a *pay-for-performance* (P4P) demonstration program with physicians to meet the needs of eligible beneficiaries through the adoption and use of health information technology (HIT) and evidence-based outcome measures. The goals of the three-year demonstration are to improve quality of care to eligible fee-for-service Medicare beneficiaries and encourage the implementation and use of HIT. The specific objectives are to promote continuity of care, help stabilize medical conditions, prevent or minimize acute exacerbations of chronic conditions, and reduce adverse health outcomes. The Centers for Medicare & Medicaid Services (CMS) is responsible for designing and operating the MCMP demonstration.

Under the demonstration, physician practices that meet or exceed performance standards established by CMS in clinical performance process and outcome measures will receive a bonus payment for managing the care of eligible Medicare beneficiaries. Practices that submit performance data electronically using a certified electronic health record (EHR) system to CMS will also be eligible for an increase in the incentive payment. The bonuses will be in addition to

the normal fee-for-service Medicare payment they receive for services delivered. In a predemonstration (baseline) year, the demonstration will be a *pay-for-reporting* (P4R) initiative to help physicians become familiar with the process of reporting quality measures. The demonstration builds on P4P models used in the private sector, most notably *Bridges to Excellence*™ (Bodenheimer et al. 2005; de Brantes 2005; Iglehart 2005).

B. DEMONSTRATION DESIGN

The MCMP demonstration will target practices serving at least 50 traditional fee-for-service Medicare beneficiaries with selected chronic conditions for whom they provide primary care. Under this demonstration, physicians practicing primary care[1] in solo or small- to medium-size group practices (practices with 10 or fewer physicians, although there may be exceptions) will be eligible to earn incentive payments for (1) reporting quality measures for congestive heart failure, coronary artery disease, diabetes, and the provision of preventive health services during a baseline (predemonstration) period; (2) achieving specified standards on clinical performance measures during the three-year demonstration period; and (3) submitting clinical quality measures to CMS electronically using an electronic health record (EHR) that meets industry standards specified by the Certification Commission for Healthcare Information Technology (CCHIT).

[1] The following physician specialties will be eligible to participate in the MCMP demonstration *if they provide primary care*: general practice, allergy/immunology, cardiology, family practice, gastroenterology, internal medicine, pulmonary disease, geriatric medicine, osteopathic medicine, nephrology, infectious disease, endocrinology, multispecialty clinic or group practice, hematology, hematology/oncology, preventive medicine, rheumatology, and medical oncology.

The MMA authorizes a total of four sites in both urban and rural areas.[2] The demonstration sites are in Arkansas, California, Massachusetts, and Utah. The Quality Improvement Organizations (QIOs) in these four states will recruit the practices on relationships built through CMS's Doctor's Office Quality - Information Technology (DOQ-IT) project. Only practices participating in DOQ-IT will be eligible to participate in the demonstration. It is expected that the demonstration will enroll 250 practices per state in California and Massachusetts and 150 practices per state in Arkansas and Utah, with an estimated 2,800 physicians participating in MCMP. These practices will represent many organizational structures, and, to participate, they must have at least 50 Medicare beneficiaries. Recruitment of demonstration practices started in January 2007. The demonstration will begin operations on July 1, 2007, and will end in June 2010.

Demonstration practices will be defined by one or more tax identification numbers (TINs). Physicians will be linked to each practice using individual Medicare provider identification numbers (PINs). Medicare beneficiaries who live in a demonstration state and who are treated by primary care providers, or those medical subspecialties likely to provide primary care, for the targeted conditions and who are covered under traditional fee-for-service Medicare for both Part A and Part B coverage will be linked to these practices.[3]

Demonstration practices will submit performance data to CMS on up to 26 clinical measures covering treatment related to congestive heart failure, coronary artery disease, diabetes, and the provision of specific preventive and screening services for all beneficiaries assigned with a

[2] In addition, the statute requires that one site be "in a state with a medical school with a Department of Geriatrics that manages rural outreach sites and is capable of managing patients with multiple chronic conditions, one of which is dementia." Appendix A contains a copy of the law.

[3] Beneficiaries for whom Medicare is not the primary source of insurance coverage or whose care a hospice program manages will be excluded from the demonstration.

chronic condition.[4] Through several contractors, CMS will collect data on all the clinical measures for the baseline period and all three years of the demonstration.

The demonstration practices will be eligible to receive up to three incentive payments. First, demonstration practices will receive an incentive of $20 per beneficiary per category (up to $1,000 per physician to a maximum of $5,000 per practice) for reporting baseline clinical quality measures. The payment will not be contingent on the practice's score on any of these measures. Second, for each of the three demonstration years, based on the clinical measures data that the practices report, CMS will calculate a composite score for each chronic condition (as well as the preventive measures) and compare it against performance thresholds. Physicians will be eligible for payments of up to $70 per beneficiary for meeting standards related to a specific chronic condition. Beneficiaries who have more than one condition will be counted in each of the relevant groups. For preventive services, physicians will be eligible for a payment of up to $25 per beneficiary with any chronic condition. Physicians will be eligible to earn up to $10,000 per year for performance on all clinical measures. The maximum annual payment to any single practice will be $50,000, regardless of the number of physicians in the practice. Third, practices with a CCHIT certified EHR system that can extract and submit performance data to CMS electronically will be eligible to increase the incentive payment by up to 25 percent, or $2,500 per physician (up to $12,500 per practice) per year during the demonstration period for electronic submission. Thus, practices could receive up to $192,500 over the three years of the demonstration (including the baseline period).

[4] In addition to three primary target chronic conditions—congestive heart failure, coronary artery disease, and diabetes mellitus—the other eligible conditions are Alzheimer's disease or other mental, psychiatric, or neurological disorders; any heart condition (such as arteriosclerosis, myocardial infarction, or angina pectoris/stroke); any cancer; arthritis and osteoporosis; kidney disease; and lung disease. These conditions will be identified through ICD-9-CM diagnosis codes available in Medicare claims data (Wilkin et al. 2007).

Finally, Congress also mandated an independent evaluation of the MCMP demonstration. The evaluation must include an assessment of P4P's impacts on improving quality of care, care coordination, and continuity of care; reducing Medicare expenditures; and improving health outcomes. The legislation specified that a final evaluation report must be submitted to Congress within 12 months of the demonstration's conclusion. CMS, with funding from the Agency for Healthcare Research and Quality (AHRQ), has contracted with Mathematica Policy Research, Inc. (MPR) to conduct this evaluation.

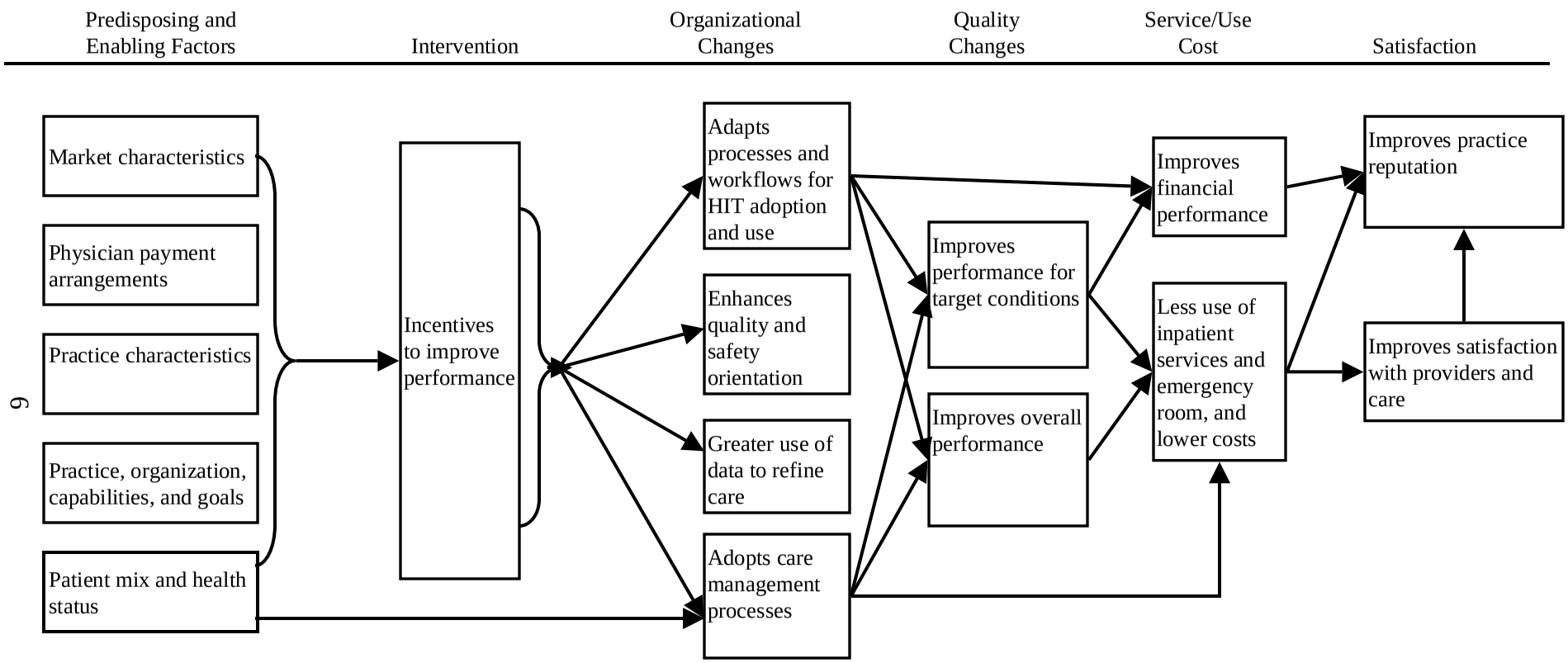
C. GOALS OF THE EVALUATION

The main goal of the evaluation is to provide CMS with valid estimates of the incremental effect, or *impact*, of providing performance-based financial incentives on the quality of care, use of Medicare-covered services, adoption and use of HIT, and Medicare costs of the chronically ill Medicare beneficiaries served by the demonstration practices. To provide this information, the evaluation must generate rigorous quantitative estimates of the intervention's impacts. In addition, the evaluation will examine the dynamics of practice response to the incentives and supports provided by the demonstration. Figure I.1 depicts a logic model for the expected effects of MCMP. It details the pathways through which the intervention (incentives to improve performance) will influence practice organizational changes that may result in improved quality of care, lower costs, and other outcomes. We will use this logic model as the framework for our evaluation.

The evaluation will include an impact analysis, implementation analysis, and synthesis analysis. We provide an overview of each analysis below, and Table I.1 summarizes the primary research questions, data sources, and planned analysis methods. These analyses will be described in detail in subsequent chapters.

FIGURE I.1

LOGIC MODEL FOR THE EXPECTED EFFECTS OF THE MCMP DEMONSTRATION



HIT = health information technology.

TABLE I.1

RESEARCH QUESTIONS, DATA SOURCES, AND ANALYSIS METHODS
FOR THE MCMP EVALUATION, BY ANALYTIC COMPONENT

Research Question	Data Source	Analysis Method
Impact Analysis		
What were the demonstration's effects on:		
Quality of care?	Medicare claims data and Beneficiary survey	Regression-adjusted comparison of demonstration and comparison group means
Medicare service use and costs?	Medicare claims data	
Implementation and use of HIT?	Physician Survey and Office Systems Survey	
Continuity of care and care coordination?	Medicare claims data, Beneficiary Survey, and Physician Survey	
Patient satisfaction?	Beneficiary Survey	
Physician satisfaction?	Physician Survey	
Implementation Analysis		
What types of practices participated?	Office Systems Survey and practice-level scores and financial payment levels for demonstration practices	Comparison of characteristics of practices submitting data to those enrolled but not submitting data
What changes have practices made in terms of HIT use in response to the demonstration?	Office Systems Survey	Comparison of the HIT use of demonstration practices over the demonstration period
What were physicians' views of the demonstration and how their practice responded to it?	Site visit data and telephone discussions with successful and unsuccessful practices (or practices that withdrew) and Physician Survey	Qualitative analysis
Synthesis Analysis		
For which types of practices were the incentives most effective?	Financial payment data, Medicare claims data, Office Systems Survey, Beneficiary Survey	Comparison of mean characteristics of successful and unsuccessful practices; regression analysis of the relationship between practice characteristics and outcomes
How did clinical outcomes vary with the incentives?	Financial payment data, Medicare claims data, Office Systems Survey	Regression analysis of the relationship between clinical outcomes and incentive payments in previous year
How did quality of care, Medicare costs, and the financial incentives vary with HIT use?	Financial payment data, Medicare claims data, Office Systems Survey	Regression analysis of the relationship between Medicare costs and HIT use; regression analysis of the relationship between HIT use and incentive payments in previous year

The impact analysis will compare regression-adjusted outcome measures for the treatment and comparison groups in order to test the hypotheses that the financial incentives (1) improve quality of care, (2) lower Medicare costs for services by enough to offset the cost of the incentives, (3) influence the adoption and use of HIT, (4) improve continuity of care and care coordination, (5) improve patient satisfaction with care, and (6) improve physician satisfaction with the demonstration. The quality-of-care analysis will assess the care delivery process and the clinical outcomes of Medicare beneficiaries. The cost analysis will include impacts on costs to the Medicare program and Medicare service use.[5] The analysis of HIT use will assess whether practices adopted or increased their HIT use in various office procedures. The continuity-of-care analysis will assess whether the adoption of P4P reduces care fragmentation. In the satisfaction analysis, patient satisfaction with care and physician satisfaction with the demonstration and its effects on their practices also will be analyzed. Subgroup analyses will test whether the intervention is more effective for certain types of beneficiaries and practices than for others. Outcome measures will be drawn from Medicare claims data, the beneficiary survey, and the physician survey.

The implementation analysis will use qualitative analysis and descriptive statistics to study the planned interventions as envisioned by a representative set of practices, practices' actual experience with the adoption and use of performance measurement technology (for example, EHRs or disease registries) and care management services, and the factors that helped or hindered the practices' efforts. The detailed description of the practices' plans will cover the background information on the range of HIT used before the demonstration and how the practices implemented the intervention (that is, the specific changes made to improve patient

[5] In addition, as required by OMB, we will monitor budget neutrality during the first 18 months of the demonstration.

adherence, refine care processes, lessen fragmentation of care, or avoid adverse drug interactions). Data sources will include site visits, telephone interviews with practices, financial payment data, practice-level scores, and the Office Systems Survey.

The synthesis will combine the practice-specific analyses, using impact estimates and implementation analysis findings, to draw inferences about the types of practices that appear to be most successful. It will use regression analyses to investigate (1) the relationship between clinical outcomes and the previous year's incentives, (2) Medicare costs and HIT use, and (3) HIT use and the previous year's incentives. It will also examine the generalizability and scalability of the demonstration. As required by CMS, the synthesis will be the basis for the report to Congress, and it will be included in the final evaluation report.

D. CHALLENGES FOR THE EVALUATION

Several technical challenges must be overcome to achieve the evaluation's objectives. The main challenges are to (1) obtain valid, comparable estimates of impacts for each state; (2) measure some qualitative outcomes; (3) link specific changes in HIT use to specific improvements; and (4) assess the scalability and generalizability of the demonstration.

1. Estimating Impacts

Three factors may complicate estimation and interpretation of impacts: (1) the feasibility of credible comparison strategies, (2) the period during which the demonstration will be operational, and (3) data completeness for linking physicians to practices.

Although random assignment is generally the strongest study design, several factors made it infeasible for the MCMP demonstration. Therefore, the impact analysis will use a *comparison group* (or *quasi-experimental*) design. To identify the comparison group, the evaluation will use DOQ-IT practices in selected nondemonstration states that match most closely to those in the

demonstration states. To be considered a valid comparison practice, the practice must have predemonstration service use and cost patterns similar to those in practices in demonstration states. It should also have comparable baseline characteristics. We will use statistical tests to ensure that the demonstration and comparison practices do not differ on preenrollment characteristics. However, comparison and demonstration practices may still differ on other observed or unobserved factors (such as interest and ability to adopt and use an EHR system) that are difficult to control for in the impact analysis. Furthermore, the operation of a national Medicare P4R program (Physician Quality Reporting Initiative [PQRI]) for physicians beginning July 1, 2007, or any future Medicare P4P programs for this type of provider, will make it even more difficult to understand which factors are responsible for the estimated impacts (even after controlling for those practices in the comparison group that decide to participate in the P4R program, since demonstration practices will be exempt from reporting quality measures to PQRI to obtain the bonus). Thus, because of the expected fluidity of the Medicare physician reimbursement environment during the demonstration period, it will be difficult to measure a credible counterfactual (that is, what would have been the experience of the physician practices in the absence of the demonstration) to the demonstration.

The period during which the demonstration will be operational is likely to constrain our ability to identify valid comparison groups. Demonstration practices will be recruited in winter/spring 2007, and many of these practices will have already received technical assistance from the QIOs on implementing EHRs for several years, have actually implemented and adopted this type of system, and may be using it to report quality measures. In contrast, DOQ-IT practices in nondemonstration states that are selected to be part of the comparison group will only be starting to decide whether to adopt an EHR system or receiving technical assistance to reorganize themselves. Controlling for length of the interval since practices first began receiving

technical assistance from the QIOs or signed a contract for purchasing an EHR system may not be enough to ensure comparability of demonstration and comparison practices on this dimension. It also may complicate the interpretation of the impact estimates because of the lag between demonstration and comparison practices in receiving technical assistance and implementing EHRs.

Data accuracy and completeness in identifying comparison practices may also present difficulties. The algorithm to be used for linking physicians to potential comparison practices will be the same as the one used for linking participating physicians to demonstration practices. However, the accuracy and completeness of the identifiers available in Medicare Part B claims data for this process will differ and, therefore, may make it difficult to identify comparable DOQ-IT practices in nondemonstration states. Likewise, while demonstration practices will be required to correctly enter the practice and physician identifiers in the claims they submit in order to receive the incentive payments, comparison practices will not be required to do so.[6] Thus, variations in the accuracy and completeness of the identifiers in claims data may result in discrepancies between the two groups in how beneficiaries are assigned to practices during the demonstration (even if the definition of the practice will remain unchanged from baseline in both the demonstration and comparison groups) and confound the interpretation of the impact estimates.

[6] However, the new Medicare P4R program probably will promote more complete and accurate data collection by practices that decide to participate, including those in the comparison group. This may dampen down the potential differences between demonstration and comparison practices in their motivation for accurately reporting identifiers, as well as other data elements.

2. Measuring Some Qualitative Outcomes

Several organizational changes that practices may undertake, such as enhancing the practice's quality and safety orientation, probably will affect important qualitative outcomes. For example, as previously discussed, it is expected that the financial incentives will improve continuity of care, which in turn could reduce *fragmentation of care* through better care coordination. Thus, it will be difficult to measure these qualitative outcomes because of the inherent complications in identifying all the providers involved in the care of beneficiaries assigned to specific practices and collecting the data needed to characterize whether the interactions among providers were appropriate, clinically meaningful, and timely.

3. Linking Changes in Specific HIT Functionalities to Specific Improvements

Demonstration practices will have considerable latitude in deciding how to use HIT, including EHRs, for measuring performance and submitting data electronically to CMS. Furthermore, the demonstration will not assess the incremental effect of specific EHR functionalities. Thus, it will be difficult to link changes in use of specific HIT functionalities to specific outcome improvements.

4. Assessing the Scalability and Generalizability of the Demonstration

The number of practice characteristics potentially related to improvement in clinical outcomes is large, so it may be difficult to link these characteristics to the success or failure of the practice. Moreover, pooling of observations across states may be inappropriate because of such factors as the differences in state regulations; physician licensing arrangements; and P4R, P4P, and EHR penetration. Thus, given the demonstration's focus on small- to medium-size practices in only four states, it will be difficult to generalize the evaluation findings to the Medicare program or to other P4P programs and to assess the scalability of the intervention.

E. GUIDE TO THIS REPORT

In Chapter II, we describe the implementation analysis objectives and approach. Chapter III discusses the impact analysis, including the hypotheses, research design, data sources, and outcome measures, as well as statistical procedures we will use to overcome the methodological challenges of the evaluation. Chapter IV explains how we will synthesize the findings from the process that the physician practices use to adopt the technology and modify their care delivery processes and impact analyses. Chapter V reviews the reports that will be produced. The appendixes contain the enabling legislation and other supporting technical materials.

II. DESIGN OF THE IMPLEMENTATION ANALYSIS

A. GOALS AND KEY QUESTIONS TO BE ADDRESSED

The goal of the demonstration is to improve quality of care for chronically ill Medicare fee-for-service beneficiaries and promote the adoption and use of HIT (see Figure I.1). The implementation analysis will first assess the demonstration's success in gaining widespread participation, and then examine the dynamics of practice response to the incentives. Specifically, we will obtain physicians' perspectives on the effects of the demonstration on their practices, and whether they as individuals have changed any practice behaviors in response. We will also identify barriers and facilitators to (1) greater use of data by the practice to refine the care process, (2) successful adoption of HIT to better manage care to improve patient outcomes, and (3) an enhanced practice orientation to quality and safety. In addition, we will examine how the practice has adapted its patient flows and documentation processes as HIT is implemented, because this has implications for efficiency and quality outcomes. Across all these topics, we will focus on the roles of financial incentives and technical assistance.

Table II.1 lists key questions to be explored related to each of these goals. The questions are grouped into seven broad topic areas:

1. ***Participating Practice Perspectives on the Demonstration.*** Questions here include features of the demonstration that the practices like and dislike, early expectations of gain from the demonstration, and experience with data submission and communication under the demonstration.
2. ***Direct Response to the Incentives.*** What, if anything, has the practice done differently due to the incentives from the demonstration?
3. ***Adaptation of Patient Flow and Documentation Processes as HIT Is Implemented.*** Questions include what operational changes the practice made as it implemented HIT; whether the demonstration influenced its thinking about making changes as it implemented HIT; and the effects of changes resulting from HIT adoption (for

TABLE II.1

KEY QUESTIONS FOR THE IMPLEMENTATION ANALYSIS

I. Participating Practice Perspectives on the Demonstration

- *Year 1 understanding of, and participation in, the demonstration:* To what extent has the practice participated to date? If the practice is not submitting data for all measures and for the baseline year, why not?
- *Submission process:* Has the data submission process gone smoothly?
- *Expectations:* How much does the practice expect to gain from participation in the next year? In the next three years? What does the practice think will be the key factors in whether its expectations are met?
- *Views on the incentives:* Which features of the demonstration does the practice particularly like or dislike? More broadly, how much does it favor linking payment to quality of care through incentives?
- *Communication about the incentives:* Has communication between CMS and the QIO and the practice about the incentives been clear? Does the practice feel the incentive payment it has received has been accompanied by enough information to understand why it received the amount it did, and whether there are specific areas in which it could improve?

II. Direct Response to the Incentives

- What, if anything, has the practice done differently due to the incentives from the demonstration?

III. Adaptation of Patient Flow and Documentation Processes as HIT Is Implemented

- *Changes made:* With implementation of HIT, what changes were made in how the practice operates day to day? Which changes were essential to make the HIT function, and which were changes that made sense more broadly to take best advantage of the HIT?
- *Factors influencing the changes (for demonstration practices):* Did participation in the MCMP influence your thinking about making changes when you implemented the HIT? What other information sources or other factors influenced your thinking about what changes you should make with HIT implementation?
- *Effects of the changes:* Have the changes you made had any effect on the time spent on each patient visit? On the time spent on administrative versus clinical functions? On the completeness of the practice's documentation? On the usefulness of the information you have immediately in hand at the start of patient appointments?

IV. Relevant Context—Other Incentive and Reporting Programs

- Have other incentives that the practice faces from other payers or other reporting programs (such as the PQRI) affected how the practice has responded to the incentives under the demonstration? If so, how?

V. Adoption of Care Management Processes

- *Speed/extent of adoption:* How quickly and extensively are the practices adopting care management processes? Why? What are the "next steps" in implementing [more] care management and what are the major factors affecting the timing of those steps?
- *Perceived costs and benefits:* What does the practice perceive as the benefits and costs of adopting care management for its practice? For its patients? How does it view the relative benefits and costs of adopting care management for different conditions?
- *Smoothness of implementation:* For those with full implementation of care management for one or more conditions, how smoothly did implementation go? Why?

TABLE II.1 *(continued)*

- *Perceived effects:* Has implementation of care management affected the functioning of the office? Is it producing any results yet for the patients?
- *System support:* Does the HIT the practice has adopted provide good support for care management? Are the care management capabilities of its current system being fully used? How do practices' implementation of, and views about, care management relate to their decisions about the type of HIT they implement?
- *MCMP demonstration role (for demonstration practices):* Has participating in the MCMP demonstration affected the practice's views on care management, its decision to adopt care management processes, or the smoothness of implementation of the processes?
- *Role of other external factors:* What, if any, factors outside the practice have influenced the practice's view on care management, its decision to adopt care management processes, or the smoothness of implementation of the processes? For example, did particular sources of information on care management influence these things, or a particular consultant or QIO staff member? Did P4P programs other than the demonstration influence them?
- *Role of practice characteristics:* What, if any, practice characteristics have influenced the practice's view on care management, its decision to adopt care management processes, or the smoothness of implementation of the processes? For example, the characteristics of its patients? The views or skills of its administrative staff? How busy the practice is at present? How profitable? Its comfort level with HIT? With care management?

VI. Greater Use of Data to Refine the Care Process

- *Number of measures available:* Has the number of measures and types available for review changed?
- *Key information sources:* Who generates clinical measures for the practice (if used), and what patient populations from within the practice are included? How often are the measures generated and at what level of detail? Are results broken down by physician? If so, does each physician in the practice see the others' scores?
- *Perception of available benchmarks:* What benchmarks are available, and how useful are the benchmarks perceived to be? Why?
- *MCMP feedback (for demonstration practices*): Has feedback on performance from the MCMP demonstration been useful? If so, how?
- *Impact on care process:* If data are being used more, has this led to any changes in the care process?

VII. Enhanced Practice Orientation to Quality and Safety

- *Awareness of performance:* How aware is the practice of its performance on quality measures? [Assess breadth and depth of this understanding.] Do the physicians in the practice meet to go over performance reports and discuss performance?
- *Perception of opportunities for improvement:* Does the practice think that there are processes that could be implemented or changes made that could further improve the quality and/or safety for patients of the practice? Which changes are viewed as potentially most important?
- *Changes made and planned:* What, if any, changes has the practice made to improve quality or safety for its patients? What, if any, additional changes are under way? Planned?
- *Probe for practices highly oriented to quality/safety:* What, if anything, has influenced the practice to increase the focus on quality improvement?

HIT = health information technology.

example, on time spent per patient visit, on documentation, on usefulness of information during a patient visit, on administrative versus clinical time).

4. ***Relevant Context—Other Incentive and Reporting Programs.*** Have other incentives that the practice faces from other payers or reporting programs (such as the PQRI) affected how the practice has responded to the incentives under the demonstration? If so, how?

5. ***Adoption of Care Management Processes.*** Questions here include how quickly and extensively the practices are adopting care management processes; what factors are affecting the adoption and speed of implementation of care management; degree of support provided by HIT used by the practice; and perceived benefits and costs of adopting care management for different conditions.

6. ***Greater Use of Data to Refine the Care Process.*** Questions include what data the practice routinely reviews on its own performance and related benchmarks, how this has changed, and whether this has had an impact on the care process.

7. ***Enhanced Practice Orientation to Quality and Safety.*** Questions are designed to assess the breadth and depth of the practice's understanding of its performance, its perception of opportunities for improvement, improvement-focused changes made and planned, and whether the practice meets as a group to review data and discuss performance.

Interviews conducted for the implementation analysis will also help us classify practices by key characteristics likely to be associated with outcomes, as discussed in Section B.8.

B. APPROACH

1. Overview

The analysis of implementation of the demonstration will rely on several data sources: (1) the Office Systems Survey; (2) site visits; (3) telephone discussions with highly successful practices; and (4) telephone discussions with unsuccessful practices, including those that withdrew from the demonstration, if any. A literature review will ensure that the site visit discussion guides are consistent with recent research and that all site visitors are knowledgeable about the latest research as we enter the field.

2. Literature Review and Review of Key Websites and Other Background

By July 2007, we will complete a literature review and review of key websites and other background material, focusing on physician practice responses to incentives and recommended changes in office practice to improve quality and efficiency by implementing EHRs or registries. We anticipate that the peer-reviewed literature will be thin but that some websites will provide case studies and tips that can help practices make positive changes as they implement EHRs, disease registries, or other information technology. In addition, we may identify (and seek to obtain) other helpful background material, such as materials used by QIOs with practices in the DOQ and DOQ-IT initiatives.

First, we will work with our library staff to search the standard databases (including DIALOG, OCLC First Search, FACTIVA, EBSCO Host, OVID, ISI's Web of Science, and PubMed) for peer-reviewed literature. In addition, we will identify and search websites that may lead us to online information designed to help physicians make the changes we are hoping to see. For example, AHRQ's National Resource Center for Health Information Technology (www.healthit.ahrq.gov), which aims to promote adoption of HIT in ways that improve quality, may contain useful background information for the study.

Key results of the literature review and review of background materials will be summarized in an internal memorandum to be shared with CMS as an appendix to our draft site visit protocols in September 2007.

3. Office Systems Survey

Data from the DOQ-IT Office Systems Survey will be important for the implementation analysis, because it can identify changes in practice across *all* the demonstration and comparison practices (that respond to the survey), as opposed to our site visits and telephone contacts, which

are limited to a subset of sites.[1] The survey will be conducted by the Maine Health Information Center on behalf of CMS twice during the demonstration for demonstration and comparison practices (at the beginning [2007] and end of the demonstration [2010]). We assume that the data from the survey will be shared with MPR for the evaluation.

Specifically, if the data are fairly complete for the demonstration and comparison sites, the survey data will allow us to identify whether demonstration practices have changed how they use electronic tools to support quality differently from the comparison practices. The hypothesis is that the demonstration's incentives will increase practices' attention as to how they can use available means, including electronic tools, to improve quality, particularly on the measured dimensions of care. While we expect them to be more advanced in their use of electronic tools than the comparison sites, due to longer exposure to assistance from the QIO in the beginning of the demonstration, it will be of interest for the evaluation to see if they progress measurably more over the demonstration period.[2] Developing a robust measure of HIT sophistication based on the survey is beyond the scope of this project. However, our hope is that CMS (or the QIOs) will already have created a summary measure based on the survey. We would use any such measure to examine relative advancement in HIT use in the demonstration group compared to the comparison group. If a summary measure is not available, we will review relative progress on the more detailed areas covered by the survey and avoid global analysis, except where we find a consistent pattern across areas.

In the final year of the demonstration, we plan to produce tables comparing the percentages of the demonstration and comparison groups performing HIT-related activities on first measurement, final measurement, and the change in these percentages, as well as the percent of

[1] Appendix B includes a copy of the latest version of the survey instrument.

[2] Most demonstration practices would have participated as DOQ pilot sites for several years.

practices with advancement in HIT use over the period for each activity. Table II.2 shows the table for demonstration-wide analysis. In addition, we will produce the table for each state for larger and smaller practices. We will include all demonstration and comparison sites that responded to both the initial and final surveys.[3] After the data for the first survey become available to us, we will calculate the "initial percentage" columns for the table for the demonstration and comparison groups (demonstration-wide) and consider whether we need to make any adjustments to the table or analysis plan. For example, there is a detailed list of activities associated with using registries, using EHRs, and e-prescribing. In each case, in addition to identifying changes in the percentage of practices undertaking each activity, we suggest summarizing by calculating the percentage of practices that conduct at least three of these activities, and at least five of them. However, the numbers "three" and "five" are somewhat arbitrary as a means to summarize overall disease registry use, use of EHRs, and use of e-prescribing; depending on practices' initial scores, it may make more sense to use "five" and "eight" instead or another alternative.

After we have conducted the broad-based analysis described above, we anticipate following up with additional analysis of two or three content areas that show especially promising results. For example, if the demonstration sites advanced especially well compared to comparison sites with respect to the EHR activities most closely associated with managing chronic conditions, we will use data on practice characteristics from the Office Systems Survey to examine results by size of practice, as well as location by state and whether the practice is located in an urban or rural area. This will allow us to identify whether the results seem to hold

[3] Separately, we will examine the implications of a dropoff in responses between the initial and final surveys, if such a dropoff occurs. Specifically, we will examine the characteristics of those in the demonstration and comparison groups that respond to both surveys compared with those that do not, to the extent our data allow.

TABLE II.2

ADVANCEMENT IN USE OF HIT DURING THE DEMONSTRATION: DEMONSTRATION-WIDE ANALYSIS

	Initial Percentage		Final Percentage		Mean Change in Percentage		Percent with Advancement in HIT Use	
Items from Office Systems Survey	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices
EHR Adoption Activities								
Percent with each activity at least half completed (3.8):[a]								
a. Perform office readiness assessment								
b. Document and analyze current office workflows								
c. Redesign office flow to meet EHR process								
d. Evaluate care management and process improvement pre-EHR								
e. Full implementation of EHR								
f. Use EHR to identify additional care management and process improvement opportunities								
Redesigning Workflows								
Percent of practices doing the following for at least 50 percent of patient visits (4.1 – 4.4):								
a. Pulling paper, charts for scheduled visits								
b. Dictate visit notes into a tape recorder or phone								
c. Dictate visit notes directly into the EHR								
d. Use a computerized (as opposed to paper) system to manage the following office workflows:								
i. Telephone calls								
ii. Prescription refills								
iii. Referrals								
iv. Results followup								

[a]Numbers in parenthesis reference the applicable question number(s) on the DOQ-IT Office Systems Survey (see Appendix B).

TABLE II.2 *(continued)*

Items from Office Systems Survey	Initial Percentage		Final Percentage		Mean Change in Percentage		Percent with Advancement in HIT Use	
	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices
Registry Use								
Percent using a registry (5.1):								
Mean number of conditions in the registry when used (5.2)								
Percentage of practices using a registry for at least half of patient visits for (5.3 – 5.12):								
a. Diabetes								
i. At least 1 task								
ii. At least 3 tasks								
iii. At least 5 tasks								
b. Coronary Artery Disease								
i. At least 1 task								
ii. At least 3 tasks								
iii. At least 5 tasks								
c. Congestive Heart Failure								
i. At least 1 task								
ii. At least 3 tasks								
iii. At least 5 tasks								
d. Hypertension								
i. At least 1 task								
ii. At least 3 tasks								
iii. At least 5 tasks								
e. Preventive Care								
i. At least 1 task								
ii. At least 3 tasks								
iii. At least 5 tasks								
Percent of practices using a registry for at least 50 percent of relevant patients for at least one condition: (5.3 – 5.12):								
a. Notify patient who are overdue for visits								
b. Prompt clinicians to order tests, etc.								
c. Remind patients about needed tests, etc.								
d. List eligible patients for each condition								

TABLE II.2 *(continued)*

Items from Office Systems Survey	Initial Percentage		Final Percentage		Mean Change in Percentage		Percent with Advancement in HIT Use	
	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices
e. List patients requiring an intervention								
f. Generate specific patient care plan								
g. Generate information for patients on their condition								
h. Create written, personalized, actions plans								
i. Prompt clinician and/or patients to review self-management plan during a visit								
j. Modify self-management plan as needed following a visit								
EHR Use								
Percentage of practices using an EHR								
a. At all (6.1)								
b. Including all patients (6.2)								
Percent of practices using an EHR for more than 50 percent of patient visits/encounters: (6.2 – 6.12)								
a. For at least 3 tasks								
b. For at least 5 tasks								
Percent of practices using an EHR for at least 50 percent of patient visits for each function: (6.2 – 6.12)								
a. Generate laboratory requisitions/orders electronically								
b. Review laboratory test results electronically								
c. Generate radiology requisitions/orders electronically								
d. Review radiology results electronically								
e. Enter data into documentation templates								
f. Review and act on reminders for care activities (for example, overdue health maintenance)								

TABLE II.2 *(continued)*

Items from Office Systems Survey	Initial Percentage		Final Percentage		Mean Change in Percentage		Percent with Advancement in HIT Use	
	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices
g. Maintain medication lists for individual patients								
h. Maintain allergy list								
i. Maintain problem and/or diagnosis list								
j. Record and review patient's family history information on the computer								
k. Trend lab and/or other test results over time								
Online Resources								
Percent of practices using online resources (7.1)								
Percent of practices using online resources for at least half the patient visits/encounters for the following tasks (7.2 – 7.5):								
a. Access online resources for patient care, review guidelines and evidence-based recommendations at the time of treatment								
b. Generate a care plan–set of mutually agreed upon goals and interventions to meet goals								
c. Produce condition-specific patient care materials								
d. Connect with patients via portal or secure email								
E-prescribing								
Percent of practices using software to generate (8.1):								
a. New prescriptions only								
b. Refills								
c. Both								
Percent of practices using electronic or handheld devices for e-prescribing (8.2 – 8:12):								
a. At least 3 activities								
b. At least 5 activities								

TABLE II.2 *(continued)*

Items from Office Systems Survey	Initial Percentage		Final Percentage		Mean Change in Percentage		Percent with Advancement in HIT Use	
	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices	Comparison Practices	Demonstration Practices
Percent of practices using electronic or handheld devices for the following e-prescribing activities (8.2 – 8:12):								
a. Identify generic or less expensive brand alternatives at the time of prescription entry								
b. Reference the drug formularies of the patient's health plans/pharmacy benefit manager to recommend preferred drugs at time of prescribing								
c. Offer guidelines and evidence-based recommendations when prescribing medication for a patient								
d. Calculate appropriate dose and frequency based on patient parameters such as age and weight								
e. Maintain a list of each patient's current medications								
f. Screen prescriptions for drug allergies against the patient's allergy information								
g. Screen new prescriptions for drug-drug interactions against the patient's list of current medications								
h. Print out prescription on a computer printer								
i. Transmit prescriptions directly to pharmacy via electronic fax (no paper printed)								
j. Transmit prescriptions directly to pharmacy via electronic means (without relying on a fax machine at either clinician's office or in the pharmacy)								
k. Provide patient-friendly information about the medication to the patient								

EHR = electronic health record; HIT = health information technology.

across the board or appear strongest for certain subgroups of the population. Although there may not be enough practices to generate enough statistical power to distinguish true differences statistically, it will still be important to mine the data we have for underlying patterns in implementation of HIT that will ultimately help us understand outcomes.[4]

4. Site Visits

Through the site visits, the evaluation (and, thus, CMS) will acquire an in-depth understanding of how practices are implementing the demonstration and how the practice changes under way may be most likely to be influencing quality of care, practice efficiency, and patient satisfaction. After careful consideration, we believe the best approach will be to conduct the site visits in two waves, with wave 1 in year 1 and wave 2 in year 3.

Visits to the Same Practices in Two Waves. A concentrated site visit effort near the beginning and end of the project (February through May 2008 and October 2009 through January 2010) will best ensure that the practices have as much time as feasible to implement demonstration-related changes and that we thoroughly understand where they were on the dimensions of interest before the demonstration. We also recommend that we visit the same practices in both time periods. This will allow us to directly observe a myriad of important factors at the two critical points in time, bringing our understanding of the demonstration implementation into sharp focus. We can directly relate what the practices, early in the demonstration, expected their goals to be to what actually occurred by near the end and how their goals changed. We can remind them of their early ambitions and probe, as appropriate, for what helped or hindered in accomplishing them. We can notice major changes in how they are using technology to support quality relative to the earlier site visit, and ask for relevant details (for

[4] A discussion of the precision of practice-level estimates from different data sources is presented in Chapter III, Section D.2.

example, perhaps they are carrying handheld devices during the wave 2 interview that we did not observe in wave 1, and we will ask them why). In addition, the physicians would likely remember the project team and early site visit positively, which could lead them to communicate more freely during the critical second wave of visits.

Conversely, if we visit different practices near the end of the demonstration, we cannot expect them to clearly recall how they operated three years earlier or what they had hoped to achieve at that point (though, of course, we would ask). The conclusions we would be able to draw for the implementation analysis would need to be considerably more tentative, because we could not be sure that any changes we were observing in the themes from the wave 1 versus wave 2 practices were influenced by selecting different practices. The recommended strategy does imply a reduction in the total number of practices visited on-site, from 76 to 40.[5] We believe that 40 is still a sufficiently large number of practices to address the implementation analysis questions credibly and provide useful information for the outcomes analysis and synthesis.

Selection of Demonstration and Comparison Sites. The 40 sites are planned to be distributed equally among the states, with 8 demonstration participant practices and 2 comparison practices for each state. The emergence of the congressionally mandated link between provider reporting of quality data and a payment increase makes it likely that both comparison and demonstration sites will be improving on the dimensions of interest during the period. Therefore, it is even more important than at the time of our original proposal that we

[5] To keep the total number of site visits the same as originally proposed, we assume that 40 practices would be visited in year 1, and 36 in year 3. We think it is likely that four of the practices originally visited will be unavailable for a second-round visit due to normal attrition (for example, retirement of physicians or consolidation of practices). If all 40 practices are available, we will interview all of them, but would expect to interview four by telephone.

perform some visits to comparison sites to see if we observe differences in the nature of change or its dynamics or pace between the demonstration and comparison sites.

Demonstration sites to be visited in each state will be selected judgmentally from among geographically feasible choices to ensure that we achieve a mix of practice sizes and urban versus rural location. We need the practices to be clustered to some degree geographically to make the site visit efficient, although they do not need to be tightly clustered. For example, the sites in a state could include four from a large urban area, one from a rural area outside that urban area, and three from or around a small urban area two hours away from the large urban area. If comparison group sites can be chosen by December 2007, we would plan to select sites so as to be able to visit comparison group sites on the same visit; that is, we would aim for areas near the border of the neighboring state where the comparison practices are located. We would also like to ensure that the selected practices vary in HIT sophistication, if any summary measure of that or proxy for it is available from CMS by late December 2007 when the selection process must begin. If no such measure is available across all the potential sites, we may tentatively select sites based on their known characteristics, then ask the QIO to comment on the relative HIT sophistication of the ones chosen to ensure substantial variation.

We have found that a good way to begin site selection with geographic components like this one is to map the potential practices, then focus on promising areas. By mapping the zip codes of all the demonstration and comparison group practices and comparing them to an atlas, one can readily find areas that would be both convenient to access and offer a sizable cluster of practices to choose from. After suitable geographic areas are located, we will construct a table with all the practices available for selection within the targeted geographic areas, showing their practice size, specialty, and urban/rural location. We assume the basic practice characteristics would come from the Office Systems Survey, with the urban/rural location identified by MPR showing the

practice's county based on its town and zip code, and then linking that to the Area Resource File. We will select a set of targeted practices and a pool of replacement practices (in case some of our initial targets fail to cooperate). We assume we would provide CMS with a memo describing the characteristics of our selected sites, as well as all the details of the process outlined here.

If comparison group sites are not identified by December 2007, we will need to move forward with selection of the demonstration sites and discuss alternative options for the comparison sites with CMS. For example, we could devote more resources than initially planned to the site visits in wave 1 by adding trips to comparison sites, then conducting the wave 2 comparison site visits by telephone. In addition, it may not be possible to meet our goal of visiting the comparison sites as well as the demonstration sites on the same visit in all the states, even if comparison group practices are identified by December 2007 (for example, in California). In such a case, we will discuss options with CMS, but would probably suggest making the first-round visit longer and adding a travel leg to it to visit the comparison sites, conducting the second round by telephone if necessary from a budget perspective.

Background Discussions with QIOs and CMS. Before each wave of visits, it would be helpful to the evaluation to have a one- to one-and-a-half hour discussion with key staff at each of the QIOs and CMS who operate the demonstration (a total of five discussions). This will allow us to enter our discussions with providers already having a solid background on the demonstration recruitment (wave 1) and operational experience (wave 2), as well as communication strategies and messages (both waves). Such discussions will greatly assist us in interpreting provider comments and probing appropriately. For example, if some providers say they are not sure how to submit the data, we will be familiar with the process and can ask whether they have reviewed communications X, Y, and Z, and, if so, what about those

communications was confusing to them. In addition, the QIO interviews will help us identify any state-specific issues to examine during the site visits.

Logistics. The first wave of site visits is planned as three four-day visits (to Arkansas, Massachusetts, and Utah) and two two-day visits (to northern and southern California). In the second wave, we anticipate one fewer practice per state (dropping from 10 to 9 practices per state). Therefore, we plan for 3 to 3.5 days on-site for Arkansas, Massachusetts, and Utah, and 2 2-day site visits for northern and southern California.

Scheduling for each visit will begin nine weeks before the target date for the visit and will be completed one month in advance. If CMS agrees, we would like to use an introductory letter signed by a CMS official to describe the study and encourage participation. The letter will be mailed nine weeks before the target date, and a follow-up call will be made a week later. Many offices will need to have the letter faxed or emailed to them again. However, we have found that it is useful to be able to reference a letter sent at least a week earlier when making the first call. A confirmation letter will be emailed to sites, and a second confirmation call will be made the week before the visit.

We anticipate that many of the practices will require us to visit during off-hours—either before clinic hours, during lunch, or after hours.[6] This leaves ample time between practice interviews for traveling to the next location, with a maximum of three practice visits probably possible on any given day. Each office visit will include a 45-minute discussion with the office manager and a 30-minute discussion with a physician.[7] We will ask to interview the physician in

[6] This is confirmed by the experience of the site visits for the evaluation of the Physician Voluntary Reporting Program.

[7] Although we will request 30 minutes, it is possible that some physicians will agree only if the time frame is shorter (for example, 15 minutes). Whether we would accept the shorter time may be a case-by-case judgment, depending on whether there are alternatives available in the area that would retain the varied mix of practices that we seek.

the practice who has been most involved with any changes in response to the demonstration. If all are equal in this regard, we will ask to interview the physician with the largest number of Medicare patients. During the discussion with the office manager, we will ask to see documentation related to the types of changes that are discussed, with any personal health information de-identified. We may selectively request copies of illustrative documents, but we will make sure they contain no identifiable information. Under no circumstances will any personal health information be recorded in our notes.

Data Collection Instrument. A semistructured interview protocol, based on the key questions provided in Table II.1, will be the central data collection instrument. We will provide a draft protocol to CMS in September 2007 and revise it following CMS comments. The protocol for the demonstration and comparison practices will be similar, except that protocol questions pertaining to the demonstration will not apply to the comparison group practices.

Staffing. Each visit will be staffed by one senior project team member and one research analyst. During the visits, the senior site visitor will be primarily responsible for covering the protocol topics. The analyst will be responsible for documenting each visit and making sure the content areas are covered.

The analyst will complete detailed notes from the site visit within two weeks of the visit, and the senior site visitor will review the notes, and modify or add to them, promptly after that. This schedule ensures the documentation is completed and reviewed while the visit is still fresh in the site visitors' minds.

5. Telephone Discussions with Highly Successful Practices

After the first and second payouts under the demonstration (not counting the payout for baseline-year reporting), we plan to conduct telephone discussions with 12 practices that benefited substantially from the program. We suggest the 12 be drawn randomly after the first

payout from among practices that received the ceiling payment amount, if there are more than 12 such practices each year. After the second payout, we would follow a similar process but ensure that the 12 selected practices are different from those we interviewed after the first payout.

As in the site visits, we will use a semistructured protocol. The protocol, which will be provided to CMS with the site visit protocol in September 2007, will focus on identifying the changes practices made to achieve their success under the demonstration, their motivation for doing so, contributing factors that may have made it easier for them to succeed relative to others, and lessons learned.

6. Telephone Discussions with Unsuccessful Practices (Including Those That Withdrew)

On a continuing basis beginning in year 2 of the demonstration, we will conduct telephone discussions with unsuccessful practices, including practices that withdrew from the demonstration, if any (up to six practices in demonstration year 2 and six in year 3). We will schedule the calls to minimize the likelihood of having an unproductive and emotional discussion with the unsuccessful practices. For instance, the best time for such calls may be around three months after a withdrawal. We expect that emotions in a practice may run high at the time of withdrawal and that waiting about three months should provide a "cooling off" period that would allow for more thoughtful discussion of all the factors leading up to the practice's decision, while still being close enough to the practice's experience under the demonstration for good recall. The protocol for the discussions with unsuccessful practices will be provided with the other protocols in September 2007. Its principal focus will be on identifying factors that led the practice to suboptimal performance or to withdraw from the demonstration, and what, if anything, could have been different that would have led the practice to remain in the study.

As noted in our original proposal, we will also use descriptive statistics to compare quantifiable characteristics between withdrawing and remaining plans. If no practices withdraw

from the demonstration, we will discuss with CMS alternative allocations of these interviews to practices that remained in the demonstration but were unsuccessful (that is, they did not receive performance payments).

7. Analysis

The implementation analysis will be conducted demonstration-wide and will include reviewing the information for qualitative and descriptive differences within subsets of practices (such as by size) and by state location. Atlas.ti software will be used to help the team organize the detailed interview information to identify themes, as well as to help identify illustrative examples as needed from among the many site visits and telephone interviews. MPR routinely uses this software to assist in analyzing large numbers of site visit interviews.

We plan to analyze the site visit data demonstration-wide on the major dimensions, and the factors influencing them, listed in Section A and on the table of key questions (Table II.1): participating practices' perspectives on the demonstration; direct responses to the incentives; adaptation of patient flow and documentation processes as HIT is implemented; relevant context; adoption of care management processes; greater use of data to refine the care process; and enhanced practice orientation to quality and safety.

For each of these dimensions, we will pull relevant blocks of text from the Atlas.ti database containing our detailed notes, and review them sorted in different ways (for example, by demonstration and comparison site, by state, by practice size, and by physician versus office manager responder). Reviewing the data in this way is the best method to identify patterns and to ensure all relevant information on an issue is considered when synthesizing across the sites. Use of Atlas.ti also allows easy exploration of alternative interpretations of a theme that may be suggested as the senior researchers on the team discuss the key findings.

As we prepare for the implementation analysis, we will establish practice classifications based on the site visit data and draw on practice characteristics identified from other data sources. (The characteristics of each practice will be stored in an Excel file.) The characteristics will then be used to explore relationships between characteristics and outcomes as those become available (see Chapter IV). The classification of practices may include the following, with exact categories set after reviewing actual demonstration data:

- Smaller versus larger practices
- Practices with higher versus lower numbers of Medicare patients with the targeted chronic conditions
- Urban versus rural practices, and state location
- Practices that were at high, medium, and low levels of HIT use at the start of the demonstration, assuming a summary-level measure is available based on the Office Systems Survey
- Practices aggressively adopting care management processes for one or more conditions in the demonstration versus those not doing so
- Practices in higher- versus lower-income areas, and/or those whose patient loads include a lower proportion of Medicaid and uninsured patients

In analyzing the demonstration's implementation experience, we will remain sensitive to the possibility that patterns of experience and response may differ across these types of practices. In addition, to support the evaluation's synthesis analysis (see Chapter IV), midway through the second wave of site visits, we will create additional classifications of practices that we believe may be associated with outcomes. The specific classifications that will be most useful are impossible to predict before the later stage of the evaluation, but our telephone interviews with especially successful practices, combined with the site visits, might suggest that practices that have adopted one particular type of system, or those that have been most aggressive with one or two dimensions of response, are likely most successful in improving outcomes. We will then

classify the site visit practices on those suspected factors for success, so that the outcomes for those with and without those characteristics could be compared.

Because the demonstration is being implemented at the state level, the practice characteristic of state location will receive special attention. Immediately after each site visit in a state (or after both the California visits), the senior site visitor and research analyst will meet to discuss highlights and key points emerging from the set of visits as a whole. After creating an outline as the output from this meeting, they will prepare a brief site visit summary highlighting key findings for that state, which will become part of our implementation report. As noted above, we will look for patterns in demonstration experience by state (as well as other characteristics) as we analyze each major topic. In addition, we will perform descriptive analysis of the Office Systems Survey data by state using the table format shown in Table II.2.

As themes emerge in the implementation analysis, we will employ several means to display and summarize them for the implementation report. Summary tables showing the numbers of interviewed practices that reported something of interest are one important way to summarize the analysis for the reader, even though such results cannot be presumed to be generalizable due to small sample size and nonrandom site selection. For example, a table may summarize the number of site visit practices that we would classify as undertaking high-, medium-, or low-level responses to the demonstration by the time of the first site visit, broken down by other practice characteristics noted above.

In addition, text tables can list illustrative examples of specifics that lend credibility and clarity to overall statements. For example, we could create a table with columns for facilitating factors and barriers, divided by the types we have found most common, and within those parts of the table dedicated to each type of barrier or complaint, list near-verbatim quotations from our

notes to illustrate more specifically how X, Y, and Z were barriers and how U, V, and W were facilitating factors.

The results from the analysis of implementation will first be presented in the implementation report, due to CMS in draft in July 2008, with a final version by September of that year. Results from the telephone discussions with successful and unsuccessful practices in demonstration years 2 and 3 will be incorporated into the second and third interim synthesis reports, drafts of which are due in October 2009 and October 2010, respectively.[8] Results from the second wave of site visits will be reported first in the site visits report, due in April 2011. The final, comprehensive analysis of implementation, assessing changes between wave 1 and wave 2 site visits, as well as all the other data sources, will be reported in the final evaluation report, due in September 2011.

[8] As discussed in Chapter V, the third interim synthesis report will be submitted as the Report to Congress.

III. DESIGN OF THE IMPACT ANALYSIS

Estimating impacts of the MCMP demonstration will require a rigorous research design, data from several sources on the outcomes the P4P intervention is expected to influence, and strong statistical models to provide unbiased and efficient estimates of program impacts. Several factors make this task challenging, including (1) the need to rely on a quasi-experimental design; (2) the need for separate impact estimates for each state; and (3) the considerable variation of many factors across states, including the timing and intensity of technical assistance for implementing EHRs, P4P and EHR penetration, physician licensure regulations, and accuracy and completeness of key identifiers in claims data.

A. RESEARCH DESIGN

The two key features for ensuring that valid estimates of impacts are obtained are (1) the comparison group strategy (identifying a sample of practices that will yield reliable estimates of what would have occurred to demonstration practices and beneficiaries without the P4P incentives), and (2) an adequate sample size. We will estimate impacts of the demonstration through a *difference-in-differences* approach. With this approach, we will compare changes in quality measures and other outcomes of *practices* in the demonstration states and comparison states before and after the start of the demonstration.

The impact analysis will use a *comparison group* (or *quasi-experimental*) design. To identify the comparison group, the evaluation will choose DOQ-IT physician practices in selected nondemonstration states that match most closely those in demonstration states on key factors likely to be associated with outcomes of interest and, where possible, on predemonstration values of the outcomes themselves. DOQ-IT practices in nondemonstration

states provide an ideal counterfactual for demonstration practices, because all demonstration practices will also participate in DOQ-IT (due to demonstration eligibility requirements). Thus, demonstration and comparison group practices will both receive the technical assistance provided by DOQ-IT to adopt HIT, including an EHR system. In this section, we discuss how we (1) identified potential comparison states, (2) will select comparison group practices, and (3) will identify comparison group beneficiaries in each practice. Finally, we provide an overview of how we will estimate impacts.

1. Selection of Comparison States

Using a reproducible process, we selected nondemonstration states using criteria that aimed to identify states with environments similar to those of the demonstration states in that they at least had EHR and P4P programs.[1] Based on this selection process, we proposed the following states be used as comparison states for the MCMP demonstration states: *Arkansas:* Nebraska, with Texas as alternate; *California:* for comparison to southern California only, Arizona; for comparison to California overall, Oregon, with Washington as alternate; *Massachusetts:* New York, with Connecticut as alternate; and *Utah*: Idaho. Although the comparison states chosen have face validity and meet the criteria used for selection, they are only an approximate match to the demonstration states.

2. Selection of Comparison Group Practices

To be considered a valid comparison practice, the practice's patients must have predemonstration service use and cost patterns similar to those of the practices in demonstration states. The practice also should have comparable baseline characteristics. Practice size and

[1] Appendix C describes in detail the process for selecting comparison states.

experience with HIT are key determinants of practice behavior (Miller and Sim 2004). Therefore, we will first stratify the sample by constructing cells defined by the combination of these practice characteristics (that is, whether the practice has one physician, two physicians, or three or more; and whether the practice has experience with an EHR system).[2] Within each cell, we will use statistical matching methods to identify the comparison practice that best matches each demonstration practice in terms of predemonstration service use measures, costs, and baseline characteristics.

Two methods for selecting the "closest" match for each demonstration practice are the *caliper* and *nearest-neighbor* methods.[3] From our experience selecting comparison groups for other demonstrations, the caliper method is a more efficient approach than the nearest-neighbor method when matching a large number of units on a limited number of characteristics, as in the case of MCMP. Therefore, we plan to use the caliper method as our primary method for identifying comparison practices in nondemonstration states. Ideally, we will have several suitable comparison practices within each stratification cell for each demonstration practice. If so, we will select the comparison practice that provides the closest predemonstration match to a demonstration practice.

We will assess the validity of the matches by testing whether there are significant differences between demonstration and comparison practices in the changes in outcome

[2] We will consider alternate ways to stratify the sample if there is not at least one comparison practice to match to each demonstration practice within each cell.

[3] The caliper method identifies all potential comparison group units whose weighted average of characteristics falls within a specified range, or "caliper," of the weighted average of the characteristics to which they are being matched. The size of the caliper is typically defined in percentage terms (for example, ± 1 percentage points of the weighted average of the target practice). The nearest-neighbor method identifies potential comparison group units with the closest absolute difference, on a composite measure, relative to the unit to which it is being matched. This method assumes that all characteristics used for matching are combined into a single score or distance, which raises the problem of determining the weights for calculating the overall score.

measures during the baseline period. (We will only be able to perform this test with outcome measures available during the predemonstration period, such as those based on Medicare claims data.) If the matches are valid, we would not expect to see significant differences.

The measures we plan to use to match practices (in addition to practice size and experience with HIT) include the number of Medicare fee-for-service beneficiaries served by the practice (ideally, for each target condition), number of evaluation and management (E&M) visits per beneficiary in the practice, number of hospital admissions per beneficiary in the practice, and Medicare expenditures per beneficiary. However, the final list of baseline characteristics will depend on the availability of specific data elements in the Office Systems Survey. To construct the measures of service use and expenditures, we will use claims data to be supplied by the financial support contractor (Actuarial Research Corporation [ARC]).

3. Identification of Comparison Group Beneficiaries

To link beneficiaries to comparison group practices, the demonstration's financial support contractor will use provider identification numbers available in claims data and the practices' demonstration or DOQ-IT application forms, as well as an algorithm for allocating beneficiaries to only one practice. For demonstration practices, this procedure avoids double-counting beneficiaries on whom the incentive payments will be based. For both demonstration and comparison practices, the algorithm assigns each beneficiary represented in the claims files to the practice that provided the plurality of E&M services during the reporting period. As a tiebreaker for beneficiaries seen by more than one such practice, the algorithm assigns the practice with the most recent E&M visit, the practice with the highest Medicare expenditures for that beneficiary in the previous year, and whether it is a demonstration practice (only for practices in

demonstration states).[4] Finally, because only fee-for-service Medicare beneficiaries are allowed to participate in the demonstration, comparison group beneficiaries will not include those in Medicare Advantage plans.

4. Estimation of Demonstration Impacts

We will use the difference-in-differences method to estimate impacts on claims-based outcomes, such as use of Medicare-covered services and costs. This regression-based method implicitly accounts for all factors—both measured and unmeasured—that do not change over time when estimating impacts, and thus is likely to yield unbiased estimates. As described in Section F, we will assess whether we need to use statistical methods to control for selection bias in models that estimate impacts from survey data (such as measures of satisfaction with care).[5] If needed, we will use selection-adjusted models that attempt to account for systematic differences between practices in the comparison group and demonstration practices in preenrollment characteristics that are difficult to measure, such as the motivation to provide higher or lower quality of care.

The practice will be the *unit of analysis,* because it is the *unit of intervention.* That is, the practice—not the beneficiary—will receive the financial incentives. Furthermore, the individual physician will *not* be the unit of analysis. Our analytic sample for estimating impacts from the demonstration will consist of all beneficiaries in the demonstration and comparison group practices. In addition, we will measure impacts for important subgroups of beneficiaries for

[4] The proposed procedure codes for identifying E&M services are: 99201 through 99215 (office or other outpatient services), 99301 through 99316 (nursing facility services), 99321 through 99333 (domiciliary, rest home, boarding home, or custodial care services), 99341 through 99350 (home services), 99381 through 99397 (preventive medicine services), and 99401 through 99429 (counseling and/or risk factor reduction intervention) (Wilkin et al. 2007).

[5] Selection bias occurs when unmeasured differences between the demonstration and comparison groups affect outcomes, which results in biased estimates of demonstration effects.

which it is reasonable to expect impacts to vary, such as which chronic condition the beneficiary has.

Because the analysis is multilevel (beneficiaries will be nested within practices), we plan to use *hierarchical (or multilevel) linear models* (HLM) to estimate impacts, which will include a range of individual and practice characteristics and their interactions. HLM allows for efficient estimation of model parameters and their variances. These models will also allow us to examine the interactions between beneficiary and practice characteristics, while accounting for clustering of beneficiaries within practices.

B. EXPECTED EFFECTS

The evaluation will measure the effects of providing financial incentives (compared to not providing such incentives) to practices that have received technical assistance to implement HIT or have adopted an EHR system. We expect that the demonstration will improve physicians' adherence to the 26 recommended care guidelines rewarded by the incentive payments. However, as discussed below, we will be able to directly assess the demonstration's impacts on only 13 of these 26 measures—those that can be captured through Medicare claims or beneficiary self-reports for both demonstration and comparison practices. Therefore, we will also examine the demonstration's impacts on other measures related to quality of care that we expect to be influenced as a result of physicians adhering to the care guidelines, including those related to the process of care (such as whether the beneficiary received examinations, preventive services, and screenings) and to health outcomes (such as health-related knowledge and behaviors and hospitalizations).

By improving quality of care, the demonstration is also expected to reduce Medicare costs for hospital and emergency room services. However, the demonstration could increase costs for other Medicare services, because evidence-based practice guidelines for the target conditions

may recommend that beneficiaries receive specific care from physicians, thereby increasing the average number of physician visits, as well as other Medicare costs. Overall, the demonstration is intended to reduce total costs for Medicare services by at least enough to offset the costs for the physician incentive payments, so that the demonstration is budget neutral. Thus, we will measure the demonstration's effects on total Medicare costs, Medicare costs by type of service, and use of selected Medicare services (such as emergency room visits and hospitalizations).

To maximize profits from bonus payments, demonstration practices must provide and report on evidence-based clinical interventions as efficiently as possible. Some practices may meet these objectives by adopting HIT or using it more efficiently. Thus, we will measure the demonstration's effects on the adoption and use of HIT, in general, and EHRs, in particular.

Physicians striving to earn financial incentives might also make changes that could improve their patients' adherence to their recommendations and, in turn, their health outcomes. Such changes might include improving their interactions with patients, spending more time on patient education, and spending more time coordinating their patients' care with other providers. As a result, physicians might be more satisfied with the care they provide, and patients might be more satisfied with the care they receive. Thus, our secondary outcome measures will include those related to continuity of care, care coordination, patient satisfaction, and physician satisfaction.

C. DATA SOURCES

The impact analysis will use data from four data sources: (1) a beneficiary survey, (2) a physician survey, (3) Medicare claims and eligibility data, and (4) practice-specific data. Together, these data sources will allow us to directly capture the demonstration's impacts on a subset of the 26 quality measures for which physician practices receive incentive payments (Table III.1), as well on a wide array of the other primary and secondary outcomes that we expect the demonstration to influence (Table III.2).

TABLE III.1

DATA AVAILABILITY OF QUALITY MEASURES RELATED TO FINANCIAL INCENTIVES

	Data Source			
Measure	Medical Records	Medicare Claims	Beneficiary Survey	Data Available for Comparison Group Practices?
Percentage of patients with coronary artery disease who:				
Were prescribed antiplatelet therapy	X			No
Were prescribed a lipid-lowering therapy	X			No
Were prescribed beta-blocker therapy, among those with prior myocardial infarction	X			No
Received at least one lipid profile	X	X		Yes
Had most recent LDL cholesterol < 130 mg dl	X			No
Were prescribed ACE inhibitor therapy, among those who also have diabetes and/or LVSD	X			No
Percentage of patients with diabetes having:				
One or more blood tests for hemoglobin A1c	X	X		Yes
Most recent A1c level > 9 percent	X			No
At least one test for microalbumin (or had medical attention for existing nephropathy or microalbuminuria or albuminuria)	X	X		Yes
Dilated retinal exam	X	X		Yes
At least one foot exam	X		X	Yes
Last blood pressure measurement below 140/90mm Hg (among those receiving a test)	X			No
Most recent LDL cholesterol < 130 mg/dl	X			No
Had at least one LDL cholesterol test	X	X		Yes
Percentage of patients with congestive heart failure who:				
Had left ventricular function results recorded	X			No
Left ventricular ejection tested (among those hospitalized with heart failure)	X	X		Yes
Had weight measurement recorded	X		X	Yes
Had patient education class on disease management and health behavior change during one or more visits within a six- month period	X			No
Were prescribed beta-blocker therapy, among those who also have LVSD	X			No
Were prescribed ACE inhibitor therapy, among those who also have LVSD	X			No
Were prescribed warfarin therapy, among those with paroxysmal or chronic atrial fibrillation	X			No

TABLE III.1 *(continued)*

	Data Source			
Measure	Medical Records	Medicare Claims	Beneficiary Survey	Data Available for Comparison Group Practices?
Percentage of those with specified chronic diseases who:				
Had blood pressure measurement during last office visit	X		X	Yes
Had breast cancer screening during current or previous year, among those under age 69	X	X		Yes
Had colorectal cancer screening during recommended period	X		X	Yes
Had influenza vaccination during September through February of year prior to measurement year, among those over age 50	X		X	Yes
Had pneumonia vaccination, among those with a chronic condition over age 65	X		X	Yes

ACE = Angiotensin Converting Enzyme Inhibitor; LVSD = left ventricular systolic dysfunction.

TABLE III.2

OVERVIEW OF TYPES OF OUTCOME MEASURES AND DATA SOURCES FOR IMPACT ANALYSIS

Measure	Data Source
Primary Outcome Measures	
Quality Measures	
Outcomes directly related to financial incentives	Medicare Claims Data and Beneficiary Survey
Process measures related to care quality	Medicare Claims Data and Beneficiary Survey
Health outcomes	Medicare Claims Data and Beneficiary Survey
Medicare service use and costs	Medicare Claims Data
Use of HIT in office procedures	Physician Survey and Office Systems Survey
Secondary Outcome Measures	
Coordination and continuity of care	Beneficiary Survey, Physician Survey, Medicare Claims Data
Physician satisfaction	Physician Survey
Patient satisfaction	Beneficiary Survey

HIT = health information technology.

We will administer a mail survey (with telephone followup) of eligible beneficiaries about 19 months after the beginning of the demonstration. This survey will measure well-being (using such indicators as health status, burden of illness, and quality of life), access to care, adherence to self-care management principles, continuity of care, satisfaction with care, and awareness of the demonstration (or the DOQ-IT program, in nondemonstration states). For physicians, we will also administer a mail survey (with telephone followup) about 14 months after the start of the demonstration. This survey will measure demonstration and comparison group barriers to transforming the practices' clinical encounters with beneficiaries and other office procedures, barriers to adoption of HIT, experience implementing this type of technology, satisfaction with HIT, and experience with P4R and P4P (in the demonstration sites only). We will obtain data on use of Medicare-covered services and expenditures, indicators for whether tests were performed related to selected clinical measures upon which the financial incentives are based, and measures of continuity of care from Medicare claims data, and demographic and eligibility data from the Medicare Enrollment Database (EDB). We will use these Medicare data both to construct outcome measures and to construct regression control variables covering the baseline period. Finally, we will obtain data for the actual financial payments made to the demonstration practices. We will work closely with the financial support contractor, the QIOs, and CMS to use any additional data that would enhance the evaluation, such as participation of demonstration and comparison practices in the PQRI.

1. Beneficiary Survey

We will administer a mail survey (with telephone followup), with a goal of completing interviews with 4,800 eligible beneficiaries (600 from the demonstration group and 600 from the comparison group in each state). The survey will start 19 months after the beginning of the demonstration's operations (in or around January 2009). The financial support contractor will

provide MPR with lists of Medicare beneficiaries classified as having a primary care physician affiliated with any of the demonstration or comparison group practices for the first year of demonstration operations. We will select a sample of 6,400 beneficiaries from these lists, evenly split across demonstration and comparison practices in each state (800 beneficiaries in demonstration practices and 800 beneficiaries in comparison practices in each state). We expect to complete interviews with three-fourths of this sample, to reach our target sample size of 4,800 respondents.

We will use a "mail first" approach, in which a self-administered questionnaire will be sent to the survey sample as part of an initial mailing package. However, we will send more than one mailing before contacting beneficiaries by telephone. Sample members will first receive a packet containing (1) a letter (printed on CMS letterhead and signed by a CMS official) describing the survey, (2) the questionnaire, (3) a fact sheet of commonly asked questions and answers, and (4) pre-paid return mailing materials. Regular mailing service will be used for this mailing. Advance materials will serve several important purposes. They will (1) provide a written description of the study; (2) legitimize the study through the use of agency letterhead or inclusion of the OMB approval number; (3) alert the sample member to an impending call; (4) provide sample members with a ready reference of names and numbers to contact for additional information and/or to complete the survey; and (5) provide MPR with information on bad addresses, through the use of return service envelopes. Our use of advance letters is supported by our past experience with CMS studies, and by a study conducted by Link et al. (2003), which found that advance letters could improve cooperation rates and reduce initial refusals.

The questionnaire will be designed with a high level of sensitivity to the age of the target population. For example, written materials will have a larger print size than is typical for use with the general population. We expect that beneficiaries will be able to complete the survey in

15 minutes or less. The questionnaire, and all accompanying survey materials, will be available in both English and Spanish. Sample members with Spanish surnames and those from areas that are known to many Spanish speakers will receive survey materials in both languages. Otherwise, materials will be mailed in English only. These materials will include a separate, Spanish-language insert containing a toll-free number to call to request Spanish-language materials.

The beneficiary survey will be fielded over a 12-month period. The initial survey mailing to beneficiaries will take place in January 2009. This mailing will be supplemented with a reminder postcard, a second full mailing to nonresponders, and a second reminder postcard. In addition, about halfway through the field period, we will send a final appeal to get more sample members to return completed surveys by mail. This final appeal will use priority mail service. We expect this multi-mail strategy to yield an estimated 50 percent response rate to the mail survey (Table III.3), as suggested by other surveys in which these types of mailings have achieved high rates of response (Hassol et al. 2003).

Overall, we are targeting 4,800 completed interviews, or a 75 percent response rate, for this survey. To complete the remaining interviews needed to achieve the targeted 75 percent response, we will focus our remaining resources on a computer-assisted telephone interviewing (CATI) collection effort. We will monitor late returns of mail surveys, and such cases will be removed from the CATI sample on an ongoing basis. Interviewers will receive project-specific training to conduct the interview by telephone, including training on sensitivity toward seniors. We will staff the project with interviewers who are experienced at interviewing similar populations. Attempts to interview beneficiaries by telephone will begin after the final mail appeal, around September 2009.

TABLE III.3

PROJECTED RESPONSE FOR THE BENEFICIARY SURVEY

Data Collection Strategy	Released Sample	Projected Completed Interviews	Cumulative Completed Interviews	Response Rate
	6,400	—	—	—
Initial mailing	—	1,280	1,280	20
First postcard	—	320	1,600	25
Second mailing	—	640	2,240	35
Second postcard	—	320	2,560	40
Priority mailing	—	320	2,880	45
CATI	—	1,920	4,800	75

CATI = computer-assisted telephone interviewing.

2. Physician Survey

We will also administer a mail survey (with telephone followup) to 1,600 physicians (200 in the demonstration group and 200 in the comparison group in each state) 25 months after the start of the demonstration (in or around July 2009). The physician survey will collect data on practice and physician characteristics not captured in the Office Systems Survey (which will be administered to all demonstration and comparison practices in 2007 and 2010), changes the physician has made in response to the incentives, barriers and facilitators to HIT adoption, use of HIT in office processes, coordination of care, satisfaction with care quality, and satisfaction with Medicare financial incentives. In addition, there will be a separate module for demonstration physicians, focusing on how participation in the demonstration influenced the practice, their perceptions of the effects of the financial incentives on their practices, and their satisfaction with the demonstration.

For our physician survey sample, we will select one or more physicians from each of the 400 treatment and 400 comparison group practices, while simultaneously attempting to minimize the design effect. For solo practices, the physician will be selected with certainty. We will select a sample of 2,376 physicians—1,144 from practices in demonstration states and 1,232 from practices in comparison states, evenly split across the four states. This sample should yield 800 respondents from demonstration states and 800 respondents from comparison states, assuming response rates are 70 and 65 percent, respectively. We are projecting a lower response rate for the comparison states because comparison group physicians will have no clear incentive to participate in a survey. Our response rate assumptions are consistent with our recent experience interviewing physicians whose patients were participating in CMS's care coordination or disease management demonstrations. No financial incentive will be offered.

As in the beneficiary survey, we will use a self-administered mail questionnaire as our first approach to surveying physicians. We selected this approach because we believe that physicians' busy schedules may make it difficult for them to respond to an unscheduled telephone survey. Furthermore, we are not confident that a "cold call" to a physician's office will get past the gatekeeper. Therefore, including the survey instrument in the essential initial mailing is most efficient. We will mail survey materials to demonstration and comparison group practices using official CMS letterhead and envelopes. These survey materials will include a cover letter signed by a CMS official, a mail questionnaire, and pre-paid return mailing materials. The advance letter will include a toll-free number giving physicians the option to call and complete the survey by telephone.

The initial mailing to physicians will occur in or around July 2009. Two weeks after the initial mailing, we will begin telephone contact to schedule appointments and conduct interviews with sampled physicians. This effort will continue throughout the 11-month field period, from July 2009 through June 2010. We will train staff experienced in interviewing physicians and other professionals to negotiate access with gatekeepers and to conduct the estimated 10-minute survey interview. About midway through the field period, we will send a second mailing appealing to physicians who have not completed surveys or scheduled appointments. We expect that about 60 percent of the completed surveys will come from CATI and that 40 percent will be completed by mail.

3. OMB Clearance

We will develop draft instruments for the beneficiary and physician surveys for CMS review and approval. Instrument content will be refined through discussions with the CMS evaluation project officer and will draw on other relevant surveys, as discussed in Section E. We will submit draft instruments to CMS in November 2007. We will revise the instruments based on

feedback from CMS and will prepare an OMB clearance package for submission in January 2008.

Both instruments will be thoroughly tested before seeking OMB approval. MPR will conduct nine pretests of the physician survey with physicians serving populations similar to those in MCMP. We will also conduct nine pretests of the beneficiary survey. The pretest sample for the beneficiary survey will consist of Medicare beneficiaries with conditions that the demonstration is targeting. The instruments will be cognitively tested to ensure that the target population will understand terms and phrases used, as well as for question sequencing, skip logic, print size, and burden. The pretests will replicate plans for the main data collection to the fullest extent possible.

4. Medicare Claims and Eligibility Files

We will obtain Medicare claims and eligibility files for all Medicare beneficiaries classified as having a primary care physician affiliated with one of the demonstration practices. The files will be obtained from two sources. First, the financial support contractor (ARC) will supply *unadjusted* claims data for all Medicare beneficiaries assigned to demonstration practices. The data will cover calendar year 2006 (the *baseline* period) and each of the three years of demonstration operations. In addition, we will receive claims data for the baseline period for all beneficiaries in nondemonstration states that we classify as being in DOQ-IT practices in these states. As described in Section A.1 of this chapter, we will use these data to match demonstration and potential comparison group practices using measures of Medicare-covered service use and expenditures during the baseline period and baseline characteristics. After we select comparison group practices, the financial support contractor will continue to supply data

for these practices for all three years of the demonstration.[6] The financial support contractor will obtain the claims data from a monthly TAP of the National Claims History (NCH) File. For demonstration practices, we expect to receive the data for the baseline period within three months of the start of the demonstration and, for each of the three demonstration years, at about the time the financial support contractor makes payments to the practices based on their performance during the preceding reporting year.[7] We estimate that there will be a 12-month lag in obtaining these data.

Second, we will use Medicare HIC numbers provided by ARC to develop a finders file, or list of beneficiaries on whom data will be requested from CMS. Our current plan is to obtain beneficiaries' demographic characteristics (age, sex, race), date of death, Medicare entitlement, HMO enrollment, reason for Medicare entitlement, and dual eligibility status from the EDB. In addition, we will use the most recent EDB file to obtain contact data for the beneficiaries in the demonstration or comparison group practices to conduct the beneficiary survey (see Section C.1).[8]

We will use Medicare claims data to construct measures of Medicare-covered service use and reimbursement by type of service (inpatient hospital, skilled nursing facility, home health, hospice, outpatient hospital, emergency room, and physician and other Part B providers) for both

[6] The definition of a comparison practice in terms of its constituent physicians *will not change during the demonstration period*. However, beneficiaries may change from period to period as they see other providers. Therefore, the claims data that we get from the financial support contractor for both demonstration and comparison practices may include different beneficiaries during the three-year demonstration period.

[7] For comparison practices, we will receive baseline data after we select comparison practices in nondemonstration states in winter 2008. For each of the three demonstration years, we will receive the data at about the same time we receive them for demonstration practices, as noted above.

[8] If necessary, we plan to supplement the claims data described above with *final-action* claims from the Standard Analytic Files and NCH/National Medicare Utilization Database. We will extract these data for both the baseline and demonstration years for beneficiaries assigned to demonstration or comparison group practices. For the impact analysis, we will assume a six-month lag between the receipt of a Medicare-covered service and its appearance on these files.

before and after the start of the demonstration. We will develop rules to assign hospitalizations (or other episodic care) to the preenrollment period if the beneficiary was hospitalized (or in the middle of an episode of care) on the day the demonstration started. We will assign expenditures and service use this way because, in practice, the physician assigned to the beneficiary would not be able to influence outcomes until the stay for that beneficiary was over. Thus, the costs of the identifying hospitalization (or other episodic care), which may be substantial, will be counted as predemonstration costs. We will use the same approach for demonstration and comparison practices.

5. Practice-Specific Data

Practice-specific measures for the impact analysis will be drawn primarily from the Office Systems Survey and from financial incentive payment data. As noted, we will work closely with the financial support contractor, the QIOs, and the Maine Health Information Center to obtain these data.

We discussed the systems assessment survey data in Chapter II. Therefore, we state here only that we will use some of these data as control variables in our impact analyses, as well as to define subgroups of interest (as described in Section F of this chapter), and in our synthesis analysis.

We will use data on the financial payments to demonstration practices to estimate the impact of P4P on Medicare costs. Subsequently, we will use these data in the evaluation synthesis (see Chapter IV). We will receive payment data from the financial support contractor at the end of each demonstration year. The payment data will include the scores for each of the 26 clinical measures on which the payments will be based. For each of the chronic conditions, the data also will include the score, the payment per beneficiary, and the number of beneficiaries on which payment was made. Finally, the database will also include any bonus payments for electronic

submission of clinical data (Wilkin et al. 2007). We will use the practice ID to link this database to other practice-specific files.

D. SAMPLE SIZES

The demonstration's budget and the number of practices likely to enroll in the DOQ-IT program influence the minimum number of beneficiaries and physicians required for detecting demonstration impacts with the desired statistical power and precision. Here, we first discuss the statistical precision that will be obtainable for the analysis of survey-based outcomes and claims-based outcomes, including subgroup analyses. We then discuss the precision for descriptive estimates of clinical outcomes among demonstration practices.

1. Minimum Detectable Differences for Impact Estimates Derived from the Beneficiary Survey, Physician Survey, and Claims Data

For binary outcomes, we will be able to detect substantively important differences with 80 percent power with the proposed sample sizes. For the beneficiary survey, we will select a sample of 6,400 of the patients served by demonstration and comparison group practices.[9] Assuming that 4,800 of these beneficiaries respond to the survey (600 in the demonstration practices in each state and 600 in comparison group practices in each state), we will be able to detect a difference in a binary outcome (with mean equal to 50 percent) of about 8 percentage points in within-state analyses and of about 4 percentage points in analyses that pool all states together (assuming 80 percent power and 5 percent level for a two-sided test; Table III.4).

[9]We assume the planned 800 demonstration practices will serve an average of 500 Medicare beneficiaries per practice.

TABLE III.4

MINIMUM DETECTABLE DIFFERENCES FOR BINARY AND CONTINUOUS OUTCOMES DERIVED FROM THE BENEFICIARY AND PHYSICIAN SURVEY AND CLAIMS DATA

Data Source	Arkansas (150 Demonstration Practices)	California (250 Demonstration Practices)	Massachusetts (250 Demonstration Practices)	Utah (150 Demonstration Practices)	All States Pooled (800 Demonstration Practices)
Binary[a]					
Physician Survey[b]	20.4	15.6	15.6	20.1	8.8
Beneficiary Survey[c]	8.1	8.1	8.1	8.1	4.2
Claims Data[d]	0.7	0.6	0.6	0.7	0.3
Continuous[e]					
Claims Data	2.3	1.8	1.8	2.3	1.0

[a]Based on the comparison between demonstration and comparison groups for a binary variable with mean equal to 0.50, given a 5 percent level for a two-sided test and 80 percent power. MDEs are expressed in percentage points.

[b]Calculations assume that there will be the same number of respondents from demonstration practices in each state as the number of practices in that state, and the same number of respondents from comparison-group practices in each state as the number of practices in that state. Calculations also assume that, in addition to the strata defined by the practices, we will stratify the physician sample according to practice size, and that the proportion of physicians selected within each stratum will be the same as the actual proportion of physicians in each practice-size stratum among practices that are eligible to participate in each state.

[c]We assume there will be an average of 500 Medicare beneficiaries per practice. From these beneficiaries, we assume there will be 600 respondents from demonstration practices and 600 respondents from comparison-group practices in each state. Respondents are not stratified by practice. We used the following two strata for chronic conditions: (1) claims that included at least one code for coronary artery disease, congestive heart failure, and/or diabetes; and (2) all other claims.

[d]Calculations assume that all claims are selected from each practice, and that there are an average of 500 Medicare beneficiaries per practice. We used the following two strata for conditions (1) claims that included at least one code for coronary artery disease, congestive heart failure, and/or diabetes; and (2) all other claims.

[e]Based on the comparison between treatment and comparison groups for a continuous variable with a 2.5 coefficient of variation, given a 5 percent level for a two-sided test and 80 percent power. MDEs are expressed as a percent of the comparison group mean.

Moreover, when using claims data, we will be able to detect even smaller differences in binary outcomes between the demonstration and comparison groups (less than one percentage point for within-state analyses and less than one-third of a percentage point for analyses that pool all states together). With the full sample of physicians,[10] and 80 percent power, our detectable differences in binary outcomes between the demonstration and comparison groups are large (about 16 to 20 percentage points for within-state analyses and about 9 percentage points for analyses that pool all states) but still sufficient for identifying major impacts. Finally, for continuous expenditure variables derived from Medicare claims data, we will be able to detect differences of about two percent between demonstration and comparison group beneficiaries (assuming a coefficient of variation of 2.5 and 80 percent power) in within-state analyses and of about one percent for the sample that is pooled across states.

2. Precision for Descriptive Estimates of Clinical Outcomes Among Demonstration Practices

To examine changes in practice performance, and the correlation of these trends with practice characteristics, we will consider the practice as the unit of analysis. The rest of this section presents precision estimates for binary and continuous outcomes for each state and for a pooled analysis of all states, as well as for subgroups representing 15, 50, and 80 percent of the full sample.

For binary outcomes, the half-width for the 95-percent confidence interval is less than one-half of a percentage point for within-state analyses, and is about one-fifth of a percentage point

[10]Our calculations assume that there will be the same number respondents from demonstration practices in each state as the number of practices in that state and the same number of respondents from comparison-group practices in each state as the number of practices in that state. Calculations also assume that, in addition to the strata defined by the practices, we will stratify the physician sample according to *practice size*, and that the proportion of physicians selected within each stratum will be the same as the actual proportion of physicians in each practice-size stratum among practices that are eligible to participate in each state.

for the sample that is pooled across states (Table III.5). For continuous variables,[11] the half-width for the 95-percent confidence interval is less than 1.5 percent for within-state analyses, and is less than 1 percent for the sample that is pooled across states.

Subgroup analyses will have less precision, especially for continuous variables. For example, in within-state analyses, the half-width of a 95-percent confidence interval for a binary variable for a subgroup comprised of half the sample is less than 1 percentage point, while the corresponding half-width for a continuous variable ranges from 1.9 to 2.4 percent of the comparison-group mean.

E. OUTCOME MEASURES

The evaluation will estimate demonstration impacts on the range of primary measures (including care quality measures directly or indirectly related to the financial incentives, health outcomes, use of HIT in office procedures, and Medicare cost and service use) and secondary outcomes (including care continuity and patient and physician satisfaction with care) that the demonstration is expected to influence (Table III.2). These outcomes are measured at the practice, physician, or beneficiary level. We will use them to construct identical outcome measures for (1) beneficiaries and physicians associated with demonstration practices, and (2) those associated with comparison practices.

To minimize burden on physicians and beneficiaries, we will construct outcomes from claims data to the extent possible. Medicare claims, however, only provide information about services that can be billed to Medicare. Due to the P4P incentives, physicians might make changes that could improve beneficiaries' health outcomes, but that cannot be billed to Medicare, such as spending more time on patient education or on communicating with other providers. For

[11] We assume the coefficient of variation for continuous variables at the practice level is 1.75.

TABLE III.5

HALF-WIDTH, 95-PERCENT CONFIDENCE INTERVALS FOR BINARY AND CONTINUOUS OUTCOMES DERIVED FROM A DESCRIPTIVE ANALYSIS OF OUTCOMES USING THE PHYSICIAN PRACTICE AS THE UNIT OF ANALYSIS

Sample	Arkansas (150 Demonstration Practices)	California (250 Demonstration Practices)	Massachusetts (250 Demonstration Practices)	Utah (150 Demonstration Practices)	All States Pooled (800 Demonstration Practices)
Binary[a]					
Full sample	0.4	0.3	0.3	0.4	0.2
Subgroup Size:					
80 percent	0.5	0.4	0.4	0.5	0.2
50 percent	0.7	0.5	0.5	0.7	0.3
15 percent	1.4	1.1	1.1	1.4	0.6
Continuous[b]					
Full sample	1.4	1.1	1.1	1.4	0.6
Subgroup Size:					
80 percent	1.7	1.3	1.3	1.7	0.7
50 percent	2.4	1.9	1.9	2.4	1.0
15 percent	4.9	3.8	3.8	4.9	2.1

Note: Calculations assume 218 beneficiaries are sampled from (on average) 500 Medicare beneficiaries per practice.

[a]Based on estimates for a binary variable with mean equal to 0.50. Half-widths for binary variables are expressed in percentage points.

[b]Based on estimates for a continuous variable with a 1.75 coefficient of variation. Half-widths for continuous variables are expressed as a percent of the mean.

example, providers might spend more time educating patients with diabetes or with coronary artery disease on diet and exercise to improve their cholesterol levels. It is essential for the evaluation to collect survey data on such indicators. Thus, we will rely on the beneficiary survey for details about certain care processes, self-care indicators, interactions with their primary care physician, and satisfaction with care. Similarly, we will use the physician survey for key information about use of HIT in office procedures and satisfaction with the quality of care beneficiaries receive. Where possible, we will use claims- and survey-based indicators of quality of care that have been developed and tested by other researchers.

1. Quality of Care

The MCMP demonstration anticipates that financial incentives from Medicare will lead physicians to adopt and use HIT and improve the quality of care they provide to beneficiaries with chronic illnesses by transforming their clinical encounters with these beneficiaries and other office procedures. Improvements in care quality should be reflected in both process and health outcome measures.

a. Process Measures

Clinical Interventions. Process-of-care measures reflect clinical interventions, such as examinations, preventive services, and screenings, that are provided to beneficiaries in ambulatory settings. In a P4P environment, we would expect beneficiaries to be more likely to receive interventions related to the quality measures upon which the incentives are based than they would be in an environment without P4P incentives. An essential group of process measures is the performance of clinical interventions that are known, or strongly believed, to be effective in preventing morbidity and mortality. A few such measures are generic (for example, influenza and pneumonia vaccinations, colorectal cancer screenings, and breast cancer

screenings in women). Other processes, such as the performance of hemoglobin A1c tests or dilated retinal exams in beneficiaries with diabetes, are disease-specific. As noted, MCMP demonstration practices will be eligible to receive bonus payments based on up to 26 clinical intervention process measures.[12] Because 13 of the 26 measures can be captured through Medicare claims or survey data, they will be available for both demonstration and comparison practices, at the practice level, and we will use them to estimate demonstration impacts on care quality.[13] We will collect six process measures, which cannot be captured through claims, in the beneficiary survey. Two measures are condition-specific: whether beneficiaries with diabetes received foot examinations from their primary care physician and whether beneficiaries with congestive heart failure and coronary artery disease said their primary care physician examined their heart and lungs with a stethoscope during their last office visit. Three additional generic measures are influenza and pneumonia vaccinations and colorectal cancer screening. Although these preventive measures could be derived from claims data, we decided to include them in the beneficiary survey to be consistent with the conventions adopted by the implementation contractor for claims-based measures. Tables III.6 and III.7 show process-of-care measures.

Use of HIT in Office Procedures. Demonstration practices may introduce or increase the use of HIT in their daily office procedures, because HIT is thought to facilitate the provision of high-quality ambulatory care. Because of the financial incentives provided by the demonstration, we expect demonstration practices will be more likely than comparison practices to invest in technology to help physicians keep medical records, access test results, consult with

[12] Table III.1 lists the 26 clinical measures to be used in the MCMP demonstration.

[13] The remaining 13 quality indicators will be available for demonstration practices only. The evaluation will use them for descriptive (as opposed to comparative) purposes. The implementation contractor (RTI International) will provide these data to MPR.

TABLE III.6

CARE PROCESSES USED IN CLINICAL INTERVENTIONS, MEASURED AT THE PRACTICE LEVEL

Measure	Source of Measure	Data Collection Method
Among All Beneficiaries with Chronic Illness	CMS	Medicare Part B Claims Processed by RTI
Breast Cancer Screening		
Among Beneficiaries with Chronic Heart Failure	CMS	Medicare Part B Claims Processed by RTI
Left Ventricular Ejection Fraction Testing		
Among Beneficiaries with Diabetes	CMS	Medicare Part B Claims Processed by RTI
Dilated retinal exam		
Blood test for hemoglobin A1c		
Urinalysis for microalbumin		
LDL cholesterol testing		
Among Beneficiaries with Coronary Artery Disease	CMS	Medicare Part B Claims Processed by RTI
Lipid profile		

Note: The outcomes in this table are a subset of 26 clinical quality measures upon which MCMP demonstration bonus payments may be based. Because they can be captured in Medicare claims, they will be available for both demonstration and comparison practices.

CMS = Centers for Medicare & Medicaid Services; RTI = Research Triangle Institute.

TABLE III.7

CARE PROCESSES USED IN CLINICAL INTERVENTIONS, MEASURED AT THE BENEFICIARY LEVEL

Measure	Source of Measure	Data Collection Method
Among all beneficiaries with chronic illness, whether blood pressure, height, and weight were measured during last visit to PCP	BRFSS	Beneficiary Survey
Whether beneficiary received appropriate colon cancer screening test within recommended time period	CoCA	Beneficiary Survey
Whether PCP asked if the beneficiary has ever received a pneumonia vaccination	CoCA	Beneficiary Survey
Whether beneficiary received influenza immunization during September through February during the previous year	CoCA	Beneficiary Survey
Among beneficiaries with congestive heart failure or coronary artery disease, whether PCP examined beneficiary's heart and lungs with stethoscope during last office visit	CoCA	Beneficiary Survey
Among beneficiaries with diabetes, whether PCP examined beneficiary's feet with monofilament during last office visit	CoCA	Beneficiary Survey

BRFSS = Behavioral Risk Factor Surveillance Survey of the Centers for Disease Control and Prevention; CoCA = patient survey developed by Mathematica Policy Research, Inc. for the evaluation of the Medicare Coordinated Care Demonstration (Ensor et al. 2003a); PCP = primary care physician.

beneficiaries outside the traditional office visit, communicate with other health care providers, issue reminders to patients, and guide their own clinical decisions and appropriate follow-up care. If the financial incentives of P4P are large enough, physicians in demonstration practices may report fewer barriers to HIT adoption than physicians in comparison practices. We plan to base measures of HIT use and barriers to HIT adoption on a recent physician survey funded by the Commonwealth Fund (Tables III.8 and III.9). That survey, like ours, was a mail survey with telephone followup.

To maximize profits from bonus payments, demonstration practices must provide and report on evidence-based clinical interventions as efficiently as possible. Some practices may meet these objectives by adjusting practice staffing and workflow—the procedures and resources used to perform clinical and nonclinical tasks. For example, practices might (1) change the composition of staff; (2) assign tasks to staff members who had not performed them before; (3) focus more attention on collecting and reviewing data on care quality for Medicare beneficiaries with chronic illness; and (4) attempt to shift the volume of patient encounters conducted as office visits, telephone calls, or emails to meet patient needs for information and consultation and increase the probability of adherence to self-care. We will examine staffing and workflow measures of these types (shown in Tables III.10 and III.11) because they may help us identify the mechanisms underlying P4P's effects, if any, on the 26 clinical quality indicators.

(As noted, we have proposed to measure the outcomes described in this section mostly for their potential contextual value in interpreting the outcomes of primary interest to CMS. However, if these measures of HIT use, barriers to use, and workflow overlap substantially with the Office Systems Survey being administered to demonstration practices in 2007 and 2010, and to comparison practices in 2007, we will omit them from our physician survey.)

TABLE III.8

USE OF HIT IN OFFICE PROCESSES, MEASURED AT THE PHYSICIAN LEVEL

Measure	Source of Measure	Data Collection Method
Whether and How Uses EHRs, Routinely, Occasionally, or Plans to Within Next Year	CWF	Physician Survey
Whether Accesses Test Results Electronically, Routinely, Occasionally, or Plans to Within Next Year		
Whether Consults with Beneficiaries by Telephone, Routinely, Occasionally, or Plans to Within Next Year		
Whether Consults with Beneficiaries by Email, Routinely, Occasionally, or Plans to Within Next Year		
Whether Communicates with Other Providers by Email, Routinely, Occasionally, or Plans to Within Next Year		
Whether Uses Clinical Decision Support Tools, Routinely, Occasionally, or Plans to Within Next Year		
Currently Issues Reminders to Patients, by Computerized System, by Manual System, or Plans to Within Next Year		
Currently Uses Follow-Up Alerts, By Computerized System, by Manual System, or Plans to Within Next Year		

Note: The CWF is the source for all the proposed measures.

CWF = Commonwealth Fund survey of physicians (Audet et al. 2005); EHR = electronic health record; HIT = health information technology.

TABLE III.9

BARRIERS TO HIT ADOPTION, MEASURED AT THE PHYSICIAN LEVEL

Measure	Source of Measure	Data Collection Method
In Deciding Whether to Implement HIT:	CWF	Physician Survey
Start-Up Costs a Major Barrier, Minor Barrier, or Not a Barrier		
Lack of Time to Acquire, Implement, and Use a New System a Major Barrier, Minor Barrier, or Not a Barrier		
Maintenance Costs a Major Barrier, Minor Barrier, or Not a Barrier		
Lack of Evidence of Effectiveness of HIT a Major Barrier, Minor Barrier, or Not a Barrier		
Patient Privacy Concerns a Major Barrier, Minor Barrier, or Not a Barrier		
Lack of Training/Knowledge of How to Use HIT Among Clinical and/or Administrative Staff a Major Barrier, Minor Barrier, or Not a Barrier		

Note: The CWF is the source for all the proposed measures.

CWF = Commonwealth Fund survey of physicians (Audet et al. 2005); HIT = health information technology.

TABLE III.10

STAFFING AND TASKS, MEASURED AT THE PHYSICIAN LEVEL

Measure	Source of Measure	Data Collection Method
In the Past 12 Months:	Draft	Physician Survey
Whether Number of Full-Time Equivalents in Physician's Practice Has Increased, Decreased, or Stayed the Same for: Physicians Physician Assistants Nurse Practitioners Registered Nurses Administrators Business Managers Office Managers Other: Please Specify		
Whether the Number of Office Locations Associated with Physician's Practice Has Increased, Decreased, or Stayed the Same		
Whether, for the Sake of Efficiency or Otherwise Improving Office Workflow, Staff Members in the Following Positions Began Performing Clinical or Nonclinical Tasks They Had Not Performed Before:[a] Physicians Physician Assistants Nurse Practitioners Registered Nurses Administrators Business Managers Office Managers Other: Please Specify		

[a]Tasks will be specified. Examples include taking medical histories, submitting prescriptions to pharmacies, and authorizing prescription refills.

Draft = questions we will draft for this survey.

TABLE III.11

OFFICE PROCESSES, MEASURED AT THE PHYSICIAN LEVEL

Measure	Source of Measure	Data Collection Method
In the Past 12 Months, Whether Physician Has Been Involved in Efforts to:	Draft	Physician Survey
Evaluate:		
How patients of the practice get their needs met during office visits, or by telephone or email (Aspects of getting one's needs met may include time spent waiting in the reception area or exam room, paperwork requirements, encounters with clinical and nonclinical staff members, and receiving notification of test results.)		
How patient information (clinical and billing) is collected and processed		
Change or Improve:		
How patients of the practice get their needs met during office visits, or by telephone or email		
How patient information (clinical and billing) is collected and processed		
Whether the Average Number of Patients Encountered by the Physician Per Day Through (1) Office Visits, (2) Telephone Calls, and (3) Email Messages Has Increased, Decreased, or Stayed About the Same		

Draft = questions we will draft for this survey.

Physician-Beneficiary Interactions. A final set of process-of-care measures that the demonstration may affect pertains to interactions between primary care physicians and beneficiaries. The primary care physician may influence (although not entirely control) beneficiary adherence to recommended therapies and self-monitoring activities. Beneficiaries associated with demonstration practices may be more likely than their counterparts in comparison practices to report that their primary care physician tried to involve them in care planning and educate them about self-monitoring, with a view toward improving adherence. Improved adherence, in turn, might lead toward improvements in the quality indicators tied to bonus payments. Again, we will examine measures of physician-beneficiary interactions because they may help us understand mechanisms underlying effects on the clinical outcomes of primary interest to CMS. Table III.12 presents four measures we will draw from the beneficiary survey.

b. Health Outcomes

Health outcome measures are the *results* of the care beneficiaries receive. These include intermediate-term outcomes (such as improved health-related knowledge and behaviors among beneficiaries), as well as longer-term outcomes (such as fewer hospitalizations of the type that could be avoided if ambulatory care is properly managed and the beneficiary practices adherence and good self-care).

We will ask beneficiary survey respondents about their adherence to several lifestyle behaviors (such as increased physical activity, smoking cessation, and moderation of alcohol intake) that are generally recommended to beneficiaries with chronic illness. We will also ask about some behaviors that are more disease-specific. For example, decreased dietary fat intake may be especially important for beneficiaries who report a diagnosis of coronary artery disease or diabetes. Similarly, control of dietary salt intake would be important for those with

TABLE III.12

PHYSICIAN-BENEFICIARY INTERACTIONS, MEASURED AT THE BENEFICIARY LEVEL

Measure	Source of Measure	Data Collection Method
Whether beneficiary reports setting health goals, and making a plan to meet goals, with PCP	CoCA	Beneficiary Survey
Whether beneficiary reports receiving education or a referral for education on self-care from PCP	CoCA	Beneficiary Survey
Whether beneficiary reports receiving explanation from PCP on what symptoms or problems to look for, and what to do if they appear	Picker	Beneficiary Survey
Whether beneficiary reports times when a health problem could have been avoided through more frequent contact with PCP	Picker	Beneficiary Survey

CoCA = patient survey developed by Mathematica Policy Research, Inc. for the evaluation of the Medicare Coordinated Care Demonstration (Ensor et al. 2003a); PCP = primary care physician; Picker = Picker Ambulatory Care Patient Interview (Lorig et al. 1996).

congestive heart failure. Beneficiaries with diabetes should inspect their feet regularly, and beneficiaries with congestive heart failure should weigh themselves daily. Beneficiaries must also know how to recognize and respond to symptoms of trouble, and what to do if a health condition worsens (Table III.13).

If physicians improve the quality of care they provide to Medicare beneficiaries with the chronic illnesses included in the demonstration, they may be able to help patients avoid the health crises that can lead to hospitalizations. Reducing hospital admissions is important because hospitalizations themselves often cause further declines in function, they are unpleasant for beneficiaries, and they are costly. Some prevention of hospitalizations is essential if P4P programs are to be cost-effective. Beneficiaries associated with demonstration practices may be less likely to need hospitalizations for preventable acute exacerbations or complications of chronic illness. Preventable hospitalizations may be disease-specific or generic. Examples of disease-specific preventable hospitalizations include those for heart failure in beneficiaries with congestive heart failure, or for lower extremity ischemia in beneficiaries with diabetes. Examples of generic preventable hospitalizations include admissions for pneumonia. We will estimate the rate of preventable hospitalizations among at-risk beneficiaries during the demonstration follow-up period. We will use the ICD-9 principal diagnosis codes in Medicare Part A hospital claims (lists of which have been developed by other researchers) to identify risk, restricting the rate calculations for disease-specific preventable hospitalizations to beneficiaries with the disease in question (Table III.13 includes illustrative examples). Assessment of beneficiaries' use of emergency room services, which also should decrease as the quality of ambulatory care improves, is discussed in subsection 5 of this section.

TABLE III.13

HEALTH OUTCOMES, MEASURED AT THE BENEFICIARY LEVEL

Measure	Source of Measure	Data Collection Method
Health-Related Knowledge and Behavior		
Self-Rated Knowledge of:	Picker	Beneficiary Survey
What to be aware of with his/her health condition		
What to do if health condition gets worse		
Whether, on Advice of PCP, Beneficiary Tried to:	Draft	Beneficiary Survey
Increase physical activity		
Stop smoking		
Lower alcohol intake		
Lower salt intake		
Lower intake of dietary fat		
Whether Beneficiaries with Diabetes Examine Feet Daily	DSCA	Beneficiary Survey
Whether Beneficiaries with Congestive Heart Failure Weigh Themselves Daily	CoCA	Beneficiary Survey
Visited PCP within 15 Days of any Hospital Discharge	DM	Medicare Part A and B Claims
Avoidable Hospitalizations		
Among All Beneficiaries, Hospitalized for:	Culler	Medicare Part A Claims
Pneumonia		
Among Beneficiaries with Congestive Heart Failure, Hospitalized for:	Culler	Medicare Part A Claims
Congestive heart failure		
Hypokalemia (potassium deficiency)		
Hyponatremia (water overload)		
Among Beneficiaries with Diabetes, Hospitalized for:	Culler	Medicare Part A Claims
Diabetes out of control or diabetic coma		
Ischemia		
Surgical debridement (removal) of infected tissue		
Lower extremity amputation		
Diabetic foot infection		
Among Beneficiaries with Coronary Artery Disease, Hospitalized for:	Culler	Medicare Part A Claims
Unstable angina, myocardial infarction, cardiogenic shock		
Coronary angiography		
Coronary angioplasty		
Coronary artery bypass surgery		

CoCA = patient survey developed by Mathematica Policy Research, Inc. for the evaluation of the Medicare Coordinated Care Demonstration (Ensor et al. 2003a); Culler = Culler et al. 1998; DM = patient survey developed by Mathematica Policy Research, Inc. (MPR) for the evaluation of the Medicare Disease Management Demonstration (Mathematica Policy Research, Inc. 2005); Draft = questions that we will draft for this survey; DSCA = Diabetes Self-Care Activities (Toobert and Glasgow 1994); PCP = primary care physician.

2. Continuity-of-Care Measures

A common criticism of the U.S. health care system is that care is fragmented. Fragmentation of care occurs when patients lack a "medical home," in which one provider is the patient's usual source of care who can be relied upon for same-day appointments during illness and knows about all the care the patient receives, or when there is a lack of coordination and communication across providers. Fragmentation can reduce overall care quality, particularly in patients with comorbid conditions. The opposite of care fragmentation is care continuity, which we expect to improve as a result of P4P. For Medicare beneficiaries with chronic illnesses, continuity of care is especially important, because these beneficiaries typically require a variety of acute and long-term services, and many prescribed medications. However, measuring continuity of care is difficult, especially in a fee-for-service environment.

a. Access to Care

We will construct both claims- and survey-based measures for whether the demonstration improves beneficiaries' access to a usual source of care. The claims-based measures we plan to construct will reflect the fraction of visits that beneficiaries had with their usual provider (referred to as the "Usual Provider Continuity Index"). For this demonstration, we will define the "usual provider" as the one to whom the beneficiary had been assigned. We will also construct variants of this measure, such as whether most of the beneficiaries' physician visits were with a single provider. To construct survey-based measures, we will ask beneficiaries who is their primary care physician, how long they had been seeing a regular doctor, whether they had a regular place of care, and whether they had a doctor they usually saw. We do not necessarily expect the demonstration to have large effects on whether beneficiaries have a usual care provider (because nearly all Medicare beneficiaries already have one). However, we do expect the demonstration might influence other measures related to access of care that we will ask

about, such as whether the beneficiary was able to schedule appointments and get referrals quickly.

b. Care Coordination

We will measure coordination of care from the perspective of beneficiaries and physicians. We will ask beneficiaries to report instances in which their primary care physician did not speak with other providers involved in his or her care, or did not have information the beneficiary thought he or she should have. We will ask physicians how often beneficiaries received the wrong drug or the wrong dose of a drug, or experienced an adverse drug-drug interaction, any of which could result from different providers not knowing about other aspects of the beneficiary's care. We will also ask physicians how often they receive timely information about results after they refer a beneficiary to a specialist, and how often they received timely information after a beneficiary had been hospitalized or had a change in drugs prescribed by a specialist. (In addition, we will use claims data to determine whether beneficiaries had a primary care physician visit within 15 days of any hospital discharges as a process-of-care measure; Table III.13). Care coordination measures, which we draw from existing surveys, are shown in Table III.14.

3. Satisfaction with Care

Greater satisfaction with care processes, physician-beneficiary interactions, and care coordination on the part of beneficiaries and physicians associated with demonstration practices could be a strong indication that P4P improves care quality. Greater beneficiary satisfaction could also be an important mediating effect for improved health outcomes and for generating health care savings, because beneficiaries who are highly satisfied with their health care

TABLE III.14

COORDINATION-OF-CARE OUTCOMES, MEASURED AT THE INDIVIDUAL LEVEL

Measure	Source of Measure	Data Collection Method
Beneficiary Perspective		
Whether beneficiary reports instances in which PCP did not speak with other providers involved in his or her care, or did not have information beneficiary thought he or she should	Picker	Beneficiary Survey
Physician Perspective		
How often beneficiaries receive wrong drug or wrong dose, or experienced drug-drug interaction	CWF	Physician Survey
How often notified by other providers of new or modified prescriptions	Draft	Physician Survey
After a referral, how often receives timely information about results	CWF	Physician Survey
In past 12 months, how often beneficiaries experienced postdischarge problems because PCP did not receive information in a timely manner	CWF	Physician Survey

CWF = Commonwealth Fund survey of physicians (Audet et al. 2005); PCP = primary care physician; Picker = Picker Ambulatory Care Interview (Lorig et al. 1996).

providers and services may be more motivated to adhere to therapies and recommended self-care.

For physicians, we will develop a set of survey questions that lead physicians to reflect on facets of the components of high-quality care. These include the amount of time they can spend with patients; beneficiary knowledge of, and adherence to, recommended therapies and self-care; and the quality of care beneficiaries receive from other providers. Although physicians clearly have a vested interest in giving themselves high ratings, certain types of survey questions are useful in eliciting objective responses. For example, we will ask physicians whether they are more likely than they were a year ago to have ready access to information about beneficiaries at the time of office visits or other encounters, and whether they are more successful than they were a year ago in encouraging Medicare beneficiaries with chronic illness to adhere to prescribed treatments and self-care. Finally, we will ask physicians how satisfied they are with their compensation from Medicare and other payers in the past 12 months; physician satisfaction with compensation may be key to the viability of P4P programs (Table III.15).

We will ask beneficiaries to rate their satisfaction with the care they received from their primary care physician in the past six months and with the care they received from all providers. Satisfaction in both realms could indicate that care is better coordinated across providers under the demonstration. The remaining measures of beneficiary satisfaction will all pertain to care received from the primary care physician, as the demonstration directly targets only these providers. We will ask beneficiaries to rate their satisfaction with the amount of time their primary care physician spends with them during office visits, with the amount of time he or she devotes to education about self-care and what to do if a health condition worsens, and with how easy it is to contact him or her by telephone or email between office visits. We will ask beneficiaries how satisfied they are with the advice their primary care physician gave them on

TABLE III.15

SATISFACTION OUTCOMES, MEASURED AT THE INDIVIDUAL LEVEL

Measure	Source of Measure	Data Collection Method
Physicians		
How Satisfied with:	Draft	Physician Survey
Overall quality of care received by Medicare beneficiaries with chronic illness		
Beneficiaries' knowledge of their conditions and behavior		
Beneficiaries' adherence to recommended self-care		
Beneficiaries' adherence to recommended therapy		
Amount of time spent with beneficiaries during office visits		
Amount of time spent educating beneficiaries		
Compensation from Medicare in the past 12 months		
Compensation from other payers in the past 12 months		
Compared to a Year Ago:		
How often has ready access to information about beneficiaries' history, conditions, and care plan during office visits and other encounters		
How often succeeds in encouraging beneficiaries to adhere to prescribed treatment and self care		
Beneficiaries		
How Satisfied with:	Draft	Beneficiary Survey
Amount of time spent with PCP during office visits		
Amount of time PCP devotes to education about self-care and what to do if a condition worsens		
PCP's accessibility by telephone or email		
Reminders from PCP to make or keep appointments for medical care		
Advice from PCP on ways to prevent illness and promote health		

TABLE III.15 (*continued*)

Measure	Source of Measure	Data Collection Method
PCP's familiarity with medical history and conditions		
PCP's involvement in overall care		
PCP's knowledge of care received from other providers		
Overall quality of health care in past six months		
Care received from PCP in past six months		

Draft = questions that we will draft for this survey; PCP = primary care physician.

how to prevent illness and promote health. Finally, to further assess care coordination, we will measure beneficiary satisfaction with their primary care physician's (1) familiarity with their medical history and conditions, (2) involvement in overall care, and (3) knowledge of care received from other providers. Table III.15 summarizes beneficiary satisfaction measures.

4. Descriptive Measures

Our evaluation will include a descriptive analysis of physicians' attitudes toward the MCMP demonstration and its features, as well as an examination of whether participation is a burden on their time or has other detrimental effects on the everyday practice of medicine. Data for the descriptive analysis will be drawn from a special module of the physician survey (administered only to physicians associated with demonstration practices). Examined alongside practice-level data on the 26 quality indicators on which bonus payments are based, the descriptive data should provide insights into why scores vary across practices (see Chapter IV for details of this analysis). The descriptive measures will also help CMS identify features of demonstration design that may warrant modification. Table III.16 contains an illustrative list of descriptive measures.

5. Costs and Service Use

Medicare costs and service use are among the most critical outcomes for the evaluation. Analysis of impacts on total Medicare costs for traditional services will indicate whether these savings are large enough to offset the cost of the intervention. Examination of impacts on various services will indicate the sources of such savings. Because hospitalizations represent the largest share of total Medicare costs, we will pay particular attention to estimating program impacts on the number of hospital admissions.

TABLE III.16

PHYSICIAN EXPERIENCES WITH THE MCMP DEMONSTRATION

Measure	Source of Measure	Data Collection Method
Physicians		
How or How Much the MCMP Demonstration Has Affected:	CoCA	Physician Survey
The way you care for Medicare beneficiaries with chronic illness		
Your practice's ability to be more in line with recommended clinical practice guidelines or evidence-based medicine		
Your practice's ability to monitor and follow up with beneficiaries		
The time you spend educating Medicare beneficiaries about self-care and monitoring		
The time you spend communicating with other providers who are treating your Medicare patients with chronic illness		
Your Medicare patients undergoing unnecessary or duplicate tests		
The quality of your relationships with your Medicare patients with chronic illness		
The overall health of your Medicare patients with chronic illness		
Beneficiaries' satisfaction with their health care		
Number of times beneficiaries had office visits in the past six months. Was the increase/decrease medically appropriate?		
Your clinical decision making? In what way?		
Overall, what impact has the MCMP demonstration had on the quality of care of Medicare beneficiaries with chronic illness?		
In what ways has the demonstration been most useful?		
In what ways could it have been more useful?		

TABLE III.16 (*continued*)

Measure	Source of Measure	Data Collection Method
Would you recommend the demonstration to your colleagues?		
Do you have experience with other pay-for-performance programs?		
How does MCMP compare to the other programs with which you have experience?		

CoCA = physician survey prepared by Mathematica Policy Research, Inc. for the evaluation of the Medicare Coordinated Care Demonstration (Ensor et al. 2003b).

The P4P incentives may also affect the use and cost of other services. We would expect that modifications to practices' workflows implemented in response to the financial incentives to improve quality of care, adoption of EHRs, or both, would result in better management of beneficiaries' chronic conditions. However, the use of some services could increase if they replace or prevent the need for hospital care. For example, evidence-based practice guidelines for the target conditions may recommend that beneficiaries receive specific care from physicians, thereby increasing the average number of physician visits and Part B costs. We will estimate the impacts on the use and cost of all major Medicare-covered services (hospital, home health care, skilled nursing facility [SNF], hospice, physician office visits, other physician costs, and emergency room visits) to determine how any overall effects are achieved. The outcome measures relating to service use and costs that the evaluation will examine include:

- The probability of receiving various Medicare services
- The amount of Medicare services received
- The cost to Medicare of those services
- The cost of the incentive payments to demonstration practices
- The net savings to Medicare (to assess whether the demonstration is budget neutral)

We will measure whether beneficiaries received any care (as illustrated on Table III.17), as well as the amount of services used among those receiving each type of service (Table III.18), for each of the following services: home health care, physician care, emergency room care, outpatient services, hospital, hospice, and SNF care among those using care. In addition to measuring impacts on the costs of each type of service (Table III.19), the analysis will estimate the demonstration's effects on Medicare Part A, Part B, Part D (if data become available), and total costs (Table III.20). All costs will be reported per Medicare-covered month, to control for

TABLE III.17

REGRESSION-ADJUSTED EFFECT OF THE DEMONSTRATION ON PERCENTAGE USING MEDICARE SERVICES, STATE A

Service Use Category	Demonstration Practices	Comparison Group Practices	Estimated Effect (p-value)
Percentage Having:			
Any home health care			
Any outpatient care			
Any physician visit			
Any emergency room visit			
Any hospital admission			
Any hospice care			
Any skilled nursing facility care			

Source: Medicare claims data.

Notes: Effects estimated using regression models controlling for predemonstration characteristics of the individual and of the practice. Truncated observations are weighted by the number of months during the follow-up period that individuals were alive and not in Medicare managed care.

TABLE III.18

REGRESSION-ADJUSTED EFFECT OF THE DEMONSTRATION ON AMOUNT OF MEDICARE SERVICES USED AMONG SERVICE USERS, STATE A

Medicare Service Type	Demonstration Practices	Comparison Group Practices	Estimated Effect (p-value)
Among Those Using Service:			
Number of home health care visits			
Number of physician visits			
Number of emergency room visits			
Number of outpatient visits			
Number of hospital admissions			
Number of inpatient hospital days			
Number of days of hospice care			
Number of days of skilled nursing facility care			

Source: Medicare claims data.

Note: Effects estimated using regression models controlling for predemonstration characteristics of the individual and of the practice. Truncated observations are weighted by the number of months during the follow-up period that individuals were alive and not in Medicare managed care.

TABLE III.19

REGRESSION-ADJUSTED EFFECT OF THE DEMONSTRATION ON MEDICARE EXPENDITURES PER MONTH ENROLLED, STATE A

Expenditure Category	Demonstration Practices	Comparison Group Practices	Estimated Effect (p-value)
Expenditures for:			
Inpatient Hospital			
Skilled nursing facility			
Home health			
Hospice			
Physician visit			
Other physician costs			
Emergency room visits			

Source: Medicare claims data.

Note: Effects estimated using regression models controlling for predemonstration characteristics of the individual and of the practice. Expenditures exclude months that beneficiaries were enrolled in Medicare managed care.

TABLE III.20

REGRESSION-ADJUSTED EFFECT OF THE DEMONSTRATION ON MEDICARE EXPENDITURES PER MONTH ENROLLED, STATE A

Expenditure Category	Demonstration Practices	Comparison Group Practices	Estimated Effect (p-value)
Expenditures for:			
Medicare Part A Services			
Medicare Part B Services			
Medicare Part D Services[a]			
All Medicare Services			
Average Incentive Payment per Practice		n.a.	
Average Medicare Savings per Practice (Effect of the demonstration on total Medicare costs per practice minus average incentive payment per practice)			

Source: Medicare claims data and program data on incentive payments.

Note: Effects estimated using regression models controlling for predemonstration characteristics of the individual and of the practice. Expenditures exclude months that beneficiaries were enrolled in Medicare managed care.

[a]If Part D claims data become available.

n.a. = not applicable.

beneficiaries who were not covered by Medicare fee-for-service for the full 12-month follow-up period.

a. Condition-Specific Measures

The evaluation will also test whether the incentives affect service use and reduce costs for services that are suitable for the target chronic conditions (congestive heart failure, coronary artery disease, and diabetes), as well as the other chronic conditions for which the incentive payment for preventive care will be made. For example, for beneficiaries with diabetes, we would examine the use and cost of services for dilated eye examinations and hemoglobin A1c tests (as described in Section E.1). We expect that the incentives will be more likely to influence care related to a specific target condition, although changes in physician practice to meet the quality targets may affect care and outcomes for other beneficiary comorbidities. Thus, our main focus will be on examining all conditions. In addition, the condition-specific estimates may be inaccurate. Which diagnoses are recorded for a particular visit or episode of care is somewhat arbitrary and has been shown to differ substantially across providers. Nonetheless, examination of service use and costs specific to the target conditions may help shed light on the sources of any cost savings. For example, for those with diabetes or coronary artery disease, we will examine the demonstration's effects on whether the participant received smoking cessation counseling.

b. Cost Savings

To assess any cost savings of P4P, the evaluation will measure the demonstration practices' net savings per beneficiary month. While it is possible that the net savings could be negative, we anticipate that net savings will be at lest zero due to the demonstration's budget neutrality requirements. To do this, we will first construct a measure of the costs of the incentives from the

annual payment data to each demonstration practice, to be supplied by the financial support contractor. Based on this measure, we will estimate the program cost per beneficiary month over the previous 12-month follow-up period for each of the three annual periods for which payments will be made. We will compare these costs with the estimated savings to Medicare per beneficiary month (based on our regression results) over the same follow-up period, to estimate the demonstration's net savings per beneficiary per month.

Due to the high variance of Medicare expenditures across beneficiaries, the analysis may find statistically significant reductions in hospitalization rates that are not accompanied by significant reductions in expenditures. In this case, we will construct an alternative measure of expenditures to determine whether savings to Medicare were produced that could not be detected statistically due to the large variance of Medicare expenditures. For example, we will look for the presence of outliers. A single high-cost outlier (such as a kidney transplant case), which could be due to chance alone, could mask savings in a state that actually reduced costs for other beneficiaries. For this reason, we will reestimate impacts with all outliers in the demonstration and comparison groups truncated at a fixed value. For example, we would set the costs for all cases above a given percentile (for example, the 98th percentile) at that percentile value, reestimate the regression models, and compare the results to those from the raw, nontruncated data to assess the sensitivity of the impact estimates to the high-cost cases.

c. Interim Assessment of Cost Savings and Potential Revisions to the Incentives

Using data for the first 18 months of demonstration operations, we will compare the demonstration cost per beneficiary month over the follow-up period to the estimated Medicare savings per beneficiary month (based on regression estimates) over the same follow-up period.[14]

[14] We will conduct the analysis with all four states pooled together.

If the savings in Medicare costs are not enough to offset the cost per beneficiary month of providing the incentive, and it appears that there is no trend toward increased savings, CMS may wish to reduce the incentive to a level that would render the demonstration budget-neutral. We will also explore whether this budget-neutrality analysis is sensitive to outliers, by running our analyses with trimmed outliers (for example, capped at the 98th percentile). If the analysis is sensitive to outliers in either direction, then CMS should consider projecting cost savings, and revising incentive payments, based on the analysis with trimmed outliers. However, any such proposed change should be incorporated at the outset into the operational protocols under which the demonstration will be implemented.

It is possible that the demonstration will yield no savings, or that the savings in Medicare costs is so small that it would be impossible for the demonstration to be budget-neutral over the entire study period even if the incentive were greatly reduced. If this is the case, we will explore whether demonstration savings might be greater in the second half of the demonstration than in the first half. For example, the demonstration could affect short-term clinical outcomes that will not translate to cost savings until later in the study period. Therefore, in addition to estimating the savings over the first half of the demonstration, we will estimate the Medicare projected savings during the last half of the demonstration, under assumptions about how the demonstration's impacts might change. We will provide CMS with the estimated incentive needed to render the demonstration budget-neutral under each of the projected savings scenarios. If reducing the incentive payment will not be sufficient to render the demonstration budget-neutral under reasonably realistic scenarios, CMS will need to consider whether to continue the demonstration.

d. Reconciling Impacts on Various Outcome Measures

To understand whether the demonstration generated cost savings in each state, we will reconcile the estimates of impacts on aggregate and service-specific costs and service use. This interpretative analysis will rely primarily on qualitative analysis. For example, we will array the service impact, cost impact, and cost impact without outliers for each service category for all target conditions and for condition-specific measures. In some states, estimates for all these outcome measures may provide evidence that the intervention reduced Medicare expenditures, or conversely, that the intervention increased Medicare expenditures. When the estimates produce conflicting evidence, we will focus on whether there were statistically significant impacts on service use for the most expensive Medicare-covered services, such as hospitalizations, SNF stays, and home health care. If the cost estimates are not statistically significant but are sizable, we will consider the statistical power to detect an effect of the estimated size, and whether there were outliers. As we reconcile the impact estimates, we will draw on the insights gathered in the implementation analysis to assess the plausibility of alternative estimates.

F. STATISTICAL METHODOLOGY FOR ESTIMATING IMPACTS

This section describes the statistical models that we will use to estimate demonstration impacts and the sensitivity and robustness tests that we will conduct to increase our confidence that the estimates truly reflect demonstration impacts. Throughout this analysis, we will estimate impacts separately for each demonstration state, because physician practice regulations, practice styles, practice settings, technical assistance to implement HIT, adoption of EHRs, and P4P penetration will differ across states. Where sample sizes permit, we will estimate impacts for subgroups defined by practice features such as size or patient mix.

Most of the analysis of claims-based quality and outcome measures will require that we construct control variables based on claims data. Therefore, these analyses will require different

models than those used for survey data (beneficiary and physician), because the control variables will be limited to what is available from claims data. Sample sizes will be much larger for the claims-based analyses due to the expectation that the number of beneficiaries classified as having a primary care physician who is in a demonstration practice will far exceed the beneficiary survey sample sizes, in addition to loss of survey observations to interview nonresponse and item nonresponse.

1. Regression Models

To estimate impacts of P4P on outcomes, we will use hierarchical linear regression models to analyze claims-based outcomes (related to quality, costs, and service use) available for both the predemonstration and demonstration periods. We will use claims-based analyses to assess whether there are likely to be unobserved differences between demonstration and comparison group practices that will bias impact estimates based on analyses that do not include predemonstration values of the outcome measure. Depending on the results of this assessment, we may need to use selection-adjusted linear and probit (or logit) models for cross-sectional survey data and clinical outcomes.

a. Hierarchical Linear Models for Claims-Based Quality, Service Use, and Cost Outcomes Available for the Predemonstration and Demonstration Periods

As noted in Section A, we will use a difference-in-differences approach to estimate impacts for outcomes for which we have claims data (including quality outcomes, costs, and service use) for a baseline period and during the demonstration for both demonstration and comparison group practices. To implement this approach, we will use a hierarchical (or nested) linear model (HLM) framework. Specifically, we will use a two-level HLM model to estimate the results for each state separately: Level 1 corresponds to the *beneficiary* and Level 2 corresponds to the *practice* (the unit of intervention). The regression model for a continuous dependent variable is:

$$(1)\ Y_{ipq} = \alpha_0 + \alpha_1 T_p + \alpha_2 X_{ip} + \alpha_3 Z_p + \sum_{q=2}^{l} \delta_{0q} F_{ipq} + \sum_{q=2}^{l} \delta_{1q} (T_p F_{ipq}) + \left[u_p + \sum_{q=2}^{l} \tau_{pq} F_{ipq} + e_{ipq} \right]$$

where Y_{ipq} is the dependent variable for beneficiary *i* in practice *p* at follow-up point *q* (*q =1,…,l*), where period *q=1* corresponds to the baseline period; F_{ipq} is an indicator variable equal to 1 for observations at the follow-up point *q;* T_p is a treatment status variable indicating whether practice *p* is a demonstration practice*;* u_p are practice-specific random error terms (at baseline) with distribution $N(0, \sigma_p^2)$; τ_{pq} are error terms that represent the extent to which practice effects vary over time during the follow-up period (relative to the baseline period) with distribution $N(0, \sigma_\tau^2)$; e_{ipq} are beneficiary-level residual error terms that are distributed independently of u_p and τ_{pq} with distribution $N(0, \sigma_e^2)$; and the remaining terms are parameters (beneficiary-specific X_{ip} or practice- specific Z_p control variables).

In this formulation, δ_{1q} represents the impact in follow-up period *q*, and is the demonstration-comparison group difference between the mean dependent variable in period *q* relative to the mean baseline dependent variable in period *1* (that is, $\left[Y_{..qT} - Y_{..1T} \right] - \left[Y_{..qC} - Y_{..1C} \right]$). The coefficient α_1 is an estimate of the predemonstration difference between the treatment and comparison practices.

We will estimate equation (1) using *xtmixed* in STATA (StataCorp 2005).[15]

[15] For binary dependent variables, we may have to use *MLwiN* (Center for Multilevel Modelling 2006).

b. Assessing the Need for Selection-Adjusted Models

Data from the beneficiary and physician surveys will be available only once during the demonstration period, precluding us from using the difference-in-differences method for accounting for all measured and unmeasured factors that do not change. Therefore, before estimating the demonstration's effects on survey-based outcomes, we will assess whether there is selection bias due to such unmeasured differences between demonstration practices and comparison group practices. Specifically, for claims-based measures, we will compare the impact estimates based on regressions that used a difference-in-differences approach to the impact estimates based on regressions that do *not* control for the predemonstration value of the outcome measure. (Such regressions would control only for predemonstration variables that will be available for all analyses, such as practice-level characteristics from the Office Systems Survey.)

If both sets of claims-based impact estimates are similar, we will assume impacts based on survey-based measures will not be biased, even though these regressions will not include predemonstration values of the outcome measure. We will then analyze survey-based outcomes using linear regression models (for continuous outcome variables) and logit (or probit) regression models (for binary outcome variables) that also account for the survey design (stratification, clustering of beneficiaries among practices, and sampling weights). These regressions would control for all relevant, available predemonstration measures, such as practice-level characteristics drawn from the Office Systems Survey or from Medicare claims data, the demographic characteristics and diagnoses of the beneficiary (for analyses of the beneficiary survey), and the demographic characteristics, and educational level of the physician (for analyses of the physician survey). However, if the difference-in-differences impact estimates are substantively different from the impact estimates that do not include predemonstration values of

the outcome measure, we will need to implement selection-adjusted regression models, as described below.

c. Selection-Adjusted Linear and Probit Models for Cross-Sectional Survey Data

If needed, we will use selection-adjusted linear and probit models for assessing the impacts of the incentives on measures derived from both the beneficiary and physician surveys. In addition, to properly account for the complex survey design, we will use estimation methods that take into account the sampling weights and other design parameters (for example, stratification and clustering within physician practice).

The challenge is to account for differences between demonstration and comparison group practices due to unobserved characteristics that affect the outcome of interest. For example, we suspect that practices that have planned to adopt an EHR, or actually began to use it, before the demonstration are more likely to provide better quality of care and therefore enroll in the demonstration to receive the incentives for improving care. Thus, comparing outcomes from demonstration and comparison group practices is likely to lead to overestimates of the effectiveness of the incentives.

To deal with this *endogeneity* (or self-selection) of practices into the demonstration, we will use the *two-part model* developed by Maddala (1983). This model requires identification of one or more variables that are likely to predict participation in the demonstration but that are not likely to influence the outcomes of interest. For example, we would need to use baseline practice characteristics (such as size or patient mix), one or more measures of the degree of sophistication with HIT before the beginning of the demonstration (according to the Office Systems Survey), or other characteristics we could measure from physician survey data to estimate the probability of a practice participating in the demonstration (that is, the *first part* of the model). We would then

estimate a regression model (linear or probit, depending on the type of dependent variable) of the outcome of interest, including an indicator for whether a practice is in the demonstration or comparison group (that is, the *second part* of the model).[16] In practice, both equations are jointly estimated (by maximum likelihood), accounting for the possible correlation between their respective error terms. The model would have the following specification:

$$
\begin{aligned}
& Y_{ip} = \beta_0 + \beta_1 X_{ip} + \delta T_{ip} + \varepsilon_{ip} \\
(2) \quad & T_{ip}^{*} = \gamma_o + \gamma_1 Z_{ip} + \mu_{ip} \\
& T_{ip} = I\left(T_{ip}^{*} > 0\right)
\end{aligned}
$$

where Y_{ip} is the dependent variable for beneficiary *i* in practice *p;* T_{ip} is a treatment status indicator variable for practice *p* (and beneficiary *i*)*;* X_{ip} are characteristics that would predict the outcome of interest*;* Z_{ip} are practice characteristics that would predict participation in the demonstration but *not* the outcome of interest; *I(.)* is an indicator-variable function that returns a value of one when the expression inside it takes a positive value, or zero otherwise; and ε_{ip} and μ_{ip} have a bivariate normal distribution with zero mean and covariance matrix $\left[\begin{smallmatrix}\sigma & \rho \\ \rho & 1\end{smallmatrix}\right]$. In this formulation, the coefficient δ is the impact of the incentives on the outcomes of interest. In addition, the term $\lambda = \rho\sigma$ (or *lambda,* as it is known in the econometric literature) is used to test the hypothesis of independence of the two equations. We will estimate equation (2) using *treatreg* in STATA.

Another challenge we face is estimating the impact estimate (δ) and its standard error accounting for the complex survey design. Unfortunately, the two-part model is not supported by standard statistical packages (such as STATA or SUDAAN) that account for the survey

[16] Both beneficiaries and physicians will be nested in a given practice. Therefore, all of them will be assigned to either the demonstration or comparison group because of their assignment to a practice.

design (StataCorp 2005; Research Triangle Institute 2006). Therefore, we would first need to estimate the model to assess whether the two equations are independent, assuming that the survey sample was randomly selected *without* allowing for the complex design (that is, stratification and clustering). If they were independent, then we would use standard linear or probit models to estimate the outcomes equation accounting for the complex survey design. If the equations were not independent, we would use another method (called *instrumental variables*), which is supported by survey data analysis packages (Johnston and DiNardo 1997), to estimate the impacts of the demonstration. In this instance, we would write a technical memorandum discussing the pros and cons of using the instrumental-variables method with survey data for discussion with CMS before we proceed with the analysis.

In sum, the most critical element for estimating the selection-adjusted models for survey data (if needed) is the identification of measures that are good predictors of practice participation in the demonstration, but not outcomes. The likelihood that the proposed measures do not predict participation well may limit our ability to produce robust impact estimates. We will revisit the identification of these variables after we review all available measures from the Office Systems Survey.

d. Practice-Level Regressions

We plan to conduct several descriptive analyses that will rely on impact estimates at the practice level, as described in Chapter IV. We plan to modify the equations described above to generate these practice-level estimates for each state.

2. Testing Strategy

We will use standard procedures and significance levels to test the many hypotheses considered in the evaluation. Most of the tests about the existence of overall demonstration

effects will be two-tailed tests of whether the coefficient of the indicator of whether beneficiaries (or physicians) are enrolled in a demonstration practice is significantly different from zero using a 0.05 significance level. We believe that the incentives most likely will improve quality and reduce costs, but impacts in the opposite direction are possible (Shekelle et al. 2006). For example, as noted earlier, changes to the physicians' workflows to accommodate EHRs may encourage physicians to order additional tests or may reduce the satisfaction of beneficiaries served by these physicians. Because we will be conducting many comparisons of outcomes between demonstration and comparison practices, we will use adjustments to the significance level (for example, the *Bonferroni* adjustment) to minimize the likelihood of finding any spuriously significant impacts. We will group the outcomes according to their substantive area (for example, cost, quality, satisfaction) and will adjust the significance levels based on the relevant number of outcomes in each analysis.

3. Sensitivity Analyses

We will perform tests of the robustness of our estimates, particularly because we will use a quasi-experimental design. For example, as discussed in Section E.5, we will examine the effects of outliers on our impacts estimates and will perform checks for consistency between cost and utilization impact estimates. Furthermore, we also plan to assess the sensitivity of our estimates to different definitions of a demonstration practice. While the definition of a comparison practice will remain unchanged during the demonstration (those physicians assigned to a practice at baseline will continue to be included in the definition of the practice, whether or not they leave the practice), the definition of demonstration practices may change. At the end of each of the three years the demonstration will run, the financial support contractor will identify those physicians who constitute the practice at that time. Thus, the *operational* definition of a practice will be dynamic, because payments to the practice need to be made only for physicians

who agreed to participate in the demonstration. The alternative definitions—one using the practice's tax identification number [TIN] and the other using the baseline definition (thus excluding physicians who join a practice)—would allow us to assess how sensitive our impact estimates are to the definition of a practice adopted for DOQ-IT practices in comparison states and, indirectly, to the completeness of the practice and physician identifier numbers in claims data. Because this analysis may involve considerable resources, we will first discuss the need for it with CMS.

4. Control Variables for Impact Analysis

The set of independent variables used to control for baseline differences between the demonstration and comparison groups will depend on whether we analyze claims-based or survey-based outcomes. In general, control variables will include both individual and practice characteristics. For beneficiaries, individual characteristics typically will include demographics and comorbidities. Most of these factors may influence beneficiaries' Medicare service use and costs and should be controlled for. For physicians, demographic characteristics may influence the way they practice and their readiness for adopting innovations in their work, including EHRs. Finally, practice characteristics will include size, and the degree of sophistication with HIT at baseline, which have been suggested as likely predictors of successful EHR adoption (Miller and Sim 2004). Table III.21 lists the control variables and their sources.

Demographic and socioeconomic characteristics of beneficiaries, including age, sex, race, original reasons for Medicare entitlement, date of death (if applicable), and HMO enrollment, will be extracted from the Medicare EDB; education, income, living arrangements, care-seeking attitudes, and language spoken will be drawn from the beneficiary survey; and diagnoses will be

TABLE III.21

CONTROL VARIABLES AND THEIR SOURCE

Medicare Enrollment Database
Age
Sex
Race
Original reason for Medicare entitlement (age or disability)
HMO enrollment (used to restrict the sample to fee-for-service beneficiaries)
Beneficiary Survey
Education
Income
Living arrangements
Care-seeking attitudes
First language other than English
Physician Survey
Age
Sex
Race
Education
Specialty and board certification
Knowledge of computers before demonstration start
Experience with EHRs or other HIT before demonstration start
Office Systems Survey
Practice size
Availability of HIT
Plans to implement an EHR system
Stage of implementation, if applicable
Length of enrollment in DOQ-IT
Practice affiliation
Scores for degree of sophistication with EHRs at baseline
Languages spoken
Medicare Claims
Number of Medicare beneficiaries served by the practice in the year before demonstration start
Average Medicare expenditures per beneficiary per practice in the year before demonstration start
Percentage of beneficiaries in practice that were hospitalized in the year before demonstration start
Number of E&M visits per beneficiary per practice in the year before demonstration start
Diagnoses (percent in practice with key diagnoses)

EHR = electronic health record; HIT = health information technology.

taken from Medicare claims data.[17] For the analysis of the ***physician*** survey, we also will draw demographic and socioeconomic characteristics, including age, sex, race, education, whether board certified, knowledge of computers before demonstration start, and experience with EHRs or other HIT before demonstration start, from the survey.

We also plan to control for several ***practice characteristics*** in the analyses of the survey data and claims data. From the Office Systems Survey, we will take practice size, availability of HIT, plans to implement EHRs, stage of implementation (when applicable), length of enrollment in DOQ-IT (to measure how long practices have received technical assistance from QIOs), practice affiliation (for example, independent or affiliation with another organization), whether at least one physician speaks languages other than English when seeing patients, and the scores for the degree of sophistication with EHRs at baseline. Finally, claims data will allow us to control for several practice-level characteristics, such as number of beneficiaries served by the practice in the year before demonstration start; average Medicare expenditures per beneficiary per practice during the same period; and number of hospitalizations and E&M visits per beneficiary in the practice in the year before demonstration start. We expect that some of these characteristics (such as practice size) would be predictive of the decision to enroll in the demonstration, but not of outcomes, so that we can minimize the likelihood that our impact estimates would be biased because practices were not randomly assigned to the demonstration or comparison group. However, this is an empirical issue that needs to be examined when we obtain the required data. Finally, as noted above, because the number of control variables from Medicare claims data will be rather limited, we will rely heavily on practice characteristics to control for important differences between practices in assessing the impacts of the demonstration on claims-based outcomes and expect the Office Systems Survey to be a key source of this information.

[17] The diagnoses for the target chronic conditions will be available from the financial support contractor (ARC), who will use these data to calculate the incentive payments (Wilkin et al. 2007).

IV. SYNTHESIS OF IMPLEMENTATION AND IMPACT ANALYSES

A. OVERVIEW OF THE SYNTHESIS

The ultimate goal of the evaluation will be to provide guidance to CMS on whether P4P incentives for improving quality of care and for adopting and using HIT in solo or small- to medium-size group physician practices serving Medicare beneficiaries with chronic illnesses should be implemented on a larger scale and, if so, how this intervention might best be structured. Whether P4P should be implemented depends on whether the demonstration leads to improved quality of care and is at least budget neutral. Structuring of the intervention requires assessing the answers to three questions: (1) For which types of practices were the incentives most effective? (2) How did clinical outcomes vary with the incentives? and (3) How did quality of care, Medicare costs, and the financial incentives vary with HIT use?

To address this goal, we will synthesize our findings for the report to Congress (and for the final evaluation report). In the synthesis, we will pull together our findings from practices in all four states and outcome measures from both the implementation and impact analyses; we will note substantial state-to-state differences as appropriate. We will use this information to draw inferences about the role that financial incentives play in improving care for Medicare beneficiaries with chronic illnesses and on the adoption and use of HIT, and about the most successful ways to implement the incentives (and the technology for performance reporting). The synthesis will entail determining how the intervention's impacts on quality of care and Medicare costs vary with practice characteristics.

We will present the findings from this synthesis in the final report to Congress, which is due in October 2010. We also will include a summary of our synthesis in the evaluation final report, which is due September 2011.[1]

To accomplish the evaluation's basic goals, we will draw on the state-specific implementation and impact analyses to describe physician practices' experiences adopting and using an EHR system, or other HIT for performance reporting, and the care management strategies they use for chronically ill fee-for-service beneficiaries to improve quality of care. Likewise, we will describe how impacts varied with many of the practice characteristics that could potentially influence the efficiency of P4P programs. Our approach to the synthesis will involve three components, all of which feed into the recommendations. In the first component, we will use *exploratory* and *confirmatory* analyses to assess which practice characteristics seem to successfully improve quality outcomes and reduce costs. In the second component, we will assess how quality outcomes vary with the incentives the practices will receive for attaining predetermined performance standards. Finally, in the third component, we will examine the association between quality outcomes and costs and the practice's level of HIT use.

In Section B of this chapter, we describe the framework for organizing the synthesis. In Sections C, D, and E, we describe how we will conduct the component parts of the synthesis. The next chapter discusses how we will report our findings and options for large-scale implementation.

B. FRAMEWORK FOR SYNTHESIZING RESULTS

As a first step in conducting the syntheses, we will report on the number of practices that appear to have met the basic demonstration goal of improving quality of care, reducing

[1]See Chapter V for a detailed discussion of the content of, and schedule for, these reports.

Medicare costs for health care services (by enough to offset the costs of the financial incentives), and encouraging the adoption and use of HIT using P4P. We will use our logic model (Figure I.1), as well as our discussion of the expected effects of the demonstration (Chapter III, Section B), to select the primary outcome measures we will use to decide whether the demonstration reached its goals. First, we will cross-classify the practices (1) by changes in quality outcomes that are directly related to financial incentives (full bonus payment for a given condition, some bonus, or no bonus); and (2) by changes in the effect on the cost of Medicare-covered services (increased, no effect, or reduced by more than enough to offset the incentive payments). Each assessment will require integrating findings from several outcome measures, with possibly conflicting evidence on the size and statistical significance of the effects. For example, a practice's estimated impact on costs may not be statistically significant even as the estimate for quality measures for the target conditions shows significant positive effects. Similarly, estimated impacts on some measures related to care quality and use of Medicare-covered services may be statistically significant, whereas others may not be. Therefore, we will base inferences on the preponderance of the evidence across practices in each dimension. In addition, we will explore constructing a composite measure for the primary outcomes to integrate the many outcome measures into a summary index that could allow us to examine practices along a continuum of specific dimensions.

After this summary of the evidence has been compiled, we will use a unifying framework to synthesize the findings across the practices in each state. The goal of the synthesis will be to identify the wide range of practice characteristics that might be related to P4P effectiveness. For the implementation and impact synthesis, we will focus our discussion on the following questions:

- For which types of practices did the incentives have the largest impacts on quality of care and costs?
- How did quality outcomes vary with the incentives?
- How did quality of care, Medicare costs, and the incentives vary with HIT use?

1. For Which Types of Practices Did the Incentives Have the Largest Impacts on Quality of Care and Costs?

A key component of the synthesis will be our assessment of the practice characteristics that seem to successfully improve quality and reduce costs as the result of P4P incentives and the likely adoption of HIT (most notably, an EHR system) for performance measurement. For example, the recent literature suggests that successful implementation of EHR systems is more the result of effective *organizational changes* in clinical practice than of the technology (Scott et al. 2005). Thus, we will assess how impacts vary with the extent to which practices had organized their workflows before they considered installing an EHR and how they are using the technology. Likewise, we will examine the variability in impacts by practice size, because the evidence suggests that larger practices are more likely than their smaller counterparts to adopt EHR systems (Miller et al. 2004).

2. How Did Quality Outcomes Vary with the Incentives?

Another key issue for the synthesis will be determining whether quality outcomes vary with the performance incentives that demonstration practices will receive. This is particularly relevant because, as the number of P4P programs continues to grow, it remains unclear how the level of payment may influence changes in quality outcomes. Thus, our analysis will be one of the first to examine the role that incentives for achieving quality performance thresholds (that is, *achievement incentives*) may play in the successful implementation of P4P among small

practices serving Medicare beneficiaries with chronic conditions (Rosenthal et al. 2005; Wilkin et al. 2007).

3. How Did Quality of Care, Medicare Costs, and the Incentives Vary with HIT Use?

A related key issue will be assessing whether changes in quality outcomes, the use and costs of Medicare-covered services, and the financial incentives vary with the practices' degree of HIT use early in the demonstration and at the end of it. Recent evidence suggests that only a few organizations have shown improvements in quality and efficiency (Chaudhry et al. 2006).

C. RELATING IMPACTS TO PRACTICE CHARACTERISTICS

A unique feature of the MCMP demonstration is the large number of practices (about 800 across the four states) that will participate in the demonstration. Having this many practices in the demonstration will make it possible to sort out the combination of many of the characteristics that explains why some practices have substantial impacts on the quality of care and costs and others have no (or smaller) impacts.[2] This analysis will be feasible at the state level and, if appropriate, for all four states combined. However, in the latter case, considerable caution would be needed to interpret the findings, because it will be possible to control for only a handful of state characteristics simultaneously due to the likely high correlation among them.

If a substantial number of practices have significant impacts on key outcomes, we will conduct both exploratory and confirmatory assessments of the sources of these differences. The exploratory assessment will be accomplished by distinguishing practices that successfully improve a given outcome from practices that do not, and by comparing the characteristics of the successful and unsuccessful practices. The characteristics we will examine are those used to

[2]For the analyses discussed in this and subsequent sections, we will estimate impacts at the practice level, which is the unit of intervention. This requires adjustments to the models described in Chapter III.

develop our classification of practices, as discussed in Chapter II. The exploratory analysis will therefore determine the extent to which practice success appears to be specific to practices with a particular characteristic. We will also use the exploratory analysis to determine whether practice success seems to be linked to combinations of measured characteristics.

The confirmatory analysis will be accomplished by examining whether the impact estimates across outcomes tend to consistently show that a given characteristic was associated with better outcomes. We will compare impacts for several outcomes, including quality outcomes directly related to financial incentives, total Medicare cost, hospital admissions, and HIT use. We will examine key practice characteristics, such as size and location.

1. Exploratory Analysis

The exploratory analysis will be useful for identifying combinations of characteristics that seem to be associated with positive impacts (assuming that some of the practices have favorable impacts). We will take advantage of the large number of practices that will enroll in each state to conduct this analysis (about 150 demonstration practices per state in Arkansas and Utah and about 250 practices per state in California and Massachusetts).

We will conduct this analysis in two steps. First, we will compare the mean characteristics of successful and unsuccessful practices. We will use several alternative definitions of "successful" practices to ensure that our inferences are robust to the definition used, as it is somewhat arbitrary. For example, we will consider defining practices as successful based on the statistical significance of impacts on some combination of key quality-of-care/cost outcomes. Alternatively, we could include any practice if it received full incentive payments for any chronic condition or if the average monthly Medicare cost for demonstration practices was more than one standard deviation below that of comparison practices. The practice characteristics

described in Chapter II provide an illustrative list of some of the characteristics that we expect to use in these comparisons.

Second, we will use logit or probit regression to assess the effect of a specific characteristic on the likelihood of being a successful practice, controlling for other practice characteristics. The dependent variable will be a binary indicator of whether a practice had a favorable impact on a specific outcome, or combination of outcomes, and the independent variables will be the characteristics described above. This analysis will complement the description of successful and unsuccessful practices by identifying which practice characteristics have the largest influence on being a successful practice. Alternatively, we will use a linear regression to assess the effect of specific characteristics on continuous measure of success, such as the sum of the *effect sizes* across outcomes.[3] An advantage of this specification is that the definition of the dependent variable does not require using an arbitrary threshold for identifying successful or unsuccessful practices. Furthermore, using an effect-size-based measure makes comparisons across practices and, if appropriate, states, much easier.

2. Confirmatory Analysis

The confirmatory analysis will be useful in summarizing for which outcomes the impacts were associated with specific characteristics, such as practice size. We will rely on descriptive methods to examine whether the impact estimates across outcomes tend to consistently show that a given characteristic was associated with better outcomes. For example, we will examine whether practice size consistently showed a positive association with quality outcomes, because

[3]The effect size is defined as the ratio of the impact estimate for a specific outcome divided by the standard deviation for that outcome.

larger practices tend to have more resources than small practices (those with one or two physicians) to change their processes and adopt performance measurement technologies.

D. RELATING QUALITY OUTCOMES TO THE INCENTIVES

The impact analysis will allow us to assess whether the P4P incentives affected specific quality outcomes for each of the three years of the demonstration, or for all three years combined. However, it will not allow us to examine whether the improvements or changes in these indicators are associated with the level of payment the practice receives in a given year. Because not all practices will receive the maximum bonuses for chronic and preventive care, it will be feasible to exploit this variability to examine how the change of clinical quality indicators from one year to the next varies with practice characteristics, especially the incentive payments for previous years. Practice-level characteristics we will consider include (1) whether a practice received a bonus payment in the prior year, (2) practice size, (3) location, and (4) average Medicare payments per beneficiary served by the practice during a given period.

We will use data on clinical quality indicators for the second and third years of the demonstration (so that the incentives for the previous year are available for the analysis) and for both periods combined. As noted, we will conduct this analysis for each state and, if appropriate, pool the data across the four states to maximize the sample size available. We will estimate a linear regression model between the score of a quality indicator in a given period and the incentive payments in the previous year, controlling for the level of the score in the previous year and other practice characteristics (and, for serial correlation, when pooling data for the second and third years).

E. RELATING QUALITY OF CARE, COSTS, AND THE INCENTIVES TO HIT USE

Another promising descriptive analysis we will conduct is to assess how impacts on key outcomes (that is, clinical indicators and Medicare costs) vary with the degree of HIT use the practices had at the beginning of the demonstration. We also plan to examine how changes in the use of EHRs or other performance measurement technologies in the practice may be correlated with impacts on claims-based outcomes since the OSS survey will be available for both demonstration and comparison practices. Finally, we will also examine whether changes in HIT use are associated with the size of the incentives.

We will use data from the Office Systems Survey to conduct this analysis, because this survey will provide measures of the degree of HIT use of each demonstration and comparison practices at the beginning and end of the demonstration (see Chapter II). Our methods will build on those proposed for the analyses described in Sections C and D.

V. REPORTING OF DEMONSTRATION FINDINGS

The demonstration evaluation will produce several reports, including an implementation report, a report on site visits, and interim and final evaluation reports that synthesize findings across states and analytic components. The evaluation reports will be adapted to develop a report to Congress. This chapter describes the purpose, timing, and content of each report. Table V.1 summarizes the schedule for the deliverables.

TABLE V.1

SCHEDULE OF DRAFT REPORT DUE DATES

	Draft Due	
Report	Project Month[a]	Calendar Month
Design Report	n.a.	February 2007
Implementation Report	13	July 2008
First Interim Evaluation Report	16	October 2008
Cost Neutrality Monitoring Report	24	June 2009
Second Interim Evaluation Report	28	October 2009
Report to Congress (Third Interim Evaluation Report)	40	October 2010
Site Visits Report	46	April 2011
Final Evaluation Report	51	September 2011

[a]Refers to months after the start of the demonstration (July 1, 2007).

n.a. = not applicable.

A. IMPLEMENTATION REPORT

The implementation report, due in July 2008 (13 months after the start of the demonstration) will provide an overview of implementation and the results of the wave 1 site visits. The overview will include a summary of demonstration activities to date in each state and summary statistics on the number of practices that enrolled and that submitted baseline data. The results of the wave 1 site visits will be synthesized across the states, with major state-to-state differences noted, and state-level site visit summaries provided as an appendix. As discussed in Section E below, the implementation report will feed into the Report to Congress.

B. SITE VISITS REPORT

The site visits report, due in April 2011 (46 months after the *start* of the demonstration), will provide the results from the second wave of site visits and draw implementation-related conclusions based on both waves of visits. Similar to the implementation report, we plan to synthesize results across the states, noting substantial state-to-state differences as appropriate. State site visit summaries will also be provided as an appendix.

C. COST NEUTRALITY MONITORING REPORT

OMB has requested that we monitor cost neutrality over the first 18 months of the demonstration. This analysis will require comparing our regression estimates of the demonstration's effects on Medicare savings to the incentive payments made to demonstration practices. Assuming we will receive the data for this analysis by month 21 (that is, 21 months after the demonstration begins), we plan to deliver a draft of this report to CMS in month 24 after the demonstration begins (that is, June 2009). This task will be particularly challenging because, as noted in Chapter III, it will difficult to assess whether there is a trend toward increasing savings, given that we will only have 18 months of data.

D. INTERIM AND FINAL EVALUATION REPORTS

One of the most important components of the evaluation will be the synthesis of the findings from the implementation and impacts analyses to determine whether the P4P incentives improved quality of care for fee-for-service Medicare beneficiaries with chronic illnesses and influenced the adoption and use of HIT and, therefore, whether P4P should be implemented on a larger scale.

We will conduct three interim evaluation reports (drafts due 16, 28, and 40 months after the start of the demonstration, respectively) and a final evaluation report (draft due 51 months after the start of the demonstration), all of which will synthesize those findings available at different times during the demonstration.

1. First Interim Evaluation Report

The first interim evaluation report, due in October 2008 (16 months after the start of the demonstration), will provide qualitative descriptions of practice changes made in response to the intervention, including changes to the processes associated with the adoption of HIT and how it is used. It will rely only on data from the first round of site visits and the Office Systems Survey, as data on claims, clinical measures, and financial incentive payments for the first year of operations will not be available until May 2009.

2. Second Interim Evaluation Report

The second interim evaluation report, due in October 2009 (28 months after the start of the demonstration), will focus on impact estimates for the first year of program operations. Although we will compare impacts on use of Medicare-covered services and costs across practices and states, we will not attempt to draw inferences from them at this stage of the evaluation. In addition, we will summarize findings from our telephone discussions with highly

successful practices and those that withdrew, if any, in year 2 of demonstration operations. This report will draw heavily on the monitoring report described in Section C of this chapter.

3. Third Interim Evaluation Report

The third interim evaluation report, due in October 2010 (40 months after the start of the demonstration), will focus on impact estimates for the second year of program operations. We also will include findings on the impacts of P4P on physician-beneficiary interactions (that is, access to care, care coordination, and satisfaction with care) from the beneficiary survey. Finally, we will summarize findings from the second wave of site visits to the practices we visited during the first year of operations, as well as telephone discussions with highly successful and unsuccessful practices (including those that withdrew, if any) in year 3 of demonstration operations. As discussed in Section E below, the Report to Congress will be the third interim report.

4. Final Evaluation Report

The final evaluation report, due in September 2011 (51 months after the start of the demonstration), will provide final impact estimates from claims data using data from the third, and final, year of demonstration operations. In addition, we will present impact estimates from the physician survey on processes associated with the adoption of HIT to improve quality of care. The report will also include our synthesis analysis, using the approaches described in Chapter IV, including data from the last wave of the Office Systems Survey and the implementation synthesis (site visits) report.

E. REPORT TO CONGRESS

We will produce one report to Congress based on our evaluation. The draft report will be due in October 2010, approximately 3 months after the *end* of the demonstration operations.

This report will analyze implementation experiences and findings of the MCMP demonstration across the four states. Because this report is due before the final evaluation report (see above), the third interim report will be submitted as the Report to Congress. This will pose a challenge because we will need to present conclusions and lessons learned from the demonstration without seeing the impact estimates for the final year of demonstration operations, given that the data for this period will not be available until May 2011. In coordination with CMS, we will start planning for the report to Congress shortly after we submit the final version of the second interim evaluation report to ensure that the focus of the report to Congress addresses the key evaluation questions with the findings available up to that point. We will write the concise report for an audience of high-level policymakers and decision makers who may not be familiar with the demonstration project or evaluation methodologies.

REFERENCES

Audet, Anne-Marie J., Michelle M. Doty, Jamil Shamasdin, and Stephen C. Schoenbaum. "Physicians' Views on Quality of Care: Findings from the Commonwealth Fund National Survey of Physicians and Quality of Care." New York: Commonwealth Fund, May 2005.

Bodenheimer, Thomas, Jessica H. May, Robert A. Berenson, and Jennifer Coughlan. "Can Money Buy Quality? Physician Response to Pay for Performance." Issue Brief no. 102. Washington, DC: Center for Studying Health System Change, December 2005.

Chaudhry, Basit, Jerome Wang, Shinyi Wu, Margaret Maglione, Walter Mojica, Elizabeth Roth, Sally C. Morton, and Paul G. Shekelle. "Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care." *Annals of Internal Medicine*, vol. 144, no. 10, May 16, 2006, pp. E12–E22 and W1–W18.

Center for Multilevel Modelling. *MLwiN, Version 2.01*. [www.mlwin.com/features/index.html]. Accessed April 6, 2006.

Culler, Steven D., Michael L. Parchman, and Michael Przybylski. "Factors Related to Potentially Preventable Hospitalizations Among the Elderly." *Medical Care*, vol. 36, no. 6, June 1998, pp. 804–817.

de Brantes, Franҫois. "Lessons Learned." Presentation on the *Bridges to Excellence* Program at the Institute of Medicine, Washington, DC, May 4, 2005.

Ensor, Todd, Arnold Chen, and Randall Brown. "Medicare Coordinated Care Demonstration Patient Survey Questionnaire." Princeton, NJ: Mathematica Policy Research, Inc., July 2003a.

Ensor, Todd, Arnold Chen, and Randall Brown. "Medicare Coordinated Care Demonstration Physician Survey Questionnaire." Princeton, NJ: Mathematica Policy Research, Inc., February 2003b.

Hassol, A., H. Harrison, R. Jarmon, B. Rodríguez, and A. Frakt. "Survey Completion Rates and Resource Use at Each Step of a Dillman-Style Multi-Modal Survey." Paper presented at the 58th annual conference of the American Association for Public Opinion Research, Nashville, TN, May 15-18, 2003.

Iglehart, John K. "Linking Compensation to Quality—Medicare Payments to Physicians." *New England Journal of Medicine*, vol. 353, no. 9, September 1, 2005, pp. 870–872.

Johnston, J., and J. DiNardo. *Econometric Methods, 4th edition.* New York: McGraw Hill, 1997.

Link, Michael, Ali Mokdad, Machelle Town, David Roe, and Suzanne Triplette. "Use of Lead Letters and Answering Machine Messages." Paper presented at the 58th annual conference of the American Association for Public Opinion Research, Nashville, TN, May 15-18, 2003.

Lorig, Kate, Anita Stewart, Philip Ritter, Virginia Gonzalez, Diana Laurent, and John Lynch. *Outcome Measures for Health Education and Other Health Care Interventions.* Thousand Oaks, CA: Sage Publications, 1996.

Maddala, G.S. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press, 1983.

Mathematica Policy Research, Inc. "Medicare Disease Management Program Evaluation: Patient Questionnaire." Princeton, NJ: MPR, July 2005.

Miller, Robert H., and Ida Sim. "Physicians' Use of Electronic Medical Records: Barriers and Solutions." *Health Affairs*, vol. 23, no. 2, March/April 2004, pp. 116–126.

Miller, Robert H., John H. Hillman, and Ruth S. Given. "Physician Use of IT: Results from the Deloitte Research Survey." *Journal of Healthcare Information Management*, vol. 18, no. 1, 2004, pp. 72–80.

Research Triangle Institute. *SUDAAN: Release 9*. Research Triangle Park, NC: RTI, 2006.

Rosenthal, Meredith, B., Richard G. Frank, Zhonghe Li, and Arnold M. Espstein. "Early Experience with Pay-for-Performance. From Concept to Practice." *Journal of the American Medical Association*, vol. 294, no. 14, 2005, pp. 1788–1793.

Scott, J. Tim, Thomas G. Rundall, Thomas M. Vogt, and John Hsu. "Kaiser Permanente's Experience Implementing an Electronic Medical Record: A Qualitative Study." *British Medical Journal*, vol. 221, 2005, pp. 1313–1316.

Shekelle, P.G., S.C. Morton, and E.B. Keeler. *Costs and Benefits of Health Information Technology*. Evidence Report/Technology Assessment no. 132. AHRQ Publication no. 06-E006. Rockville, MD: Agency for Healthcare Research and Quality, April 2006.

StataCorp. *Stata Statistical Software: Release 9*. College Station, TX: StataCorp, LP, 2005.

Toobert, D.J., and R.E. Glasgow. "Assessing Diabetes Self-Management: The Summary of Diabetes Self-Care Activities Questionnaire." In *Handbook of Psychology and Diabetes: A Guide to Psychological Measurement in Diabetes Research and Practice*, edited by C. Bradley. Newark, NJ: Harwood Academic Publishers, 1994.

Wilkin, John C., C. William Wrightson, David Knutson, Erika G. Yoshino, Anahita S. Taylor, and Kerry E. Moroz. "Medicare Care Management Performance Demonstration. Design Report." Columbia, MD: Actuarial Research Corporation, January 5, 2007.

APPENDIX A

ENABLING LEGISLATION FOR THE DEMONSTRATION AND THE EVALUATION

MEDICARE PRESCRIPTION DRUG, IMPROVEMENT, AND MODERNIZATION ACT OF 2003

TITLE VI—PROVISIONS RELATING TO PART B
Subtitle D—Additional Demonstrations, Studies, and Other Provisions

SEC. 649. MEDICARE CARE MANAGEMENT PERFORMANCE DEMONSTRATION

(a) ESTABLISHMENT.

(1) IN GENERAL.—The Secretary shall establish a pay-for-performance demonstration program with physicians to meet the needs of eligible beneficiaries through the adoption and use of health information technology and evidence-based outcomes measures for

(A) promoting continuity of care;
(B) helping stabilize medical conditions;
(C) preventing or minimizing acute exacerbations of chronic conditions; and
(D) reducing adverse health outcomes, such as adverse drug interactions related to polypharmacy.

(2) SITES.—The Secretary shall designate no more than 4 sites at which to conduct the demonstration program under this section, of which

(A) 2 shall be in an urban area;
(B) 1 shall be in a rural area; and
(C) 1 shall be in a State with a medical school with a Department of Geriatrics that manages rural outreach sites and is capable of managing patients with multiple chronic conditions, one of which is dementia.

(3) DURATION.—The Secretary shall conduct the demonstration program under this section for a 3-year period.

(4) CONSULTATION.—In carrying out the demonstration program under this section, the Secretary shall consult with private sector and non-profit groups that are under taking similar efforts to improve quality and reduce avoidable hospitalizations for chronically ill patients.

(b) PARTICIPATION.

(1) IN GENERAL.—A physician who provides care for minimum number of eligible beneficiaries (as specified by the Secretary) may participate in the demonstration program under this section if such physician agrees, to phase in over the course of the 3-year demonstration period and with the assistance provided under subsection (d)(2)

(A) the use of health information technology to manage the clinical care of eligible beneficiaries consistent with paragraph (3); and
(B) the electronic reporting of clinical quality and outcomes measures in accordance with requirements established by the Secretary under the demonstration program.

(2) SPECIAL RULE.—In the case of the sites referred to in subparagraphs (B) and (C) of subsection (a)(2), a physician who provides care for a minimum number of beneficiaries with two or more chronic conditions, including dementia (as specified by the Secretary), may participate in the program under this section if such physician agrees to the requirements in subparagraphs (A) and (B) of paragraph (1).

(3) PRACTICE STANDARDS.—Each physician participating in the demonstration program under this section must demonstrate the ability

(A) to assess each eligible beneficiary for conditions other than chronic conditions, such as impaired cognitive ability and co-morbidities, for the purposes of developing care management requirements;
(B) to serve as the primary contact of eligible beneficiaries in accessing items and services for which payment may be made under the medicare program;
(C) to establish and maintain health care information system for such beneficiaries;
(D) to promote continuity of care across providers and settings;
(E) to use evidence-based guidelines and meet such clinical quality and outcome measures as the Secretary shall require;
(F) to promote self-care through the provision of patient education and support for patients or, where appropriate, family caregivers;
(G) when appropriate, to refer such beneficiaries to community service organizations; and
(H) to meet such other complex care management requirements as the Secretary may specify.

The guidelines and measures required under subparagraph (E) shall be designed to take into account beneficiaries with multiple chronic conditions.

(c) PAYMENT METHODOLOGY.—Under the demonstration program under this section the Secretary shall pay a per beneficiary amount to each participating physician who meets or exceeds specific performance standards established by the Secretary with respect to the clinical quality and outcome measures reported under subsection (b)(1)(B). Such amount may vary based on different levels of performance or improvement.

(d) ADMINISTRATION

(1) USE OF QUALITY IMPROVEMENT ORGANIZATIONS.—The Secretary shall contract with quality improvement organizations or such other entities as the Secretary deems appropriate to enroll physicians and evaluate their performance under the demonstration program under this section.

(2) TECHNICAL ASSISTANCE.—The Secretary shall require in such contracts that the contractor be responsible for technical assistance and education as needed to physicians enrolled in the demonstration program under this section for the purpose of aiding their adoption of health information technology, meeting practice standards, and implementing required clinical and outcomes measures.

(e) FUNDING.

(1) IN GENERAL.—The Secretary shall provide for the transfer from the Federal Supplementary Medical Insurance Trust Fund established under section 1841 of the Social Security Act (42 U.S.C. 1395t) of such funds as are necessary for the costs of carrying out the demonstration program under this section.

(2) BUDGET NEUTRALITY.—In conducting the demonstration program under this section, the Secretary shall ensure that the aggregate payments made by the Secretary do not exceed the amount which the Secretary estimates would have been paid if the demonstration program under this section was not implemented.

(f) WAIVER AUTHORITY.—The Secretary may waive such requirements of titles XI and XVIII of the Social Security Act (42 U.S.C. 1301 et seq.; 1395 et seq.) as may be necessary for the purpose of carrying out the demonstration program under this section.

(g) REPORT.—Not later than 12 months after the date of completion of the demonstration program under this section, the Secretary shall submit to Congress a report on such program, together with recommendations for such legislation and administrative action as the Secretary determines to be appropriate.

(h) DEFINITIONS.—In this section:

(1) ELIGIBLE BENEFICIARY.—The term ‘‘eligible beneficiary’’ means any individual who—

(A) is entitled to benefits under part A and enrolled for benefits under part B of title XVIII of the Social Security Act and is not enrolled in a plan under part C of such title; and
(B) has one or more chronic medical conditions specified by the Secretary (one of which may be cognitive impairment).

(2) HEALTH INFORMATION TECHNOLOGY.—The term ‘‘health information technology’’ means email communication, clinical alerts and reminders, and other information technology that meets such functionality, interoperability, and other standards as prescribed by the Secretary.

APPENDIX B

DOQ-IT OFFICE SYSTEMS SURVEY

Office Systems Survey

June 5, 2006

QIO Assigned Practice ID Number: _____ Date: _____

Thank you for volunteering to participate in the Centers for Medicare & Medicaid Services (CMS) Office Systems Survey (OSS). The goal of this CMS Doctors Office Quality Information Technology (DOQ-IT) initiative is to unite technology and clinical practice in the physician office setting. This is a unique opportunity for your practice to contribute to a large-scale national effort to improve the quality of ambulatory health care. The survey asks about three types of electronic clinical information tools/functions that you may be using in your practice to help manage your patient's health needs. These tools allow for the systematic application of evidence based medical guidelines to your patient population with a goal of developing care plans for any given patient.

In the survey you will be asked if you are currently using or are in the process of obtaining a:

- Electronic Health Record (EHR)
- Electronic registry software
- Electronic prescribing software

Throughout the survey we will ask you to provide information about the **functions** of the systems you currently have in place. The goal is to use this information to help CMS develop additional programs that can assist physicians move toward the common goal of improving care.

Please complete all sections of the survey unless directed within it to skip a section.

Again, we thank you for your participation and look forward to continuing to work with you.

DOQ-IT
Doctors' Office Quality - Information Technology

Office Systems Survey

SECTION 1 - General Information - Practice

1. Please review your practice information below for accuracy. Please make corrections where necessary:

1.1. Legal Name of Practice _____

1.2. Address: _____

1.3. City: _____ 1.4 State _____ 1.5. Zip Code: _____

1.6. Telephone No.: _____

1.7. Fax No.: _____

1.8. E-mail Address: _____

1.9. Practice (Group) Medicare Billing Number (PIN): _____

(If unknown, please check with your billing manager or HCFA 1500 Form - field 33)

1.10. Federal Tax ID for this practice: _____

Please check here if all of the above information is correct. ☐

1.11. Is your practice affiliated with an Independent Practice Association (IPA), Physician Hospital Organization (PHO) or medical group?

☐ No ☐ Yes - please indicate which IPA, PHO or medical group: _____

1.12. Preferred Method of Contact: ☐ Telephone ☐ Fax ☐ E-mail (check all that apply)

DOQ-IT Doctors' Office Quality - Information Technology **Office Systems Survey**

SECTION 2 – Provider Profile

Your Quality Improvement Organization (QIO) provided the following information. Please review the information below for accuracy and make corrections/additions where necessary. Please note that physician identifiers are being requested in this survey to ensure that the correct information corresponds with the correct physician practice. The information you provide will be used by CMS internally, for the purposes of this project. This information will not be shared or disseminated outside of the project staff.

First Name	MI	Last Name
UPIN[1]		(NPI) National Provider Identification Number (If known)
Credentials (MD, DO)	Specialty[2]	Language(s) spoken (Other than English)
Primary Practice Location (Y/N)[3] Yes No	PIN # (Individual Medicare Billing Number)[4]	

First Name	MI	Last Name
UPIN[1]		(NPI) National Provider Identification Number (If known)
Credentials (MD, DO)	Specialty[2]	Language(s) spoken (Other than English)
Primary Practice Location (Y/N)[3] Yes No	PIN # (Individual Medicare Billing Number)[4]	

First Name	MI	Last Name
UPIN[1]		(NPI) National Provider Identification Number (If known)
Credentials (MD, DO)	Specialty[2]	Language(s) spoken (Other than English)
Primary Practice Location (Y/N)[3] Yes No	PIN # (Individual Medicare Billing Number)[4]	

Footnotes:

1 Unique Physician Identification number, a six place alphanumeric identifier assigned to each physician/practitioner

2 Please use the following codes to indicate specialty: Cardiology (C); Endocrinology (E); Family Practice (F); Geriatrics (G); Internal Medicine (I); Other (please specify)

3 Please indicate whether the provider listed primarily practices at this office location (50% or greater = practices primarily at this site).

4 Please provide the Individual Medicare Billing Number (PIN) that is assigned by the Medicare Carrier in your state for use by this physician/clinician at this practice site only. (HCFA 1500 form field 24K or 33).

SECTION 3 – QIO Experience

The purpose of this section of the survey is to learn about your experience working with your local Quality Improvement Organization (QIO).

3.1 How **satisfied** is your practice with the QIO work in the following areas:

QIO Assistance	N/A	Very satisfied	Somewhat satisfied	Neutral	Somewhat dissatisfied	Very dissatisfied
a. **Timeliness** of the QIO's response to questions or requests for assistance						
b. The **professionalism, courtesy and respectfulness** of the QIO staff						
c. The **ease of access** to the QIO staff (when you try to contact them)						
d. Thinking about all interactions with the QIO, how satisfied are you with their services?						

3.2 Please indicate your level of agreement with the following statement about the **value** of the services your practice received from the QIO:

QIO Assistance	N/A	Strongly agree	Agree	Neither agree or disagree	Disagree	Strongly disagree
a. The **assistance** we received from the QIO was worth the time or effort required on the part of our staff.						
b. We could not have gotten where we are in the adoption and use of health information technology (EHR or e-prescribing and registry) without the QIO's help						
c. We could not have gotten where we are in care management process improvement without the QIO's help						

3.3 Did you **know** about any of the following before today, not know this before today, or weren't sure about this?

QI activity	Knew this before today	Did not know this before today	Not sure
a. CMS is currently testing pay for performance or incentive programs as a means to improve quality.			
b. The QIO also works with nursing homes, hospitals, and home health agencies in quality improvement projects.			

3.4 Using a scale of 10 to 0, please rate the **contribution** of the QIO to your EHR efforts:

	10	9	8	7	6	5	4	3	2	1	0	

Where 10 = "The QIO contribution was indispensable"												Where 0= "The QIO did not contribute at all"

3.5 How **satisfied** is your practice with the assistance provided by the QIO in the following areas:

QIO Assistance	N/A	Very satisfied	Somewhat satisfied	Neutral	Somewhat dissatisfied	Very dissatisfied
a. Assessing your practice's technology needs						
b. Providing information on technology options						
c. Helping with vendor selection processes						
d. Preparing for EHR implementation						
e. Helping to improve quality of care in your practice						
f. Helping to improve practice efficiency (e.g. workflow analysis and redesign, etc.)						
g. Overall assistance with adoption of EHR in your practice						

3.6 When did you first begin actively working with the QIO in the planning and implementation of EHR in your practice?

_____
(MM/DD/YY)

3.7 The following is a list of activities related to EHR adoption. How much had already been completed on each activity when your practice first began working with the QIO, or on August 1, 2005, whichever came later:

Activities	Not started 0	Some activity completed 1	About half way to completion 2	Nearing completion 3	Completed 4
a. Perform office readiness assessment					
b. Document and analyze current office workflows					
c. Redesign office flow to meet EHR process					
d. Evaluate care management and process improvement pre-EHR.					
e. Full implementation of EHR					
f. Use EHR to identify additional care management and process improvement opportunities					

3.8 What has been completed on each activity to date (as of the date of the survey)?

Activities	Not started 0	Some activity completed 1	About half way to completion 2	Nearing completion 3	Completed 4
a. Perform office readiness assessment					
b. Document and analyze current office workflows					
c. Redesign office flow to meet EHR process					
d. Evaluate care management and process improvement pre EHR					

e. Full implementation of EHR					
f. Use EHR to identify additional care management and process improvement opportunities					

SECTION 4 – Office Practice

The implementation of information technology (IT) presents many operational challenges. As the transition from paper to computer takes place, there are opportunities to redesign existing workflows to gain maximum efficiencies. These questions focus on current workflow processes.

* This series of questions refers to patient visits to ANY and ALL clinicians in your practice **over the past month**.

Please estimate the proportion of patient encounters/visits for which clinicians **or others** in your practice engage in each of the following activities.

Clinicians or others in your practice:	**None 0**	**About ¼ 1**	**About ½ 2**	**About ¾ 3**	**All or nearly all 4**
4.1 - Pull **paper** charts for scheduled patient visits					
4.2 - Dictate visit notes into a tape recorder or phone.					
4.3 - Dictate visit notes directly into the EHR					
4.4 - Use a computerized (as opposed to paper) system to manage the following office workflows:					
a. Telephone calls					
b. Prescription refills					
c. Referrals					
d. Results follow-up (lab, diagnostic test, x-ray)					

SECTION 5 - Electronic Health Record

The Electronic Health Record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. This record may include patient demographics, diagnoses, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. The EHR has the capability of generating a complete record of a clinical patient encounter, as well as supporting other care-related activities, such as evidence-based decision support, quality management, and outcomes reporting. *(The EHR covers all conditions that the patient might have and is distinct from a registry that covers a specific disease or a limited set of diseases).* Implementation of the EHR may vary based on the goals set by a practice and the intended functions such as: enter progress notes; provide decision support within the patient encounter; and utilize computerized physician order entry for laboratory and prescriptions.

This section asks about the use/planned use of an EHR in your practice.

* This series of questions refers to patient visits to ANY and ALL clinicians in your practice **over the past month**.

5.1 Does your practice have an Electronic Health Record (EHR)?

☐ **Yes**

When was the vendor contract signed? _____(mm/dd/yy)

When was the system installed? _____(mm/dd/yy)

What is the name and version of the EHR system you use? _____

Are you currently using the system?

☐ **Yes Please proceed to question 5.2.**

☐ **No Please proceed to Section 6 – Patient Registry/Care Management Processes**

☐ **No**

If no, when do you plan to implement an EHR? ☐ Within 1 year ☐ 1-2 years ☐ 3-4 years ☐ Not known at this time

Please proceed to Section 6 – Patient Registry/Care Management Processes

Please estimate the proportion of patient visits/encounters for which clinicians or others in your practice use the EHR to perform each of the following tasks.

Clinicians in your practice use the EHR to:	**None 0**	**About ¼ 1**	**About ½ 2**	**About ¾ 3**	**All or nearly all 4**
5.2 - Generate laboratory requisitions/orders electronically					
5.3 - Enter/retrieve laboratory test results electronically					
5.4 - Generate radiology requisitions/orders electronically					
5.5 - Enter/retrieve radiology results electronically					
5.6 - Enter data into documentation templates					
5.7 Review and act on reminders for care activities (e.g. overdue health maintenance)					

Clinicians in your practice use the EHR to:	None 0	About ¼ 1	About ½ 2	About ¾ 3	All or nearly all 4
5.8 - Maintain medication lists for individual patients					
5.9 - Maintain allergy list					
5.10 - Maintain problem and/or diagnosis list					
5.11- Trend lab and/or other test results over time					

5.12 Does your EHR include ALL or essentially all patients in your practice?

☐ **Yes** ☐ **No**

SECTION 6 – Patient Registry/Care Management Processes

For purposes of this survey, a registry is defined as an electronic system that is designed to identify patients with specific diagnoses or medications; identify patients overdue for specific therapies; prompt ordering of specific laboratory tests or recommended drugs, and prompt communication with patients requiring follow-up. For example, a practice may use a diabetes registry to document care at visits, and to create reports that indicate which patients are due for certain blood tests, or are not meeting specific treatment goals for diabetes. A registry may also be used to ensure all suggested preventive screenings take place. A Registry is usually a stand-alone system that tracks specific information regarding a limited number of disease states, but otherwise lacks additional functionality. An EHR can also be used for Patient Registry/Tracking purposes. **If your practice uses either an EHR, or a Registry, answer as appropriate the questions in this section.**

These next questions ask about the existence and use of electronic registries in your practice.

* This series of questions refers to patient visits to ANY and ALL clinicians in your practice **over the past month**.

6.1 Does your practice have or use a freestanding e-registry to track patients who have a specific chronic illness, or receive preventive care for at least one condition? **Note - if your practice uses an EHR for this purpose, please be certain that question 5.1 was completed and begin with question 6.2.**

☐ Yes	
	When was the e-registry contract signed? _____(mm/dd/yy)
	When was the e-registry system installed? _____(mm/dd/yy) What is the name of the e-registry system? _____
	Are you currently using the e-registry system?
	☐ Yes **Please proceed to question 6.2**
	☐ No **Please proceed to Section 7**
☐ No	
	If No when do you plan to start a registry? ☐ Within 1 year ☐ 1-2 years ☐ 3-4 years ☐ not known at this time
	Please proceed to Section 7

*** Preventive care is defined as immunizations, mammography and other cancer screening.**

6.2 Which of the following conditions are included in your practice's registry/EHR:

Diabetes	☐ Yes	☐ No	Adult Asthma	☐ Yes	☐ No
Coronary Artery Disease	☐ Yes	☐ No	Depression	☐ Yes	☐ No
Hypertension	☐ Yes	☐ No	Anticoagulation	☐ Yes	☐ No
Congestive Heart Failure	☐ Yes	☐ No			
*Preventive Care	☐ Yes	☐ No			
Other	☐ Yes	☐ No			

If Others, please list: _____

Following is a list of tasks that may be performed by registries. For each task, please estimate the proportion of patients or patient encounters for which clinicians **or others** in your practice use each type of registry.

	Types of Disease/Condition Registries 0= none 1= about ¼ 2= about ½ 3= about ¾ 4= all or nearly all				
Registry Tasks	**Preventive Care**	**Diabetes**	**Coronary Artery Disease**	**Congestive Heart Failure**	**Hypertension**
6.3 - Prompt your practice to notify patients who are overdue for office visits	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.4 - Prompt clinicians to order tests, studies, and other services (e.g., immunizations)	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.5 - Produce reminders for <u>patients</u> about needed tests, studies, and other services immunizations)	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.6 - Generate a list of eligible patients for each disease/condition	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.7 - Generate a list of patients requiring intervention	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.8 - Generate a specific patient care plan.	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.9 - Generate written or electronic information to help patients understand their condition	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.10 Create written action plans (personalized to patient's condition) to help guide patients in self-management at home/school/work.	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.11- Prompt clinician and/or patient to review self-management plan together during a visit.	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4

6.12- Modify self management plan as needed following a patient visit	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.13 Generate laboratory requisitions/orders electronically	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
6.14 Enter/retrieve laboratory test results electronically	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4

SECTION 7 - Electronic Prescribing

With electronic prescribing tools, clinicians can generate prescriptions electronically using either a freestanding product, or as a component of the EHR. The next series of questions ask to what extent your practice uses an electronic prescribing tool and whether that tool is stand-alone, or part of your EHR.

* This series of questions refers to patient visits to ANY and ALL clinicians in your practice **over the past month**.

7.1 Does your practice use electronic software to generate the following types of prescriptions (as part of an EHR or a freestanding e-prescribing):

☐ Yes	
	☐ New prescriptions only ☐ Refills ☐ Both
	Is e prescribing accomplished within your EHR?
	☐ Yes **Please skip to question 7.2**
	☐ No What is the name and version of the e-prescribing system you use?
	_____
	When was the contract signed? _____(mm/dd/yy)
	When was the system installed? _____(mm/dd/yy)
	Please skip to Question 7.2
☐ No	
	When do you plan to implement e-prescribing? ☐Within 1 year ☐1-2 years ☐ 3-4 years
	☐ Not known at this time
	Please skip to Section 8

Please estimate the proportion of patient visits/encounters for which clinicians or others in your practice use an electronic or hand-held device for each of the following e-prescribing activities.

0= none 1= about ¼ 2= about ½ 3= about ¾ 4= all or nearly all

E-prescribing activities:	None 0	About ¼ 1	About ½ 2	About ¾ 3	All or nearly all 4
7.2- Identify generic or less expensive brand alternatives at the time of prescription entry					
7.3 - Reference the drug formularies of the patient's health plans/pharmacy benefit manager to recommend preferred drugs at time of prescribing					
7.4- Offer guidelines and evidence-based recommendations when prescribing medication for a patient					
7.5- Calculate appropriate dose and frequency based on patient parameters such as age and weight					
7.6 - Maintain a list of each patient's current medications					

E-prescribing activities (continued):	None 0	About ¼ 1	About ½ 2	About ¾ 3	All or nearly all 4
7.7 - Screen prescriptions for drug allergies against the patient's allergy information					
7.8 - Screen new prescriptions for drug-drug interactions against the patient's list of current medications					
7.9 - Select individual medication for prescription					
7.10 - Print prescriptions on a computer printer					
7.11 - Transmit prescriptions directly to pharmacy via electronic fax (no paper printed)					
7.12 - Transmit prescriptions directly to pharmacy via electronic means (without relying on a fax machine at either clinician's office or in the pharmacy)					
7.13 - Provide patient-friendly information about the medication to the patient					

SECTION 8 - Data Attestation

I have reviewed the data submitted in this survey and agree that it is a correct assessment of this practice.

☐ **Agree** ☐ **Disagree**

Name: _____

Signature: _____

Title: _____

SECTION 9 - Attestation

I understand that I may be chosen to participate in an on-site validation of this survey.

☐ Agree ☐ Disagree.

Comments

This material was prepared by MassPRO, the Medicare Quality Improvement Organization for Massachusetts, under contract with the Centers for Medicare & Medicaid Services (CMS), an agency of the U.S. Department of Health and Human Services. The contents presented do not necessarily represent CMS policy. 8sow-ma-OSS-06-01 survey-jan-5pilot

APPENDIX C

COMPARISON STATE SELECTION PROCESS

MATHEMATICA
Policy Research, Inc.

MEMORANDUM

P.O. Box 2393
Princeton, NJ 08543-2393
Telephone (609) 799-3535
Fax (609) 799-0005
www.mathematica-mpr.com

TO: Lorraine Johnson

FROM: Lorenzo Moreno and Judy Ng

DATE: 7/8/2005
MCMP-032 (Revised)

SUBJECT: Proposed Comparison States

This memorandum describes the process that we used to select potential comparison states for the evaluation of the Medicare Care Management Performance (MCMP) demonstration. We developed this process based on the assumption that Centers for Medicare & Medicaid Services (CMS) wants us to use a quasi-experimental design in the evaluation of MCMP in each of the four states participating in the demonstration, namely Arkansas, California, Massachusetts, and Utah. Specifically, the evaluation will rely on practices in the Doctor's Office Quality—Information Technology (DOQ-IT) in *non-demonstration states*.[1] Still unresolved, however, is how to select physician practices in non-demonstration states.

This memorandum is a revised version of a draft document we sent to you on June 7, 2005. In revising this document, we included the comments we received from CMS during a telephone conference on July 5, 2005.

A. RATIONALE

Our approach to the selection of comparison states relied on selection criteria discussed with CMS staff, as well as on information provided by representatives of a health plan in Utah. As noted below, these criteria aim at identifying states with similar environments than the demonstration states in that they at least had electronic health records and pay-for-performance programs. The selection process also was designed to be reproducible and open to inspection by stakeholders at CMS and in the demonstration states. The comparison states were then selected from a list of 18 that were most likely to reasonably compare to the demonstration states in terms of high-priority characteristics. These states are listed in Table 1:

---

[1]See the description of Comparison Design 3 in memorandum MCMP-027 (Revised), page 1, dated February 16, 2005.

TABLE 1

LIST OF 18 STATES IDENTIFIED AS POTENTIAL COMPARISON STATES

Arizona	New York
Colorado	New Hampshire
Connecticut	Oklahoma
Idaho	**Oregon**
Kansas	Tennessee
Louisiana	Texas
Maine	Vermont
Missouri	Washington
Nebraska	
Nevada	

Note: Entries in bold correspond to the recommendations for Utah made by Intermountain Health Care (IHC) representatives in February 2005. Entries in italics correspond to the recommendations made by CMS in July 2005.

B. SELECTION CRITERIA

The selection of comparison states was based on four high-priority criteria (see entries marked with an A in Table 2), as well as three other criteria for which we need to obtain information from CMS.[2]

TABLE 2

CHARACTERISTICS FOR SELECTING COMPARISON STATES, BY TYPE AND PRIORITY RATING

Characteristic	Priority Rating
Physician Practice	
1. Small practices (3 to 9 physicians) as a percent of group practices (3 or more physicians)	A
2. Ratio of specialists to general practice/family medicine physicians	A
3. Percentage of office-based physicians using electronic health records	B[1]
State	
1. Medicare physician and other professional services expenditures per beneficiary	A
2. Medicare managed care penetration rate	A
3. Geographic representation	B+
4. Number of health plans that have implemented pay-for-performance programs	B
5. Whether the state has a *Bridges to Excellence* program in operation	B

[2]Table A.1 lists the selected characteristics of the 18 states identified in Table 1 for selection as comparison states.

Characteristic	Priority Rating
6. Whether the state has a participant in the Medicare Benefits Improvement and Protection Act (BIPA) Disease Management Demonstration	B
7. Whether the state has a participant in the Chronic Care Improvement Program (CCIP)	B
8. Whether the state has a participant in the Care Management for High Cost Beneficiaries (HCB) demonstration	B
9. Number and amount of AHRQ grants for implementing electronic health records	C
Other	
1. Whether QIOs would participate in MCMP in the comparison state	A[2]
2. Number of practices in non-demonstration states interested in electronic health records	A[2]
3. Input from the QIO in the demonstration state	A[2]
4. Start date of DOQ-IT program in non-demonstration state	C

[1] Only regional or nationwide estimates are currently available (Burt and Hing 2005).
[2] No data are currently available for these characteristics. Thus, we excluded them from the selection criteria.

AHRQ= Agency for Healthcare Research and Quality; DOQ-IT = Doctor's Office Quality—Information Technology; QIO = Quality Improvement Organization.

We do not currently have information on whether the Quality Improvement Organizations (QIO) in the non-demonstration states are willing to participate in the demonstration, the number of practices in non-demonstration states interested in implementing electronic health records, and the reactions to our selection from the QIO in demonstration states. Therefore, these three characteristics were not used to select the proposed comparison states. Should any of these three criteria make our proposed comparison states unacceptable, however, we will select an alternate state matched as closely as possible in physician-practices and state characteristics.

C. COMPARISON STATE SELECTION PROCESS

The process for selecting comparison states used a hierarchical, non-probability (reproducible) method based on the priority rating of the characteristics shown in Table 2. With this method, we arrived at our preliminary list of comparison states in five steps. These steps are described below and summarized in Table 3.

Step 1. Stratification of States. Each demonstration state was viewed as a separate stratum. This stratification accounts for differences across states on the adoption of health information technology, economic and regulatory environments, demographic features, physician licensing and board certification, and practice patterns. The stratification also accounts for differences in how the QIOs are implementing DOQ-IT in the four demonstration states.

Step 2. Identification of States with Closest Priority Characteristics to the Demonstration State. We listed potential comparison states based on the four highest-priority characteristics that are available: (1) small physician practices (3 to 9 physicians) as a percent of all group practices (3 or more physicians); (2) ratio of specialists to general practice/family medicine

physicians; (3) Medicare physician and other professional services expenditures per beneficiary; and (4) Medicare managed care penetration rate. For each characteristic, we rank-ordered the potential comparison states in terms of their similarity to the demonstration state.

Step 3. Selection of States Identified with Highest-Priority Characteristics. We identified potential comparison states based on their similarity to the demonstration states in their highest-priority characteristics (see Table 3, column [5]). For Arkansas, we selected *Missouri* or *Nebraska,* because these states tied on their sum of draws. For California, we selected *Arizona* (for comparison to Southern California only) and *Oregon* (for comparison to California overall). For Massachusetts, we selected *Connecticut* or *New York,* because they, too, tied on their sum of draws. For Utah, we selected *Idaho*.

Step 4. Selection of States to Ensure Similar Geographic Representation. We selected potential comparison states that were in the same geographic area as the demonstration states. Most states selected in Step 3 are in the same census subregion as the demonstration state, with the exception of Arizona, Missouri, Nebraska, and New York.[3] However, Missouri and Nebraska are in the same census region as Arkansas – that is, the Midwest; Arizona is in the same census region as California – that is, the West; and New York is in the same census region as Massachusetts— that is, the Northeast. Thus, keeping Arizona, Missouri, Nebraska, and New York as potential comparison states do not affect substantively the geographic representation of the comparison states.

Step 5. Selection of States that Have Adopted Pay-for-Performance Programs, Participate in Bridges to Excellence or in Three CMS Demonstrations. The last step required is the identification of states that implemented pay-for-performance programs, at least one *Bridges to Excellence* program, or participate in the Medicare Benefits Improvement and Protection Act (BIPA) Disease Management demonstration, the Chronic Care Improvement Program (CCIP), or the Care Management for High Cost Beneficiaries (HCB) demonstration.[4] As noted above, these five criteria ensure that (1) the comparison states have environments similar to those of demonstration states with regard to electronic health records and pay-for-performance programs

[3] Missouri and Nebraska are in the West North Central subregion, whereas Arkansas is in the West South Central subregion; Arizona – which is to be compared to Southern California only – is in the Mountain subregion, whereas California is in the Pacific subregion; New York is in the Mid-Atlantic subregion, whereas Massachusetts is in the New England subregion (Table A.1)

[4] We also examined the amount and scope of funding in FY 2005 by the Agency for Healthcare Research and Quality (AHRQ) (See columns [12] and [12a] in Table A.1). However, it is difficult to establish similarities between demonstration and potential comparison states, because of considerable variability in the grantees, the amount of funding, and the scope of the grants.

and (2) exclude CMS demonstrations that target similar populations as the MCMP demonstration:

- For *Arkansas*, which neither has a pay-for-performance program nor participates in CCIP or HCB, *Nebraska* is preferred over Missouri, despite both having tied with the highest scores in Step 3, since the former also neither has a pay-for-performance program nor participates in CCIP, BIPA, or HCB. Since Texas resembled Arkansas more than either Nebraska or Missouri on the census subregion and pay-for-performance criteria (neither has a program), we considered it as an alternate comparison state.[5]

- For *California*, which has the largest number of pay-for-performance programs among the 22 states examined and participates in *Bridges to Excellence*, BIPA, and HCB, *Arizona* is preferred over Oregon or Washington as a comparison state to *Southern California only* because the former scored the highest in Step 3. As a comparison state to California *overall*, *Oregon* is preferred over Washington because it scored higher on Step 3, despite Washington resembling California more on the pay-for-performance criterion (both have programs).

- For Massachusetts, which has five pay-for-performance programs and participates in *Bridges to Excellence* and HCB, *New York* is preferred, because it is the closest match in terms of pay-for-performance programs.[6] But because New York is not in the same subregion as Massachusetts, and participates in CCIP and HCB, we considered *Connecticut*, which is in the New England subregion and participates in none of the CMS demonstrations, an alternate comparison state.

- For Utah, which has one pay-for performance program and participates in *Bridges to Excellence*, *Idaho* is preferred over the other four states because it scored highest in Step 3 and participates in none of the CMS demonstrations, although it differs from Utah with regard to the presence of pay-for-performance programs and participation in *Bridges of Excellence*.

---

[5]As noted in Table 3, and following CMS's advice, we will exclude the Houston and San Antonio metropolitan regions from the potential comparison areas in Texas.

[6]As noted in Table 3, we will exclude New York City (except Manhattan Borough) and Suffolk and Nassau Counties in Long Island from the potential comparison areas in New York.

D. SUMMARY

Based on the selection process described above, we propose that the following states be used as comparison states for the MCMP demonstration states:

- ***Arkansas:*** Nebraska, with Texas as alternate
- ***California:*** For comparison to Southern California only, Arizona; for comparison to California overall, Oregon, with Washington as alternate
- ***Massachusetts:*** New York, with Connecticut as alternate
- ***Utah:*** Idaho

This list of states appears to be face valid and to meet the criteria shown in Table 2. However, given the numerous dimensions in which states differ, that are not included in the selection criteria, the proposed comparison states are expected only to reasonably match the demonstration states.

We look forward to receiving your comments on our proposal and to discuss alternatives, if needed.

REFERENCES

Burt, Catharine, and Esther Hing. "Use of Computerized Clinical Support Systems in Medical Settings: United States, 2001—03." *Advance Data for Vital and Health Statistics*, no. 353, March 15, 2005. Available online at [www.cdc.gov/nchs/data/ad/ad353.pdf]. Accessed on June 3, 2005.

cc: Sheldon Retchin (VCU), J. Milliner-Waddell, S. Felt-Lisk, L. Foster, A. Bloomenthal, A. Zambrowski, File

TABLE 3
PROCESS OF SELECTION OF COMPARISON STATES

State	Selected Based on Practice Size (1)	Selected Based on Ratio of Specialists to GP/FM Physicians (2)	Selected Based on Medicare Physician Expenditures per Beneficiary (3)	Selected Based on Managed Care Penetration Rate (4)	Sum of Draws (5)	Census Region (6)	Census Subregion (6a)	Adoption of P for P (7)	Participates in B to E (8)	Participates in CCIP (9)	Participates in BIPA (10)	Participates in HCB (11)	State Selected (12)
Arkansas						**Midwest**	**WSC**		**Y**	**N**	**Y**	**N**	
Kansas	0	0	0	1	1	Midwest	WNC		Y	N	N	Y	
Louisiana	0	0	0	0	0	Midwest	WSC		Y	N	Y	N	
Missouri	1	0	1	0	2	Midwest	WNC		Y	N	N	Y	
Nebraska	0	1	0	1	2	Midwest	WNC		Y	N	N	N	**Nebraska**
Oklahoma	0	0	1	0	1	Midwest	WSC		Y	Y	N	N	
Tennessee	1	0	0	0	1	Northeast	ESC		Y	Y	N	N	
Texas	1	0	0	0	1	Midwest	WSC	MS	Y	N	Y	Y	***Texas****
California						**West**	**PA**		**Y**	**N**	**N**	**Y**	
Arizona*	1	1	1	1	4	West	MT		Y	N	Y	N	**Arizona**
Oregon	1	0	0	1	2	West	PA		Y	N	N	Y	**Oregon**
Washington	1	0	0	0	1	West	PA	MS	Y	N	N	Y	***Washington***
Massachusetts						**Northeast**	**NE**		**Y**	**N**	**N**	**Y**	
Connecticut	0	1	1	0	2	Northeast	NE		Y	N	N	N	***Connecticut***
Maine	1	0	0	0	1	Northeast	NE		Y	N	N	N	
New Hampshire	0	0	0	0	0	Northeast	NE		Y	N	N	N	
New York	0	0	1	1	2	Northeast	MA	MS	Y	Y**	N	Y	**New York****
Vermont	0	0	0	0	0	Northeast	NE		Y	N	N	N	
Utah						**West**	**MT**		**Y**	**N**	**N**	**N**	
Arizona	0	0	0	0	0	West	MT	MS	Y	N	Y	N	
Colorado	0	1	0	0	1	West	MT	MS	Y	N	N	N	
Idaho	1	0	1	1	3	West	MT		N	N	N	N	**Idaho**
Nevada	0	0	0	0	0	West	MT		Y	N	N	N	
Oregon	1	1	0	0	2	West	PA		Y	N	N	Y	

Sources: Table A.1

Notes: NE = New England; ESC = East South Central; MA = Mid-Atlantic; WNC = West North Central; WSC = West South Central; MT = Mountain; PA = Pacific
B to E = Bridges to Excellence; BIPA = Benefits Improvement and Protection Act Disease Management Demonstration; CCIP = Chronic Care Improvement Program; FM = family medicine; GP = general practice; HCB = High-Cost Beneficiaries Demonstration; PGP = Physician Group Practice Demonstration; P for P = Pay for Performance

Y = yes; N = no; MS = most similar

*For comparison to *Southern* California only.
**Only in New York City.
***Excluding Houston and San Antonio.
****Excluding New York City (except Manhattan Borough) and Suffolk and Nassau Counties in Long Island.

Shaded cells represent states selected at a specific stage. Bold-italic entries in column (10) correspond to alternate states.

TABLE A.1
SELECTED CHARACTERISTICS OF THE 15 STATES IDENTIFIED FOR SELECTION OF COMPARISON STATES

State	Small Practices (3 to 9 Physicians) as a Percent of Group Practices (3 or more Physicians) (1)	Ratio of Specialists to GP/FM Physicians (2)	Non-Federal PCPs as a Percent of Total Physicians (3)	Medicare Expenditures per Beneficiary (4)	Medicare Managed Care Penetration Rate (5)	Census Region (6)	Census Subregion (6a)	Number of State P-for-P Programs* (7)	Participates in B to E (8)	Participates in CCIP (9)	Participates in BIPA (10)	Participates in HCB (11)	AHRQ HIT Funding: Number of Grants ($ in millions) (12)	EHR (12a)
Arkansas	**78%**	**3.43**	**43%**	**$1,222**	**0%**	**Midwest**	**WSC**	**0**	**Y**	**N**	**N**	**N**	**1 ($1.5)**	**N**
Kansas	72%	4.14	43%	$1,381	2%	Midwest	WNC	0	Y	N	N	Y	none	N.A.
Louisiana	81%	7.41	38%	$1,666	7%	Midwest	WSC	1	Y	N	Y	N	4 ($1.7)	Y
Missouri	77%	8.42	39%	$1,226	6%	Midwest	WNC	1	Y	N	N	Y	1 ($1.5)	Y
Nebraska	83%	3.45	43%	$1,017	2%	Midwest	WNC	0	Y	N	N	N	2 ($0.4)	Y
Oklahoma	70%	4.41	43%	$1,231	5%	Midwest	WSC	1	Y	Y	N	N	3 ($1.3)	N
Tennessee	76%	6.54	40%	$1,243	7%	Northeast	ESC	1	Y	Y	N	N	4 ($6.8)	Y
Texas	76%	6.05	40%	$1,588	4%	Midwest	WSC	0	Y	N	Y	Y	2 ($3.0)	N
California	**77%**	**6.99**	**40%**	**$1,927**	**16%**	**West**	**PA**	**17**	**Y**	**N**	**Y**	**Y**	**6 ($6.2)**	**Y**
Arizona**	78%	6.52	40%	$1,744	10%	West	MT	1	Y	N	Y	N	none	N.A.
Oregon	72%	5.82	42%	$1,340	21%	West	PA	0	Y	N	N	Y	5 ($4.8)	N
Washington	73%	4.36	42%	$1,345	9%	West	PA	1	Y	N	N	Y	3 ($2.5)	N
Massachusetts	**75%**	**19.13**	**35%**	**$1,590**	**13%**	**Northeast**	**NE**	**5**	**Y**	**N**	**N**	**Y**	**10 ($13.4)**	**Y**
Connecticut	79%	18.25	37%	$1,732	3%	Northeast	NE	1	Y	N	N	N	2 ($2.7)	Y
Maine	76%	4.60	44%	$1,020	0%	Northeast	NE	0	Y	N	N	N	4 ($3.2)	N
New Hampshire	83%	5.88	42%	$1,150	0%	Northeast	NE	1	Y	N	N	N	1 ($0.2)	N
New York	81%	16.38	39%	$1,775	12%	Northeast	MA	11	Y	Y+	N	Y	4 ($4.5)	Y
Vermont	80%	5.57	42%	$852	0%	Northeast	NE	1	Y	N	N	N	2 ($1.7)	Y
Utah	**73%**	**5.86**	**39%**	**$1,007**	**3%**	**West**	**MT**	**1**	**Y**	**N**	**N**	**N**	**3 ($8.0)**	**N**
Arizona	78%	6.52	40%	$1,744	10%	West	MT	1	Y	N	Y	N	none	N.A.
Colorado	79%	5.36	41%	$1,476	13%	West	MT	1	Y	N	N	N	1 ($5.0)	Y
Idaho	74%	3.18	44%	$989	7%	West	MT	0	N	N	N	N	2 ($1.1)	Y
Nevada	67%	6.61	42%	$1,731	13%	West	MT	2	Y	N	N	N	none	N.A.
Oregon	72%	5.82	42%	$1,340	21%	West	PA	0	Y	N	N	Y	5 ($4.8)	N

Sources:

(1) American Medical Association. *Medical Group Practices in the US, 2005 Edition.*
(2) American Medical Association. *Physician Characteristics and Distribution in the US, 2005 Edition.*
(3) Kaiser Family Foundation State Health Facts. *Non-Federal PCPs as a Percent of Total Physicians.* Available online at [www.statehealthfacts.org]. Accessed June 3, 2005.
(4) CMS, Office of the Actuary, National Health Statistics Group. *State Estimates - 1998 Medicare Expenditures per Enrollee Physician & Other Professional Services* (most recent year available).
(5) CMS. *March 2005 Medicare Managed Care Quarterly State/County Penetration Rates.* (state estimates = arithmetic average across all state-county rates for that state).
(6 and 6a): Bureau of the Census. *Census Regions and Divisions of the United States*. Available online at [www.census.gov/geo/www/us_regdiv.pdf]. Accessed on June 1, 2005.
(7) Med-Vantage, Inc. *Case Studies in Health Plan Pay-for-Performance Programs* (edited by James Gutman), November 2004.
(8) *Bridges to Excellence. Program Participants*. Available online at [www.bridgestoexcellence.org/bte]. Accessed June 3, 2005.
(9) CMS list of sites participating in the *Voluntary Chronic Care Improvement Program (CCIP,*, July 5, 2005.
(10) CMS list of sites participating in the *Physician Group Practice (PGP) Demonstration*, July 5, 2005.
(11) CMS list of sites participating in the *High-Cost Beneficiaries (HCB) Demonsration*, July 5, 2005.
(12) and (12a) Agency for Healthcare Research and Quality. *Health Information and Technology Programs*. Availabe online at [www.ahrq.gov/research/hitfact.htm]. Accessed on November 11, 2004.

NE = New England; ESC = East South Central; MA = Mid-Atlantic; WNC = West North Central; WSC = West South Central; MT = Mountain; PA = Pacific

Notes: B to E = Bridges to Excellence; BIPA = Benefits Improvement and Protection Act Disease Management Demonstration; CCIP = Chronic Care Improvement Program; FM = family medicine; GP = general practice;
HCB = High-Cost Beneficiaries Demonstration; PCP = primary care physician; P for P = Pay for Performance
CMS = Centers for Medicare & Medicaid Services; AHRQ = Agency for Healthcare Research and Quality
HIT = health information technology; EHR = electronic health records

Y = yes; N = no; N.A. = not applicable

*This table does not include P-for-P programs operating at a national or regional level; it is possible that national/regional programs also have operations in the states listed above.
**For comparison to *Southern* California only.
+ Only in New York City.