Contract No.: 500-95-0047(09)
MPR Reference No.: 8756-110

MATHEMATICA
Policy Research, Inc.

Research Design for the Evaluation of the Medicare Coordinated Care Demonstration

February 13, 2001

Randall S. Brown
Sherry Aliotta
Nancy Archibald
Arnold Chen
Deborah Peikes
Jennifer Schore

Submitted to:

Health Care Financing Administration
Office of Strategic Planning
C3-20-17
7500 Security Boulevard
Baltimore, MD 21244

Project Officer:
Barbara Silverman

Submitted by:

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

Project Director:
Randall Brown

CONTENTS

Chapter **Page**

Chapter **Page**

CONTENTS *(continued)*

Chapter **Page**

EXECUTIVE SUMMARY

Many researchers, policymakers, and medical care providers have observed that a lack of appropriate management of care for chronically ill Medicare beneficiaries and inadequate coordination of their health care leads to poorer outcomes and higher costs for these patients. These care management problems, combined with poor communication among the multiple providers often seen by patients such as these, frequently lead to conflicting or inappropriate prescriptions on diet, medication, exercise, or self-care. This situation, which is exacerbated if communication between provider and patient is poor, can confuse patients, who consequently may fail to adhere to recommended behavior. As a result, patients may experience potentially avoidable adverse outcomes that require the use of expensive services.

To address these problems, the Health Care Financing Administration (HCFA) is conducting a demonstration of coordinated care programs for beneficiaries with chronic illnesses who are covered by the traditional fee-for-service Medicare program. The demonstration, mandated by the Balanced Budget Act of 1997, is to include at least nine demonstration sites. A Request for Proposals (RFP) was issued in July 2000, and 15 of the 58 programs submitting proposals were selected in January 2001 to receive awards. The programs differ widely on target population, interventions, sample sizes, experimental designs, sponsoring organization, and many other characteristics. Mathematica Policy Research, Inc. was awarded a five-year contract in September 2000 to conduct an independent evaluation of the demonstration programs (assuming there would be only nine) and two disease management demonstration programs operated by Lovelace Health Systems.

This report describes the basic research design for the evaluation. Each site will be evaluated separately, and the findings will be summarized and compared in three synthesis reports. Each of the first two synthesis reports, in turn, will form the basis for a Report to Congress. Here, we describe the data sources, samples, implementation analyses, statistical models, outcome measures, methods for synthesizing findings, and the timeline and work schedule for the evaluation.

Readers should bear in mind that we will adapt this design for the site-specific analyses to take into account differences across programs in the experimental design, target population, intervention goals, sample sizes, available data, recruiting and intake procedures, and timing. For example, one site proposes a comparison group design, whereas the others propose some form of randomization of patients. The programs propose sample sizes ranging from the minimum (309 each for treatment and control groups) to 5,500 in each group. We will describe the details of the required site-specific adaptations in site-specific analysis plans, which will be prepared after HCFA, the program operators, and we (the evaluator) agree on each program's basic design features. That process will take place in February and March of 2001, based on telephone discussions and assessments of each program's research design by MPR and HCFA.

Throughout the following discussion, we discuss the evaluation of 17 demonstration sites, although MPR's current contract is to evaluate only 11 sites—9 coordinated care demonstration programs and the 2 Lovelace Health Systems disease management demonstrations (1 for diabetes and 1 for congestive heart failure [CHF]). The discrepancy arises because the evaluation contract was issued prior to selection of the demonstration sites, when it was presumed that only nine coordinated care demonstration programs would be selected. After the number of funded programs that are able to comply with the terms and conditions of participation has been determined, HCFA will decide how to modify the evaluation contract.

THE KEY EVALUATION GOAL IS TO ASSESS IMPACTS ON QUALITY AND COST.

The primary goal of the evaluation is to determine whether care coordination programs can decrease cost without lowering quality, improve quality of care without increasing net costs, or improve quality *and* lower net cost. To try to ascertain why some programs failed and others succeeded, and how successful programs might be replicated, we will have to understand in detail how each program was implemented. Thus, the quantitative and qualitative components of the evaluation will center on addressing the following questions:

- What interventions were delivered, and how did they affect the quality and quantity of patient care?
- Who was targeted to receive program services, and for what types of patients did these services work best?
- What types of organizations provided coordinated care, and how did impacts vary with these organizational features?
- What did the programs cost, and how should an ongoing program be financed?

Assuming that at least some of the programs exhibit favorable effects, the biggest challenge for the evaluation will be attempting to determine what program characteristics work best and for what target populations. The difficulty lies in the fact that there are many program features and combinations of features that could conceivably influence program effectiveness, but only 17 programs from which to draw these comparisons (assuming all 17 are actually implemented). We will use the organizing framework of the four questions presented above to describe and classify programs, and we will draw on both the implementation analysis and the impact analysis to make inferences about associations between program characteristics and program effects.

THE IMPLEMENTATION ANALYSIS WILL DESCRIBE THE INTERVENTIONS.

A critical component of the evaluation will be to describe in detail how the programs were designed and how they were implemented, and the reasons for any differences between these two stages. The descriptions will include a thorough explanation of the interventions, the target populations served, structural characteristics of the organizations implementing the program, and the costs of the programs. The description of the interventions will cover how the programs attempt to improve (1) patient self-care, (2) physician performance, (3) communication and coordination, and (4) service arrangement, and how well these interventions are designed and implemented. We will describe and evaluate how the programs assess patients' needs, how they develop care plans, the services they provide or arrange for, how they monitor these services, and how they reassess and update care plans. Descriptions of the target populations will include explanations of the eligibility criteria, the rationale for these criteria, the recruitment and intake processes, and the programs' level of success in enrolling and retaining the targeted population. The discussion of the programs' structural characteristics will cover the types of organizations sponsoring the programs, the number of staff hired as case managers, the training of these staff, and the degree of integration between case managers and health care providers. The program costs to be described will include start-up costs, monthly fees, costs related to the length of time patients are retained in the program, and payments to providers for their participation. We will also compare costs incurred by the program with HCFA's payments for program services.

In addition to describing how the programs were planned and implemented, this analysis will lead to a system for classifying care coordination programs on a number of dimensions. This classification framework will help clarify how policymakers, providers, and researchers think about care coordination and will be essential for synthesizing our findings across programs.

Data for the implementation analysis will come primarily from four sources: (1) telephone and in-person contacts with demonstration programs and the implementation contractor; (2) program documents (such as proposals, operational protocols, marketing materials, and staff training materials); (3) program records (such as patient-level enrollment and disenrollment records and program cost reports); and (4) Medicare data (to compare participants with eligible nonparticipants).

Findings from site-specific implementation analyses will be presented in three sets of site-specific reports: the Case Study, and the First and Second Site-Specific Interim Evaluations. Each report will focus on evolution of a program at a different point in time. The Case Study, due six months after the start of patient enrollment, will focus on describing the history and design of the program and its early implementation experiences. It will be based on telephone contacts with program and implementation contractor staff and reviews of program documents. The First Site-Specific Interim Evaluation, due six months after that, will provide a detailed description of program features as implemented, problems encountered, and changes made in response to those problems. This interim evaluation will also provide an overview of the health service environment in which the program was implemented. The report will be based on day-long in-person discussions with program staff and on analyses of enrollment and cost data. The Second Site-Specific Interim Evaluation, which will be based on telephone contacts with program staff, will focus on changes made to the program through the end of the demonstration, features that appear to be associated with program success or failure, and lessons for future care coordination/disease management programs.

THE IMPACT ANALYSES WILL ESTIMATE PROGRAM EFFECTS ON COSTS AND QUALITY OF CARE.

The primary objective of the impact analysis is to assess whether the demonstration programs were able to achieve the goals of improving patient well-being and reducing costs. This assessment will require a rigorous experimental design, examination of a large number of outcome measures from several sources, sufficient sample sizes, strong statistical models that generate unbiased estimates of program impacts, and formal tests of hypotheses about these impacts. In addition, we will conduct various tests of the sensitivity of the estimates to the model, outliers, nonresponse bias, contamination, and other potential threats to the validity of the estimates. Each program will be evaluated separately.

The research design calls for randomization, if possible, and minimum samples of 309 beneficiaries per group.

The basic approach to estimating impacts will be to use regression models to compare outcomes for the treatment and control groups in each site. Only 1 of the 17 programs to be evaluated does not propose to conduct random assignment of eligible patients into a treatment or control group, and that site has indicated a willingness to consider random assignment. However, it is possible that our assessment of the programs' research designs will indicate that a randomized design has the potential to produce unacceptable amounts of contamination in some sites, thus necessitating a comparison site design. For each program site using random assignment, we also plan to draw a comparison group of beneficiaries who meet the eligibility criteria but who are excluded from the demonstration because they do not reside in the demonstration catchment area. The comparison of impact estimates on claims-based outcome measures from the randomized design with impact estimates obtained from the comparison group approach will provide an indication of the ability of comparison designs to produce valid estimates of program impacts, and how these comparison groups should be structured.

It will be necessary to establish exactly how the programs that will use random assignment will conduct randomization, as well as who will do so. We strongly prefer that MPR conduct the randomization in all sites, to ensure that the process is not inadvertently corrupted. Programs will be expected to obtain consent forms from willing participants, and to fax the forms to MPR for randomization. We are currently investigating ways to ensure that randomization can be conducted at virtually any time of the day or week, with the results returned to the site within a few hours, to avoid any delays in beginning care coordination.

The process of selecting a comparison group for programs that will not implement random assignment is likely to be more complex because we will have to survey both treatment and comparison group members six months after the treatment group has enrolled. This schedule is necessary because the comparison site cases and the treatment group must be interviewed at comparable points in time to obtain valid estimates of impacts on survey-based measures. Thus, if a program identifies three-fourths of its potential enrollees at the time of hospital discharge, it will be necessary to identify three-fourths of the comparison group in the same way from claims data, on an ongoing basis throughout the enrollment period. Selection of the external comparison group for the random assignment sites will be much simpler, because we will not collect survey data on these cases. Selection of the comparison group can therefore be done at a later point in time, after enrollment is complete.

The minimum sample size for the evaluation is 309 cases each for the treatment and control groups, assuming random assignment of all beneficiaries who are eligible and agree to participate in the study. This sample size yields 80 percent power for detecting effects of 10 percentage points on a binary variable with a mean of .50, using a one-tailed test at the .05 significance level. This minimum precision level was selected because the proportion of Medicare beneficiaries with chronic illnesses, such as CHF or chronic obstructive pulmonary disease, who are hospitalized in a given year is about .50, and reductions in this rate of about 20 percent (10 percentage points) may be necessary to cover the cost of the interventions. Our review of best practices found many studies, including a number with strong research designs, reporting care coordination program impacts substantially greater than 10 percentage points (Chen et al. 2000). Thus, we are not concerned about the relatively low probability of detecting smaller impacts on hospitalization rates with this sample. Similarly, although smaller effects of other outcomes, such as patient satisfaction or symptom relief, may also go undetected, such modest improvements in these outcomes may not warrant establishment of a care coordination benefit unless there is evidence of cost saving.

Programs should attempt to enroll a minimum of 343 beneficiaries in each group, so that approximately 309 cases will be available for analysis of survey outcome measures, assuming a 90 percent survey response rate. (See the next section for our explanation of the expected response rate.) Programs are expected to enroll these sample members during the first 12 months of operations, in order to allow adequate time for the evaluation to be completed on schedule. Many programs plan to continue enrolling patients over a longer period, raising the possibility that programs failing to enroll the minimum sample size within 12 months may be able to reach it within a few months afterward. However, given the evaluation time frame, longer intake periods are associated with correspondingly shorter follow-up periods.

The minimum sample size will *not* be sufficient to detect impacts of 20 percent on Medicare cost for subgroups of program participants. The variance of costs is so large that we can be confident of identifying statistically significant treatment-control differences in these samples only if the true impact is nearly 50 percent. Similarly, with this sample size, it is highly possible that we will fail to detect program impacts that are concentrated in a subset of the program's target population, unless these impacts are somewhat larger than 10 percentage points.

A much larger sample size is required to produce comparable precision when a comparison group approach is used. The necessary sample size increases dramatically as the participation rate among eligible beneficiaries drops. That fact, the possibility that a well-matched comparison group cannot be identified, and the difficulty of identifying a comparison group early enough to collect the survey data on comparison site cases at a comparable point in time as for participants are strong incentives for programs to use random assignment.

Of the 15 programs selected for the demonstration, 9 propose sample sizes of 309 to 350 per group, consistent with the minimum specified in the RFP. The other six programs propose samples ranging from 500 to 5,500 per group. We are currently in the process of reviewing each of the research designs to determine whether the proposed sample sizes and designs are feasible and efficient.

The impact evaluation will rely on data from surveys, Medicare claims, intake forms, and program sites.

The evaluation will draw on data from several key sources. We will collect survey data at six months after enrollment on all sample members in sites with target sample sizes at or near the minimum; the survey will be conducted by telephone. We will attempt to interview a sample of approximately 343 cases for programs enrolling substantially more than this minimum. The survey will collect data on intermediate and final outcome measures (discussed in the next section). It will also collect data on patient characteristics that will serve as control variables in the regression models to adjust for chance differences between the treatment and control/comparison groups. Because the consent forms that all participants must complete will contain both contact information and the participants' agreement to complete our survey when contacted in six months, we expect to be able to complete interviews with 90 percent of the sample members (or with the proxies of sample members who are too ill to complete the interviews themselves).

Medicare claims data will be drawn from HCFA files for all sample members for at least 12 months and for as much as 24 months after enrollment, for use in the analyses. These data will provide a wide array of cost and service use outcome measures. Claims covering the 12-month period immediately preceding enrollment will also be drawn and will be used to construct preenrollment service use variables for use as control variables in the regression analyses. Claims data for the preenrollment period will also be used to identify the best comparison sites for each demonstration site.

Data collected by the program sites will also be important data for the evaluation. Intake forms completed along with patient consent forms at enrollment may contain useful variables, such as stage or severity of illness, to include as control variables in the regression models. Data on program costs will be needed for the cost-effectiveness and implementation analyses. The implementation analysis will also require data on enrollment and disenrollment and on the use of special services covered by the program but not by Medicare. These data will be collected periodically from the implementation contractor.

We will also survey 50 physicians serving program participants in each of the demonstration sites, to obtain data on physician satisfaction with the intervention. These surveys will be collected in two waves, with half the sample being selected and surveyed after the 9th month of program operations and the other half being drawn and interviewed after the 21st month. Collecting these data in two waves will enable us to identify changes over time. Response rates are likely to be substantially lower for physicians than for patients, so we will select samples of 80 physicians from each program, including a mixture of primary care physicians and specialists.

Outcome measures will capture quality of care, service use, and costs.

The outcome measures for the evaluation will include measures that reflect patients' well-being and other measures of the quality of care received, as well as their Medicare cost and service use. The quality-of-care measures will include both intermediate and final outcome measures. Intermediate outcomes will include such measures as access to care; adherence to recommended self-care, diet, and drug regimens; knowledge of disease; access to necessary services and information; and receipt of preventive care. We will also collect information on beneficiaries' receipt of some care-coordination services, such as reminder calls about appointments, explanations of how and when to take medications, and help with obtaining home care services. Final outcomes will include such measures as patient satisfaction, patient well-being (for example, functioning, self-rating of health, and symptom relief), mortality, and preventable events (for example, hospital admissions for pneumonia or readmissions for target conditions within 30 days of discharge). The survey will include some disease-specific modules, to capture measures that reflect the quality of care for particular targeted illnesses. Program participants will also be asked about their satisfaction with the intervention.

Cost and service use measures will be constructed from the Medicare claims data. Multiple measures will be constructed for each of the major types of Medicare services. The outcomes will be measured over short-term and longer-term intervals, to determine whether program impacts persist, and to provide early results on program impacts. Program impacts on these outcome measures will be estimated on all study group members, including those who did not respond to the survey and those in high-enrollment sites who were not selected for the survey sample. Impacts on the service use measures will be weighted by average cost per unit of service, to construct an alternative estimate of impacts on Medicare costs that may be less sensitive to outliers.

THE SYNTHESES OF RESULTS WILL RELATE IMPACTS TO CHARACTERISTICS OF PROGRAMS AND PARTICIPANTS.

We will combine the findings from all the site-specific analyses of program implementation and impacts to draw conclusions about whether care coordination programs can achieve the goals of improving patient outcomes and reducing Medicare costs, and about the types of programs that were the most successful at doing so. Synthesizing the results across sites and outcome measures is likely to be the most difficult component of the evaluation—but one of the most important. As noted, given the limited number of sites, it will be difficult to estimate the relative importance of the many intervention components and potentially important structural and operational features of the programs. We will summarize which programs seemed to be successful in improving patient well-being and reducing Medicare costs enough to be at least cost neutral and will then test for whether mean impacts on a few key outcome measures differ for subgroups of programs defined by the characteristics of interest. This assessment will focus on identifying the relationship of impacts to program characteristics that can be specified and monitored. In addition, to identify visible patterns that suggest relationships between favorable impacts and program features, we will present characteristics of the programs, with the programs ordered by the size of the impact on key outcomes.

The syntheses will focus on how program impacts vary with (1) intervention features and quality, (2) the programs' structural characteristics, (3) beneficiaries' characteristics, and (4) program costs. A wide range of measures will be used to capture the different approaches demonstration programs use in attempting to improve patient outcomes, as well as to identify differences in how the programs perform the three basic steps of assessment/care planning, service delivery/monitoring, and reassessment/revision. Structural characteristics that we will use to group programs include type of organization, integration of care coordinators with providers, staffing, and other features assessed in the implementation analyses. We will estimate models on data pooled from multiple sites to assess how impacts varied with patient characteristics, because too few observations will be available to test these relationships for individual programs.

THE INTERIM AND FINAL REPORTS WILL PROVIDE SITE-SPECIFIC ESTIMATES AND SYNTHESES.

We will present the results from the evaluation in a series of reports, with the findings for each site reported at fixed intervals from the time of startup, using a standard format. The case studies, due in month 6 after program startup, will provide findings from the implementation analysis for each site. The first interim site-specific report, due in month 12 after program startup, will update the implementation results for the site and will provide estimates of program impacts on outcomes during the first two months after enrollment for beneficiaries enrolled in the study during the first four months of operation. The second interim site-specific report (due in month 33 after program startup) will provide impact estimates on all outcomes for virtually the entire sample and will update the implementation findings with information obtained during the last round of telephone interviews with site staff.

We will prepare two interim syntheses and one final synthesis, with the interim syntheses drawing from the interim site-specific reports. The first interim synthesis will be delivered 16 months after the first coordinated care demonstration site begins enrolling patients. This report will focus primarily on differences in the programs' implementation, as little data on outcomes will be available for the first interim site-specific reports (see below). The second interim synthesis will be delivered in the 40th month after the first program begins operations and will include virtually all sample members and outcome measures. Each of the two interim syntheses will serve as a mandated Report to Congress, the first due in the 18th month after the first program begins enrollment, and the second due in the 42nd month after enrollment begins.

Given this timing, we will exclude from the first synthesis report any site beginning operations more than three months after the first site begins. The schedule for the Second Synthesis report will allow us to include results from all sites beginning operations within seven months of the first one. We expect that all or nearly all of the program sites will be included in the second synthesis report.

The final synthesis report will present final site-specific impact estimates for all sites and sample members, using the longest follow-up period possible. The report will use all these data to draw final conclusions and to make recommendations about whether adding a care coordination benefit to Medicare appears likely to improve care and reduce costs, and if so, how this benefit should be structured and paid for.

The following schedule presents the due dates for all project reports, assuming a start date for the first coordinated care site of July 2001.

SCHEDULE OF DRAFT REPORT DUE DATES

	Draft Due	
Report	Project Month	Calendar Month
Design report	5	2/01
Site methodologic evaluation	6	3/01
Draft site-specific analysis plans	8	5/01
Site case studies	6 months after site enrollment begins	1/02-7/02*
First interim site-specific evaluation	12 months after site enrollment begins	7/02–1/03*
Second interim site-specific evaluation	33 months after site enrollment begins	4/04-10/04*
First interim synthesis	26*	11/02*
First report to Congress	28*	1/03*
Second interim synthesis	50*	11/04*
Second report to Congress	52*	1/05*
Final synthesis	57	6/05

*Assumes first coordinated care demonstration program starts enrolling in July 2001 (month 10).

I. INTRODUCTION

This report describes the evaluation design for the Medicare Coordinated Care Demonstration. In it, we discuss our approach to the impact analysis, including minimum acceptable sample sizes, statistical methods, data sources, and outcome measures. We also describe the goals and framework to be used in the implementation analysis. These approaches are applicable to all demonstration sites. Many of the details of the design, such as the data available from the sites, timing, exact sample sizes, and methods of intake and randomization, will vary across sites and cannot be specified until the programs' designs have been finalized. The details of the evaluation design for individual sites will be discussed in site-specific evaluation plans to be developed this spring.

A. RATIONALE FOR THE DEMONSTRATION

Beneficiaries with chronic illnesses account for a high proportion of total Medicare expenditures. In 1996, for example, 12.1 percent of all Medicare enrollees accounted for 75.4 percent of all Medicare program payments (Health Care Financing Administration 1998). Much of the high cost of care for these enrollees is due to repeated hospitalizations. Their health care is often fragmented and poorly coordinated across multiple provider types and settings, with insufficient time devoted to education about their condition and appropriate self-care. The suboptimal frequency, timing, mix, and intensity of health care services often leads to poor clinical outcomes, dissatisfaction with care, and higher costs to individual beneficiaries and to the Medicare program. This demonstration project is based on the premise that improving care coordination will substantially reduce the cost of services these beneficiaries receive.

Several recent studies have shown that well-designed coordinated care programs can improve outcomes in substantially commercial populations by better organizing care across providers and providing support to the chronically ill (Rich et al. 1995; Naylor et al. 1994; Wasson et al. 1992; and Aliotta 1996). For example, Rich et al. (1995) showed that better care coordination reduced the rate of readmission for patients with congestive heart failure (CHF) by 43 percent. Wasson et al. (1992) found that regular telephone followup after hospital discharge reduced total medical costs by 28 percent over a two-year follow-up period. The demonstration will test whether models like these can produce comparable results for a Medicare fee-for-service population.

The Balanced Budget Act (BBA) of 1997 requires the Secretary of the U.S. Department of Health and Human Services to conduct a demonstration project testing whether existing models of coordinated care can improve outcomes for targeted Medicare beneficiaries, and whether they can reduce expenditures in the Medicare fee-for-service program. In response, the Secretary, through the Health Care Financing Administration (HCFA), released a Request for Proposals (RFP) in July 2000 for applicants to the Medicare Coordinated Care Demonstration project.

B. DEMONSTRATION SOLICITATION AND DESIGN

In its RFP, HCFA outlined several requirements of programs to be funded under the demonstration that sought to ensure the programs would successfully improve care and reduce costs. Applicants were to be existing providers of coordinated care services; in other words, they must have provided coordinated care services similar to or identical to the coordinated care services proposed for the demonstration for at least one year preceding the date of the RFP (July 28, 2000). Applicants were required to provide evidence, based on prior performance, that Medicare expenditures for their enrollees would be no higher than they would have been in the absence of the demonstration (that is, programs must be cost neutral to Medicare). In addition, programs were required to show evidence that they would increase the quality of care provided and increase beneficiary and physician satisfaction. Programs were required to detail the processes they plan to use to identify, recruit, select, enroll, and discharge participants from the program. They also were asked to describe how they will determine patient eligibility. In addition, applicants must have demonstrated that they have sufficient infrastructure, including personnel, to carry out the demonstration. Finally, they were required to provide evidence that they previously had been able to reduce medical service use for individuals with the target conditions.

HCFA received 58 proposals in response to its RFP.[1] Of these, a diverse mix of 15 programs were funded as demonstration sites.[2] Nine of the 17 demonstration sites are disease management programs; 8 are case management programs. Most of the sites will be sponsored by hospitals, health systems, or academic medical centers, although several other types of organizations also are represented (Table I.1).

The demonstration sites will serve Medicare beneficiaries with a number of chronic conditions. Many programs will target multiple conditions. Three programs will focus on providing services to the frail elderly. Nine will target CHF, six will target other types of heart disease, five will focus on diabetes, and two will target pulmonary disease. The sites are located throughout the country, but most will serve metropolitan areas and surrounding counties. Eight programs will serve urban areas only; seven will serve both urban and rural areas; and two will serve rural areas only. In terms of study design, 16 programs plan to use random assignment and 1 program plans to use a comparison group design. The program fees requested by the sites range from $382 per beneficiary per month (pbpm) to $85 pbpm (six sites, <$200 pbpm; five sites, $200 to $300 pbpm; two sites, $301 to $400 pbpm; and two sites, fees unspecified).

[1] See Archibald and Brown (2000) for a complete description of the characteristics of the applicant programs.

[2] In addition, two programs operated by Lovelace, one for CHF and one for diabetes, were previously funded under a different demonstration and are to be included in the evaluation.

TABLE I.1

NUMBER OF DEMONSTRATION SITES,
BY PROGRAM TYPE AND SPONSOR

	Case Management	Disease Management	Total
Commercial Vendor	0	3	3
Health System[a]	3	0	3
Academic Medical Center	1	3	4
Coalition	2	1	3
Hospice	1	0	1
Retirement Community	1	0	1
Total	**8**	**7**	**15**

[a] Health systems are organizations that include hospitals and affiliated physician groups. Some of these organizations also have home health agencies and/or skilled nursing facilities.

C. GOALS OF THE EVALUATION

The goals of the demonstration evaluation are to: (1) provide HCFA with unbiased estimates of the ability of the 17 care coordination demonstration sites to provide better and more cost-effective care for chronically ill Medicare beneficiaries; (2) assess the extent to which the effectiveness of care coordination depends on patient and program characteristics; and (3) provide guidance on the feasibility and desirability of establishing a Medicare coordinated care benefit, and on how that benefit should be structured and administered. To provide this information, the evaluation must generate both rigorous quantitative estimates of the programs' impacts and qualitative analyses of the programs' processes; the processes to be studied include program design, implementation, and operation.

The impact analyses will test the hypotheses that the demonstration programs (1) lower costs, (2) improve quality of care, and (3) improve patient and physician satisfaction. The cost analysis will include impacts on costs to the Medicare program (including care coordination program costs), Medicare service use, and beneficiaries' out-of-pocket costs. The analysis of the quality of care will assess the care delivery process and the clinical outcomes of Medicare beneficiaries. In the satisfaction analysis, both patient and physician satisfaction will be covered. Subgroup analyses will test whether care coordination interventions are more effective for certain types of patients than for others.

The implementation analysis will study the planned interventions as envisioned by each site, each site's actual experience, and the factors that impeded or facilitated each site's efforts. The detailed descriptions of each site's planned interventions will cover the types of organizations that are implementing the interventions, the groups targeted to receive the services, and the focus of the interventions (that is, improving patient adherence, improving physician practice, improving communication among multiple providers and patients, or improving arrangement of non-Medicare services). The descriptions will also include the programs' approach to the basic steps that care coordination programs follow: outreach, initial assessment and care plan development, delivery of interventions, and periodic monitoring and reassessment of patient progress (Chen et al. 2000). Other key program elements to be described are structural features of the program, such as staff composition; staffing ratios; program organization; and the presence, nature, and effectiveness of a quality improvement program. The analyses of each site's actual experience will assess its success in implementing the planned interventions. We also will appraise each site's performance in the areas of patient enrollment, physician "buy-in", staff recruitment and training, and its local service environment.

Finally, the synthesis will combine the findings from the site-specific analyses, using both impact estimates and implementation analysis findings, to draw inferences about the types of programs that appear to be most successful, and for which groups. The interim and final synthesis reports will show how program effects differ with various characteristics of the interventions and patients. As required by HCFA, the synthesis also will be the basis for two reports to Congress about the feasibility of creating a coordinated care benefit in the Medicare fee-for-service program.

D. CHALLENGES FOR THE EVALUATION

A variety of technical and logistical challenges must be overcome to achieve the evaluation's objectives. The primary logistical challenge is the programs' different start dates will require careful planning of the evaluation tasks to ensure that the steady stream of site-specific results are produced on time. The main technical challenges are to obtain valid, comparable estimates of impacts for each program, and to determine which patient and program characteristics are associated with effectiveness.

1. Estimating Impacts

Four factors may complicate estimation and detection of impacts: (1) the sample sizes likely to be achievable by each site; (2) the time frame for the demonstration; (3) the feasibility of credible comparison strategies; and (4) the development of study protocols that accommodate the needs of individual programs while allowing for comparison across sites.

Some sites may find it challenging even to identify and enroll the minimum sample size of 686 patients. Obtaining patient and physician buy-in to care coordination programs is difficult, and random assignment may increase the difficulty of completing this task. Although 686 cases should be adequate to detect overall impacts of policy-relevant size, a sample of this size has very limited power to meet the project goal of determining whether an intervention is more effective for particular types of patients.

The timeline for the evaluation imposes a strict time constraint on our analyses. As each demonstration site is awarded, we must promptly evaluate its proposed research design and protocol and, if necessary, quickly help the site overcome any identified weakness. Furthermore, the analysis must be sensitive to the fact that program impact may not occur until several months after a patient enrolls, or that some programs may reach peak effectiveness several months after startup.

The research designs selected by the sites may present difficulties. Random assignment of patients is generally the strongest study design, but if the program leads physicians to change practice patterns for all their patients, the programs' impacts may be underestimated. Furthermore, ensuring the integrity of the randomization process can be difficult. If physicians are randomized, control physicians' practices may be affected through normal professional interactions with treatment group physicians.[3] The best option in sites in which random assignment of patients or physicians is undesirable, is a matched-site comparison design. However, this design requires substantially larger sample sizes than does a random assignment design and yields biased estimates if other differences between the two sites cause outcomes to differ.

The analyses must be standardized sufficiently such that results can be compared across sites, despite differences in experimental designs, yet flexible enough to take advantage of data available only for a particular site. The sites are likely to vary by study design, type of program, intake procedures, data availability, and target patient population. The key outcome measures and measurement strategies will have to accommodate the features of individual sites, while meeting the goals of reporting common measures for cross-site comparisons.

2. Determining Patient and Program Characteristics Associated with Effectiveness

The number of program characteristics potentially related to program effectiveness is large relative to the number of programs in the demonstration. Thus, it will be difficult to identify the program features or combinations of features that are important for success. Observation from multiple sites will be pooled in order to detect differences in impacts across *patient* subgroups. However, pooling may mask important subgroup differences that exist only for sites serving certain target populations or using more effective interventions.

[3] Care coordination is not widespread in the Medicare fee-for-service population. It is unlikely that patients enrolled in comparison groups will receive any care coordination interventions from outside sources. Contamination of the comparison groups in this way should not be a challenge to precisely measuring differences in impacts between beneficiaries receiving care coordination interventions and beneficiaries in the comparison groups.

E. GUIDE TO THIS REPORT

In Chapter II, we describe the implementation analysis objectives and approach, and then discuss the impact analysis. Chapter III describes the hypotheses, research design, data sources, outcome measures, and statistical procedures that we will use to overcome the methodological challenges of the evaluation. Chapter IV explains how we will synthesize the findings from the site-specific process and impact analyses. Chapter V reviews the reports that will be produced, and Chapter VI provides a timeline. The Appendix (to be sent under separate cover) contains the site visit protocols.

II. DESIGN OF THE IMPLEMENTATION ANALYSIS

A key element of the evaluation will be to describe in detail how each demonstration program was implemented, and to assess what program features appear to be associated with its success or failure. The impact analysis for the evaluation will estimate program effects on patient health, patient costs, and patient and provider satisfaction. However, because the number of program features that might affect success is much greater than the number of programs being evaluated, a strictly quantitative approach to identifying program features associated with effectiveness is not feasible. Thus, the task of identifying and describing program features potentially associated with success lies largely with the implementation analysis. To accomplish this task, the implementation analysis will provide a detailed description of each program and its evolution during the demonstration, as well as an overview of the service environment in which it functioned. The analysis will then classify each program according to its key features; the classification, in turn, will be used in evaluation syntheses to describe associations between program features and effectiveness.

In this chapter, we present the goals and key questions addressed by the implementation analysis, a conceptual framework of program classification that will provide the underpinnings of the analysis, and a description of the data collection activities that will support the implementation analysis. Chapter IV describes how we will combine the findings from the implementation analysis with those of the impact analysis in the evaluation syntheses.

A. GOALS AND KEY QUESTIONS ADDRESSED

The implementation analysis has two primary goals: (1) to describe the key features and target population of each program, determine whether the program was implemented as designed, and identify any changes made to the design and reasons for those changes; and (2) to compare key features across programs and provide insights about why the programs succeeded or failed. These insights can then inform decisions by HCFA and others for future care coordination/disease management programs.

We will present the findings of the implementation analysis in sets of three evaluation reports, one set for each program in the evaluation. Each of the three reports (Case Studies, First Site-Specific Interim Evaluation, and Second Site-Specific Interim Evaluation) will focus on evolution of a program at a different point in time. The Case Study will focus on describing the history and design of the program and its early implementation experiences. The First Site-Specific Interim Evaluation will provide a detailed description of features of the program as implemented, problems encountered, and changes made in response to those problems. It will also provide an overview of the health service environment in which the program was implemented. The Second Site-Specific Interim Evaluation will focus on changes made to the program through the end of the demonstration, features that appear to be associated with program success or failure, and lessons for future care coordination/disease management programs. Synthesis reports that follow the two interim reports will compare features across programs to develop lessons about characteristics that appear to be associated with the ability of programs to achieve their goals. A third and final Synthesis report will make recommendations for future care coordination efforts.

The evaluation's Site-Specific Analysis Plan, a program-specific supplement to this design report, will identify necessary changes to the data collection instruments and procedures for the implementation analysis for each program. It also will present the data collection plan and schedule for the program. (Table II.1 summarizes the planned content of each report with respect to the implementation analysis.)

TABLE II.1

ISSUES ADDRESSED BY IMPLEMENTATION ANALYSIS DELIVERABLES

Site-Specific Analysis Plans

Data Source: site proposal	Due in Draft: 2 months after site award

- Intervention design and classification of program using working classification scheme
- Necessary site-specific modifications to site contact protocols or other data collection activities
- Site-specific work plan (schedule of site contacts, evaluation team and program staff involved, site-specific report due dates, collection of program data)

Case Studies

Data Sources: first telephone contact and program documents	Due in Draft: 6 months after start of patient enrollment

Relationship between program, host, and providers

- Relationship of program to host organization
- Relationship between program and other providers who will serve its patients

Intervention history and key intervention features

- History of intervention prior to demonstration: who designed it, where was it used previously, how effective was it, how was it adapted to the demonstration
- Proposed intervention features
- Intervention delivered to date

Target population and program goals

- Target population, expected program size
- Proposed goals for patients, providers, and health care system as a whole

Major start-up problems

- Early problems encountered: enrollment shortfalls, screening criteria not yielding desired target group, contamination of control group, difficulties hiring or retaining staff, physician opposition
- Early changes to planned targeting or intervention, if any, and reasons changes made

Other topics covered

- Acquisition of staff, office space, and equipment
- Implementation of proposed experimental design
- Implementation of evaluation data collection: intake records, other data via implementation contractor and HCFA

First Site-Specific Evaluation

Data Sources: site visit; program documents; analysis of enrollment and disenrollment records; analysis of special service use records or reports (if such services offered)	Due in Draft: 12 months after start of patient enrollment

Targeting

- Target criteria used to date
- Participation to date: enrollment process; participation rates; reasons for participation and refusal to participate; comparison of participants with eligible nonparticipants; drop-out rates and reasons for dropout
- Major changes to target criteria or screening and outreach procedures since Case Study; reasons for changes

Intervention implemented

- Intervention delivered to date

Table II.1 *(continued)*

- Physician acceptance of or resistance to intervention
- Other major barriers to and facilitators of implementation
- Major changes since Case Study; reasons for changes

Program staff

- Types of staff used and precise nature of staff contact with patients and providers

Quality assurance

- Quality assurance procedures

Other topics covered

- Service environment overview
- Progress to date in achieving goals and desired outcomes
- Record keeping and data management

Second Site-Specific Evaluation

Data Sources: follow-up telephone contact; program documents; analysis of enrollment and disenrollment records; analysis of special service use records or reports (if such services offered); analysis of program cost reports	Due in Draft: 33 months after start of patient enrollment

Changes since First Evaluation

- Target criteria used and intervention delivered to date; major changes since First Evaluation; reasons for changes
- Changes in participation and drop-out rates; reasons for changes
- Major changes in service environment since First Evaluation
- Changes in physician acceptance of or resistance to intervention since First Evaluation
- Other major barriers to and facilitators of implementation

Features (internal and external) associated with success or failure

Program costs

Recommendations for an ongoing program

First, Second, and Final Syntheses

Data Source: cross-site comparison of Site-Specific Evaluations	Due in draft: First Synthesis, 16 months after start of patient enrollment; Second Synthesis, 40 months after start of patient enrollment; Final Synthesis, 57 months after evaluation contract award

- Update classification scheme and compare programs along classification dimensions
- Program features associated with success and failure
- Recommended structure for future care coordination programs or Medicare care coordination benefit

The Case Studies and Interim Evaluations will seek to answer the following overarching research questions:

- ***Case Studies***
  - What was the relationship between the demonstration program and its host organization, and between the program and other providers?
  - Who designed the intervention, where was it used prior to this demonstration, and how was it adapted for the demonstration?
  - What were the key features of the intervention as designed?
  - Whom did the program target, and what were its goals for patients, their providers, and the larger health care system?
  - What were the major start-up problems, and how were they resolved? How will those problems or solutions affect the evaluation?
- ***First Interim Site-Specific Evaluations***
  - How (and how well) were the targeting criteria and features of the intervention implemented? If implementation differed from design, why were changes made?
  - What types of staff did the program use, and what was the nature of staff contact with patients and providers (particularly care coordination/disease management staff)?
  - How did the program ensure the quality of its intervention?
- ***Second Interim Site-Specific Evaluations***
  - What features appeared to be associated with the success or failure of the program? If program features were not implemented as planned, how important was the change to program success or failure?
  - What factors external to the program affected its success or failure?
  - What was the monthly cost of the program (including its level of profit or loss and cost to HCFA)?
  - What should an ongoing care coordination program look like? If the program was not implemented as planned, what lessons do these changes provide for the future?

B. CLASSIFICATION OF PROGRAMS

The implementation analysis will develop a classification scheme for care coordination/disease management programs that will serve two purposes. First, it will simplify comparisons of the demonstration and other programs and has the potential to provide a common language for the many activities currently labeled "care coordination/case management/disease management." Second, it will provide a conceptual framework for organizing the evaluation's implementation data collection and analysis and integrating it with the impact analyses. This framework will facilitate comparing programs in order to identify features associated with program success and failure.

The vast range of program features we will consider for classifying and assessing programs can be categorized by using terminology of traditional health care quality evaluation: structure, process, and outcome (Donabedian 1980). For the purposes of this evaluation, outcomes will refer primarily to patient outcomes; that is, knowledge of and adherence to recommended treatment, satisfaction with care, health and functioning, and service use and costs. Patient outcomes, based on survey and claims data, will define program success or failure as determined by the evaluation's impact analyses. The primary sources of success or failure are likely to be structural and procedural features and how well these features are implemented; however, environmental factors outside a program's control may also affect success.[1] Measures of structural and procedural features will be based primarily on information from telephone and in-person contacts with the demonstration program staff and review of program documents.[2]

[1] For example, environmental factors, such as the labor market, can affect program impacts. MPR is currently evaluating a random assignment-based consumer-directed intervention for Medicaid beneficiaries receiving personal assistance services. The very tight labor market for personal assistance workers may have a major effect on intervention impacts because many control group members are unable to receive the level of personal assistance they have been assessed as needing, whereas treatment group members have been able to hire family and friends to provide assistance. In the absence of information about the labor market, consumer-directed care would appear to be much more costly than traditional agency-provided assistance.

[2] Although some program features are clearly either structural (like staff size and composition) or procedural (like the type of intervention provided), the distinction is less clear for other features and not particularly germane to the goals of the implementation analysis. Thus, we do not use the structure/process distinction in the discussion that follows.

1. Recent Research Describing Successful Care Coordination

Recent research *suggests* a number of features may be associated with the success of care coordination/disease management programs but is by no means conclusive on the subject. HCFA's Best Practices in Coordinated Care project (Chen et al. 2000) and Medicare Case Management Demonstrations evaluation (Schore et al. 1997) identified the following features related to physician behavior, background of care coordination staff, certain intervention components, and financial incentives as likely predictors of success:

- ***Physicians.*** Obtaining physician buy-in for care coordination and involving them in it. Improving medical treatment with the use of evidence- or consensus-based guidelines
- ***Care Coordinators.*** Using nurses who have at least baccalaureate degrees in nursing or using nurses who have experience with community nursing and relevant clinical expertise
- ***Intervention.*** Using a comprehensive, multidisciplinary assessment whose end product is a written care plan that is used to monitor patient-specific goals throughout the life of the intervention. Providing feedback to care coordinators about patient progress toward these goals. Providing patient education in self-care; providing support for lifestyle modifications. Integrating fragmented care. Arranging for community services. Taking a proactive, preventive approach to patient problems
- ***Financial Incentives.*** Providing incentives for the program to both meet patient goals and reduce total health care costs

In addition to associating the same features with successful programs, other recent research has also stressed the importance of careful targeting, the ability to change patient behavior, and provision of a disease management intervention that is not too narrowly focused.

- ***Targeting.*** Identifying the highest-risk patients for whom the program is likely to be effective, rather than taking a population-based approach, as a necessary condition for achieving cost-effectiveness (Rector and Venus 1999). Screening for prevalent geriatric syndromes, such as physical inactivity, falls, depression, incontinence, misuse of medications, and undernutrition (Fox 2000)
- ***Changing Patient Behavior.*** Rather than relying solely on cognitive intervention (that is, factual patient education), providing an intervention that changes patient self-care behavior and teaches patients to manage their own care (Williams 1999; Lorig et

al. 1999; and Vernarec 1999). Factors associated with success in changing behavior include (1) patient understanding of the benefits of treatment adherence; (2) patient access to needed transportation (LeDuc et al. 1998); and (3) use of an approach that addresses cognitive, behavioral, and affective issues related to chronic illness and behavioral change (Roter et al. 1998; and Aubry 2000).

- ***Integrating Disease Management.*** Many elderly program participants will have more than one chronic illness; thus, providing an intervention that addresses all comorbidities, as well as prevention and psychosocial barriers to good health care. Programs "carving out" treatment of a single disease may increase fragmentation in the system if they focus on the treatment of that disease to the exclusion of coexisting illnesses and move the locus of care away from the primary care physician (Bodenheimer 1999; and Hagland 2000).

In developing the evaluation's classification scheme, we will consider a broad list of features. We will do so because the state of current research leaves us unsure of exactly which program features will lead to success in this demonstration and to support the goal of fully documenting the design and implementation of the demonstration programs.

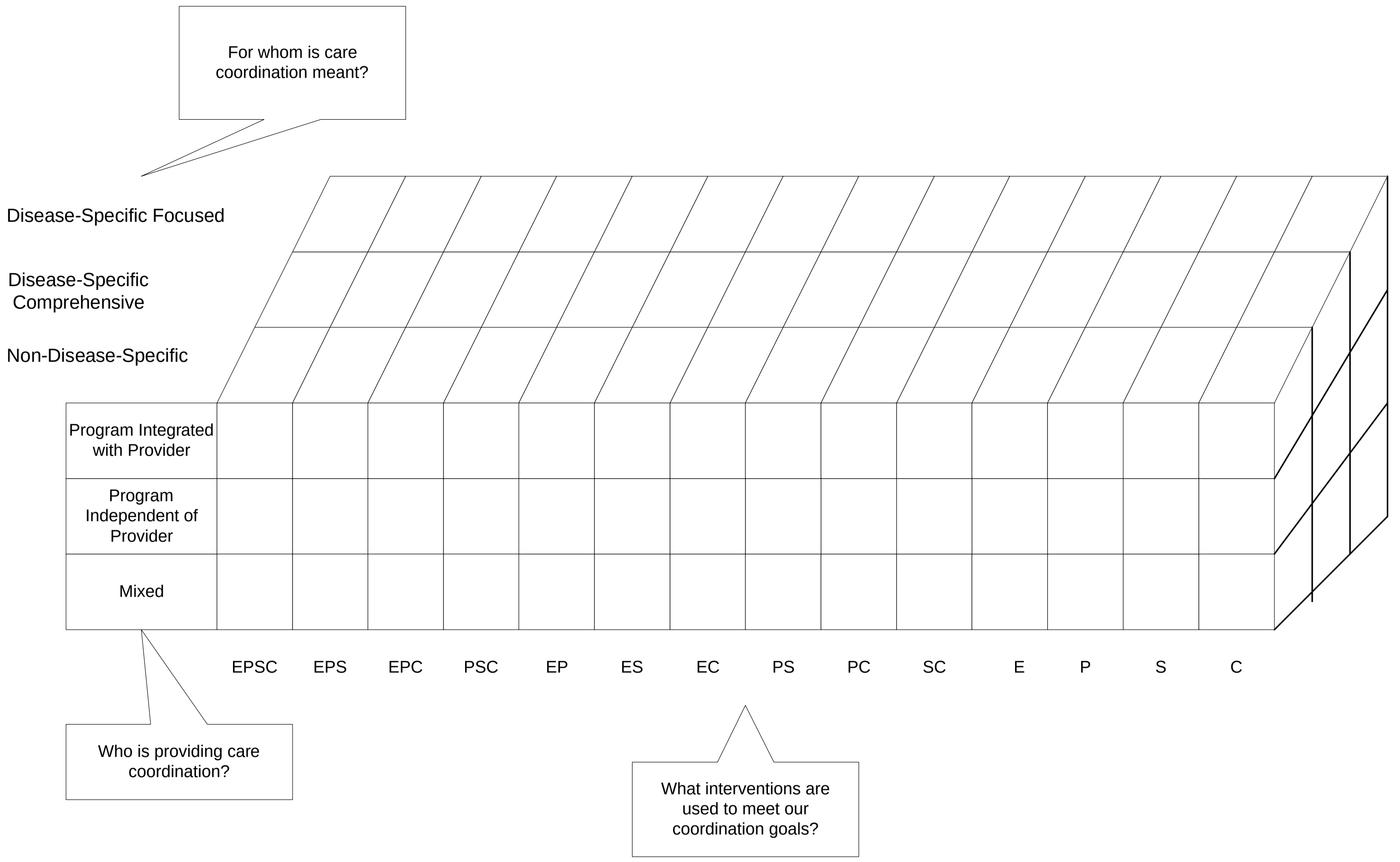
2. Program Classification Features

The evaluation has two interrelated tasks related to program classification. The first is the development of the comprehensive list of program features to provide a data collection framework for the implementation analysis. The second is the development of a parsimonious classification composed of a few salient care coordination/disease management program features to compare programs in and outside the demonstration. To accomplish the latter task, we will begin by describing (1) who is implementing the program, (2) for whom the program is implemented, and (3) what the basic program approach is (see Figure II.1).

- ***Who: Relationship of Program to Providers.*** Whether the program is full integrated with its patients' providers (for example, a care coordinator employed by a physician or physician group), independent of providers (for example, if the program provided disease management by a commercial vendor), or some combination placing it between full integration and independence (for example, a care coordinator employed by the program sent to work with physicians employed by the host organization)

Figure II.1

Initial Care Coordination Classification Scheme


For whom is care coordination meant?
Disease-Specific Focused
Disease-Specific Comprehensive
Non-Disease-Specific
Program Integrated with Provider
Program Independent of Provider
Mixed
EPSC
EPS
EPC
PSC
EP
ES
EC
PS
PC
SC
E
P
S
C
Who is providing care coordination?
What interventions are used to meet our coordination goals?

E = Improve patient education and adherence
P = Improve provider practices
S = Provide/arrange for service not covered by Medicare
C = Improve communications and coordination among and between providers

- ***For Whom: Target Population and Focus.*** Whether the program targets beneficiaries with a specific disease and focuses on managing that disease; beneficiaries with a specific disease, but also addresses comorbid conditions and psychosocial needs; beneficiaries who do not have a specific disease, providing, for example, a care coordination intervention for a chronically ill or frail population
- ***What: Intervention Approach.*** Which of the 15 possible combinations of the following four approaches the program uses to achieve its goals: (1) improving patient education and adherence to treatment recommendations, (2) improving provider practice, (3) providing or arranging for non-Medicare-covered service, and (4) improving communication and coordination among providers and between providers and patients

We will modify this initial classification scheme as the evaluation progresses to reflect what we learn, using the comprehensive list of features described below to guide our data collection. We also allow for the possibility that we will learn of still other important program features to add to this list.

We have grouped the many features of care coordination/disease management programs (including the three described above) within the following broad dimensions: program context, target population and outreach procedures, intervention features, staffing and intensity of staff contact with patients, quality assurance procedures, financial issues, and record keeping. Clearly, not all dimensions or features within dimensions will apply to all programs. Moreover, in assessing which features are associated with successful programs, it will be important to distinguish between a program simply having a particular feature and how well program staff implement that feature. Table II.2 provides an overview of program features within the seven dimensions.

Knowledge of ***program context***, which includes a program's goals, its relationship to its organizational host and other area providers, and the history of its program design, is key to understanding why program staff made certain decisions and why programs were implemented in a particular way.

TABLE II.2

DIMENSIONS OF PROGRAM CLASSIFICATION AND ASSOCIATED FEATURES

Program Context

- Host organization: type of organization; host's reasons for applying as demonstration site; program relationship to host; history of demonstration design

- Relationship of program with providers: integrated with providers; independent of providers; mixed; physical location of program relative to providers

- Goals: overarching program mission or goals; specific patient outcomes expected

- Service environment: important events in local service environment or important features of environment that could affect program operations or evaluation (for example, health care labor market, availability of transportation, home care and other support services)

Target Population and Outreach Methods

- Target population: numbers and types of beneficiaries targeted and how targeted; that is, not disease-specific, disease-specific/nonintegrated (disease management "carve-out" model), disease-specific/integrated (comorbidities also treated, patient psychosocial needs also addressed); if disease-specific, which diseases targeted; if not disease-specific, what targeting criteria used

- Outreach and intake: eligibility criteria (formal vs. informal); method of case finding; marketing; sources of referral to program; extent to which screening criteria effective in identifying target population, extent to which outreach activities effective in reaching enrollment targets; use of informed consent; random assignment (if applicable)

- Program participation: participation rates, reasons for participation and refusal to participate, comparison of participants with eligible nonparticipants; length of stay in program, drop out rates, and reasons for dropping out

Intervention Features

- Intervention means to achieving goals, broadly described: improve patient education/adherence; improve physician/provider practice; provide/arrange for non-Medicare services; improve communication and coordination among providers and between providers and patients

- Assessment and care coordination planning: how done, to whom, and by whom; time from intake to assessment, time from assessment to care plan implementation

- Monitoring, reassessment, care plan revision: process, frequency, who performs

- Case close-out: performed under what circumstances, how, and by whom

- Care coordinator communication with physicians and other providers about patient: how (formal/informal, telephone/inperson/ written, group meetings/individual conversations); how often; with whom

- Care coordinator role in sequencing care and providing for needed information for patient appointments with providers

- Patient education: how conducted and by whom

- Support for lifestyle changes: how provided and by whom

- Consumer empowerment/self-management: any attempt to teach patient to act as own care coordinator in the long run?

- Provider education: how conducted and by whom

- Service arrangement or provision: range of services; if services used, do care coordinators need prior authorization; complexity of paperwork that requests services; whether care coordinators monitor receipt of services; whether care coordinators provide any hands-on care

- Patient advocacy and provision of emotional support

- Degree to which practice is standardized (for example, uses protocols/guidelines, problem lists, forms); degree to which standardized process can be tailored to individuals when necessary

- Use of automated communication/reminder devices

Table II.2 (*Continued*)

Staff and Staff Contact with Patients
• Program staff (for example, whether uses care coordinators/case managers; other types of staff, such as program director, medical director, IT staff, financial staff, care coordinator extenders); staff background and education; staff roles and responsibilities (in particular, who has overall responsibility for patient)
• Patient contact: staffing ratio; mode of patient contact (mail, telephone, in-person, on-line); frequency of contact; anticipated length of program duration for each patient; whether intensity of contact related to perceived level of patient risk
Quality Assurance
• Training and supervision of staff
• Means for ensuring intervention delivered as designed
• Means for determining whether program is meeting goals and achieving patients outcomes
• Procedures for receiving and resolving patient and provider complaints
• Efforts to develop and maintain physician buy-in and involvement
• Efforts to provide data and other types of feedback to care coordinators and physicians
• Efforts to maintain patient participation and satisfaction
• Approach to making changes to improve quality
Financial Issues
• Financial incentives to achieve program goals and desired patient outcomes
• Program costs, by major activity; funding in addition to that provided by HCFA
Record Keeping
• Information systems: how records are kept; which records are automated; how and to what degree care coordinators, providers, and other program staff share information

As suggested by its inclusion in our initial classification scheme, we are particularly interested in the degree to which a program is integrated with the health care providers its patients will use; it seems reasonable that a closer organizational relationship between the program and providers will facilitate the flow of information about the patient, increase provider buy-in for the program, and, as a result, promote care coordination.

A program's ***target population*** is closely linked to the type of intervention implemented and the intervention's focus. For example, a program may target beneficiaries with a specific disease and thus provide an intervention that focuses on improving management of that disease. If the program targets beneficiaries seen to be more generally at risk of high costs, its intervention may focus on reducing risk and overcoming a fragmented care delivery system. Moreover, the use of a particular criterion may render certain program features unnecessary for most targeted patients. (For example, a program targeted to patients with CHF may not need to develop close links with community service providers.)

It will be critically important to know precisely what screening criteria are employed to identify targeted beneficiaries, and how program ***outreach*** was conducted (that is, how patients were recruited for or referred to the study), as well as how accurately outreach and screening procedures identified the target population. Indicators of the effectiveness of outreach activities include the proportion of "false positives" identified by screening criteria, participation rates, and, to a lesser extent, disenrollment rates. Reasons for participation, refusal to participate, and dropping out, as well as comparisons of participants with eligible nonparticipants will also shed light on the effectiveness of outreach and screening. Moreover, the participant/nonparticipant comparisons will suggest which types of eligible beneficiaries are not attracted to the intervention; this information, in turn, has implications for both the number of beneficiaries who might be served nationally and estimates of any expected cost savings for the Medicare program.[3]

A program's ***intervention features*** include its broad approach to achieving stated goals and the specific activities it undertook to implement that approach. As noted, we envision four basic approaches to care coordination/disease management: (1) improving patient education and adherence to treatment recommendations, (2) improving provider practice, (3) providing or arranging for non-Medicare-covered services, and (4) improving communication and coordination among providers and between providers and patients. We expect that each demonstration program will use 1 of the 15 mutually exclusive and exhaustive combinations of these four fundamental approaches, depending on its target population and goals.

In assessing intervention features, it will be particularly important to distinguish between a program merely providing an approach and how well it implements the approach, as revealed by a close examination of the activities undertaken to implement the approach (and care coordination more generally) and how well those activities were conducted. For example, if a program had the goal of improving patient education and adherence, we would examine the type of curriculum used, whether the intervention was solely cognitive or more comprehensive, whether education was provided in a group setting or one-on-one, what types of staff provided the education, whether the intervention taught self-management for the longer term, and whether the program was followed up to determine that patients understood the material being taught and were making desired behavioral changes. We would then assess how well patient education was implemented, perhaps on a three-point scale (excellent, adequate, or poor).

[3] As described in more detail in Chapter III, estimates of participation rates and comparisons of participants with eligible nonparticipants will be based on Medicare eligibility and claims data. Eligible nonparticipants will be identified, to the extent possible, by translating program screening criteria into measures that can be constructed from data items on the Medicare files.

Staff and staff contact with patients describes the experience and background of program staff and the roles and responsibilities of staff within the program, particularly care coordinators, if the program employs them. (If the program does not employ care coordinators, it will be important to determine who has overall responsibility for ensuring that patients receive the program intervention, and that program goals are being met.) This dimension also will provide information about the nature and intensity of staff contact with patients.

Quality assurance includes procedures for training and supervising staff and for ensuring that overall and patient-specific program goals are being met. This dimension also includes efforts to ensure interventions are being implemented as planned; to provide feedback to case managers and physicians about how they and their patients are doing; and to develop and maintain physician buy-in, patient participation, and physician and patient satisfaction.

The last two dimensions cover features pertaining to ***financial issues*** and ***record keeping***. Financial features include whether the program used financial incentives to achieve its goals (such as profit sharing with HCFA after the first demonstration year) and the level and type of program costs. It will be necessary to examine the level and type of costs to determine the cost effectiveness of a program. We will also describe how programs maintain and share patient records, because care coordinators and providers who can access information about a given patient are likely to provide less fragmented, better coordinated care.

We do not anticipate developing a single summary score that combines assessments across features or dimensions, as any weighting of these components would be arbitrary. Rather, we believe it will be most useful when developing our synthesis reports to identify which programs were effective (as measured by impact estimates), and, after taking their cost into account, to determine which features the effective programs had in common (and similarly, which features ineffective programs had in common).

Finally, we reiterate that it will not be possible to attribute the success or failure of programs to certain features (or their absence), as there are many more program features than there are programs in the evaluation. At best, the evaluation will only be able to identify *associations* between some program features and program success.

C. DATA COLLECTION

We will obtain data for the implementation analysis from semistructured, in-person and telephone contacts with program staff; review of program documents; and descriptive analyses of patient- and program-level data collected by the program.

1. Program Staff Contacts

Data collection and reporting for the implementation analysis will reflect a variety of professional perspectives, drawing on the backgrounds of the following staff, who comprise the implementation analysis team: care coordination research and implementation analysis (Jennifer Schore), care coordination practice/consulting and nursing (Sherry Aliotta), medicine and care quality (Arnold Chen), and health care administration and care quality (Nancy Archibald).

About a month after program award, we will send a letter to the program director that provides an overview of the implementation analysis and its data collection activities, and that requests program documents, such as operational protocols, marketing materials, staff training materials, patient education materials, standardized treatment protocols/guidelines, and assessment and other forms, as they are produced. (We will also speak to the implementation contractor prior to each visit to identify program-specific problem areas or other issues of particular interest to the evaluation.) The following table provides an overview of the type, timing, and purpose of the three program staff contacts for the implementation analysis.

Type of Contact* (Time/ Program)	Timing of Contact	Deliverable Supported by Contact	Type of Information Collected During Contact
Telephone (about 3 hours)	About 2 months after program begins enrolling patients	Case Studies	Start-up and early experiences with implementation
Site visit (1 full day)	About 6 to 7 months after program begins enrolling patients	First Interim Site-Specific Evaluations	Program features as implemented, reasons for changes from proposed features, implementation barriers, and challenges
Telephone (about 3 hours)	About 24 months after program begins enrolling patients	Second Interim Site-Specific Evaluations	Changes to program features over past 18 months, features believed to be associated with success, features that would change in the future, lessons

*Program staff to be contacted for the telephone discussions include project directors, medical directors, care coordination supervisors, and financial staff. Participants for the in-person visits include these staff plus enrollment coordinators, care coordinators, and physicians.

As the table illustrates, we will formally contact program staff at three points in time. (We may have informal contact more frequently to follow up on emerging issues.) The first and last of those contacts will be by telephone: about 2 months after the program begins enrolling patients, to provide input to the Case Study, and about 24 months after the program begins enrolling patients, to provide input to Second Site-Specific Evaluation. The second contact will be in person, about six months after the program begins enrollment, to provide input to the First Site-Specific Evaluation. One of three implementation team members (Ms. Aliotta, Ms. Archibald, and Dr. Chen) will be assigned to each program and will have responsibility for all three program contacts.[4] (The team leader, Ms. Schore, will also participate in all in-person site visits.)

We believe that information from telephone contacts with selected program staff combined with information from program documents (such as proposals and materials programs prepared following the award of their contracts) will be more than sufficient to support the Case Studies. The first round of contacts will occur just after training site visits by the implementation contractor and while program staff will be extremely busy starting to enroll patients; thus, program staff are likely to find telephone contact less burdensome than on-site visits. The calls also will give the implementation team members the opportunity to introduce themselves to staff before they make their in-person visits, four to five months later. Finally, it will be more efficient to collect the information required for the Second Interim Evaluations by telephone, rather than in in-person visits.

Each contact will be guided by a semistructured protocol to ensure that implementation team members collect all necessary information in the most uniform way possible, while leaving some leeway to pursue issues that may be relevant only to a particular program or only to cover unanticipated developments. Before each round of site contacts, the implementation team leader will train team members in the use of the protocols. This training will promote inter-rater reliability in the use of the protocols and will ensure that each team member shares a common understanding of the goals of the program contacts. The implementation team will meet after each program contact (or group of contacts occurring at roughly the same time) to review findings and to identify any information that requires a call back to the program. (Appendixes A

[4] This responsibility includes arranging for, conducting, and writing up notes from all telephone and in-person contacts and writing the Case Study report for their own programs (based on an outline developed by the team leader).

through C include draft protocols for the three contacts, which have been developed by Ms. Aliotta and Ms. Schore, with input from Dr. Chen and Ms. Archibald.)

Each round of telephone contacts will total roughly three hours per program. We anticipate contacting the following program staff: program director (one hour), medical director (one-half hour), care coordination supervisor (one hour), and financial staff (one-half hour). The in-person site visit will last one full day. We plan to meet with the same staff with whom we spoke during the first telephone contact, as well as with the enrollment coordinator, care coordinators, physicians, and information systems staff.

The team leader will develop outlines for the parts of evaluation deliverables pertaining to the Implementation Analysis; review the Case Study reports written by team members; write sections of the First and Second Site-Specific Evaluations and Syntheses pertaining to implementation analysis (with Ms. Aliotta); and, as noted, participate in all in-person site visits to facilitate uniformity across site-specific reports. Dr. Chen and Ms. Archibald will review the findings of the implementation analysis presented in the Site-Specific Evaluations and Syntheses.

2. Program Data

In addition to information obtained during telephone calls and visits to programs and from the review of program documents, the implementation analysis will conduct descriptive analysis of the following types of data: (1) patient-level enrollment and disenrollment records, (2) Medicare eligibility and claims data for participants and eligible nonparticipants, (3) intake data for random assignment and identification of survey respondents, (4) data on the use of non-Medicare services (if any such services are included as part of the intervention), and (5) program-level cost data. We anticipate that these data will be available from HCFA or from the implementation contractor (with the exception of the intake data, which the programs will send directly to MPR). Enrollment and disenrollment records will be a key component of our examination of program participation and costs, as will Medicare data, which we will use to compare program participants with eligible nonparticipants. (Voluntary disenrollment rates are also indicators of patient satisfaction with the program.) We expect that we will have to rely on program staff to inform us in the aggregate of reasons for participation and nonparticipation and for dropping out of the program.[5] Data on the use of non-Medicare services will enhance our assessment of how well a program that offers these services increased their use. The program cost data, in addition to being a key component of the evaluation's cost-effectiveness analysis, should provide information about the proportions of different costs (for example, the cost of care coordinators versus administrators versus other direct costs, such as rent and travel). The Site-Specific Analysis Plans will include schedules for collecting these data from each program, the implementation contractor, or HCFA, and for then processing them for the site-specific reports.

[5] We will also ask programs to conduct exit interviews with disenrollees or to complete forms describing reasons for enrollment. Programs may find this information useful for their own purposes; at this point, however, we have no idea whether they will be amenable to collecting it.

III. DESIGN OF THE IMPACT ANALYSIS

Estimating the impacts of the demonstration programs will require a rigorous research design, data from several sources on the outcomes the program is expected to influence, and strong statistical models to provide unbiased and efficient estimates of program impacts. Our need to obtain separate impact estimates for each site, as well as the considerable variation of many factors across sites, including the intervention, intake procedures, data available, time frame, sample size, and potential for contamination, increases the difficulty of this task.

The research design for the impact analysis will be described in three stages. In this report, we describe the general approach that we will take and how it may have to be modified for individual sites. After completing this report, we will prepare a memorandum for each site assessing the potential problems the program design and proposed comparison strategy pose for the evaluation of that site. The memorandum will examine the potential for contamination of the control group, possible enrollment problems, randomization issues, data collection issues, and any other aspect of the demonstration site's program that could threaten the validity of the evaluation. Once the design issues have been resolved for each site, we will prepare site-specific analysis plans that will describe in detail each of these design issues. The plans will be completed in March through May of this year.

A.   RESEARCH DESIGN

The two key features for ensuring that valid estimates of program impacts are obtained are (1) the comparison group strategy (that is, how we select the group to estimate what would have happened to demonstration participants in the absence of the intervention), and (2) the sample size. Program impacts will be estimated by comparing outcomes for the comparison group with outcomes of the demonstration participants. The comparison group will be carefully selected to yield unbiased estimates of program impacts. Having adequate sample sizes will ensure that the probability of type 1 and type 2 errors is small enough that impacts of policy-relevant size will not go undetected.[1]

1. Experimental Design for the Impact Analysis

Fourteen of the 15 demonstration sites selected for the Medicare Coordinated Care demonstration and the 2 disease management demonstration sites (operated by Lovelace) propose to use random assignment, the preferred research design for the evaluation. We can obtain unbiased estimates of program impacts with a known degree of statistical precision by randomly assigning each beneficiary who meets all the eligibility requirements and is interested in participating to either the treatment group or the control group. Features of the demonstration program can distort the estimates and introduce biases, and care must be taken in generalizing the findings to the population of interest. Nonetheless, an "experimental design" featuring random assignment is a much stronger research design for assessing the impact of demonstration programs than are any comparison strategies that can be devised.

One of the sites did not propose to conduct random assignment. We will attempt to develop a feasible and acceptable randomized design for that site, but if we cannot do so, we will select a comparison group. We will also select comparison groups for sites that proposed random assignment if our assessment of their research designs suggests that random assignment is likely to yield distorted estimates due to contamination of the control group. Contamination can occur if the

[1] Type 1 errors involve incorrectly concluding that a program has impacts when it does not (in other words, incorrectly rejecting the null hypothesis of no impacts). Type 2 errors arise from incorrectly *failing* to reject the null hypothesis when it is false, thereby failing to conclude that an effective program has favorable impacts.

intervention cannot be confined to those in the treatment group and therefore affects outcomes for control group members. Contamination can occur, for example, if the intervention influences the behavior of providers serving both treatment and control group patients.

To assess how well our estimates measure program impacts for the sites in which random assignment is not possible, we will use a similar approach to select comparison groups for the sites in which random assignment *does* take place and will compare the estimates obtained from the two approaches. Although there are a number of potential problems with this analysis, it provides a unique opportunity to assess the validity of the comparison group strategy we might have to use, as well as the size of any likely biases. If a valid comparison group strategy can be identified, it would also provide a basis for ongoing monitoring of program impacts.

The random assignment and comparison group strategies are described in some detail in the following sections. We also identify the issues to be elaborated on later, in the site-specific analysis plans.

a. Random Assignment Sites

We strongly prefer that random assignment be conducted by MPR, rather than by the sites. Sites will identify beneficiaries with the target conditions (either through outreach or referrals) and will assess their eligibility for and interest in the program. Some sites will also require that the beneficiary's primary care physician consent to his or her participation in the demonstration.

After eligibility and interest have been assessed, the site will obtain the beneficiary's signature on a patient consent form that will describe the terms of participation in the demonstration and the way the randomization process works. The consent form will include the beneficiary's name, Medicare health insurance claim (HIC) number, date of birth, gender, telephone number, and address, as well as the telephone number of someone else who will know how to reach the beneficiary. The consent form will explain that, in order to be eligible for enrollment in the program, the beneficiary must agree to be interviewed by MPR six months after enrollment, and must allow the data collected to be used for this evaluation. The form will explain that each beneficiary will be randomly assigned to either the treatment group or the control group, as well as the implications of the assignment for the services he or she will be eligible to receive. The form will also explain that all data about the beneficiary that MPR collects will remain confidential, with his or her identity masked from everyone except the interviewers who must administer the survey and the researchers who will use identifiers to link survey data and Medicare claims data about the beneficiary that are required for the evaluation. All beneficiaries will be required to sign the form, indicating that they understand and agree to the terms and conditions of participation. Program sites may have IRBs that will have to approve these procedures.

After the consent form is signed, the site will fax it to MPR, whose randomization staff will first confirm that the beneficiary has not been previously enrolled and then obtain the computer-generated random assignment of the case. The random assignment procedures will be set up using "strings" of all possible orderings of (say) six assignments, three of which are treatment group cases and three of which are controls. This approach guarantees that no more than a certain number of consecutive cases (six, in this case) can be assigned the same status, to avoid discouraging outreach staff or the appearance of favoritism. MPR will randomize cases within a few hours of receiving a fully completed consent form, so as not to delay initiating the intervention.

Because the consent form may be completed while some of the eligibility criteria are being assessed, it may contain additional information that would be valuable for the evaluation. For example, a program targeting a particular disease may serve only patients whose disease stage exceeds a certain level of severity (such as class II or higher on the NY Heart Association's heart disease scale). This information must be assessed from the patient's record and should be noted on the consent form. All relevant consent-form data will be data entered by MPR.

Three issues raise potential concerns about the randomization process. First, sites may want to conduct the randomization themselves. We will discourage them from doing so, to ensure that proper procedures are used consistently across sites, and to enhance the face validity of the whole evaluation. There are many ways that the randomization process can be subverted, either intentionally by well-meaning staff who want to be sure certain patients receive the intervention, or accidentally, as a result of inadequate quality controls and procedures. In our experience, sites often prefer that the research firm bear responsibility for the randomization, to protect the site operators from claims of favoritism. However, we are sensitive to programs' potential need for rapid assignments, including on weekends, and will work with them to develop an approach that addresses both methodologic and operational concerns.

A second concern is that it may be necessary to stratify the sample (that is, separately randomize cases within each stratum) to ensure that the treatment and control groups are well-matched on particularly important criteria. Although randomization should generate approximately equal numbers of treatment and control group cases in each stratum, formal stratification would ensure it. For example, if heart disease stage is a strong predictor of outcomes for CHF patients, comparing outcomes for the treatment and control groups could yield distorted estimates of impacts in a particular sample if, by chance, the distribution of disease stage differs markedly for the two groups. However, stratification complicates the assignment process substantially.

We believe stratification is unnecessary, given the target sample sizes and the proposed approach of using regression analysis (described below) rather than simple comparison of means to estimate program impacts. With the minimum sample size of 309 completed interviews per group (618 total) specified for each site in this demonstration, the groups are unlikely to differ substantially on any key characteristic. The regression models should adequately control for any differences. Thus, we suggest that no stratification be performed in general but will consider this issue for each of the demonstration sites.

A third concern is that some sites (for example, rural sites) may not be able to enroll enough cases to meet the sample size targets for a randomized design. In that case, it might be preferable to use a comparison group design. Suppose, for example, that 600 beneficiaries are eligible for a demonstration program, but only half (300) are interested in participating. In that case, rather than randomly assign 150 of the interested beneficiaries to each group, a better design might include all willing participants in the treatment group and use a comparison group design. These choices involve complex tradeoffs, however, so it will be important for MPR and each site to realistically assess the number of cases in the target population and the expected participation rates. It will also be important to discuss with HCFA alternative arrangements under which intake could be allowed over longer than the specified 12-month time frame, in order to reach the minimum sample size and retain the highly desirable randomized design. These discussions must also take into account the adverse implications that extending the intake period would have for the evaluation schedule and budget.

b. Comparison Group Approach

For sites in which a comparison group approach is deemed necessary or preferable, it will be necessary to determine (1) the geographic area from which to draw the comparison group, (2) the method for identifying beneficiaries for the comparison group, and (3) the methodology for estimating program impacts. We discuss each of these issues here.

Selecting Comparison Sites. It is well known that practice patterns differ widely across geographic areas. Thus, in selecting a comparison area for a particular demonstration site, we will first determine whether we can identify a comparison area within the geographic area in which the demonstration site operates. This may entail defining the comparison site as (say) other hospitals within the same metropolitan area, for a program that intends to draw all or virtually all of its patients from a particular set of hospitals. We could use a similar approach for programs that draw patients from a particular medical group or set of providers. For demonstration programs that draw their caseloads from certain counties within the metropolitan area, we would first examine other counties within the metropolitan area as a source for comparison group members. In some cases, it may be necessary to select cases from a different metropolitan area entirely.

One approach that we do *not* intend to use is to draw the comparison group from the set of individuals who were invited to participate but who declined (or whose physician declined). This group of patients is likely to be systematically different from participants on observed and unobserved characteristics likely to be highly related to outcomes of interest and would therefore lead to biased estimates of program effects. Although we would be able to control for measured characteristics on which the two groups differed in the analysis, differences on unobserved factors, such as willingness and ability to comply with recommended medication regimens, diets, and behaviors, are likely to be important and cannot be controlled for statistically.[2]

Regardless of whether the comparison "site" is a set of hospitals, a group of counties, or a separate metropolitan area, we will first identify a set of the most obvious candidates, in consultation with the demonstration site. We will then compare the predemonstration Medicare service use of all beneficiaries who met the program eligibility criteria in the demonstration and

[2] Econometric procedures to control for such "selection bias" (see, for example, Heckman 1976) often produce unreliable and imprecise estimates and therefore require larger sample sizes to yield the same level of precision.

potential comparison sites in the predemonsration period.[3] To be considered as a valid comparison site area, the area's target population must have predemonstration service use patterns that are similar to the predemonstration service use patterns in the demonstration site. Statistical tests will be used to ensure that the demonstration and comparison areas do not differ on predemonstration service use measures. Key outcome measures used in these comparisons will be the proportion of cases with hospital admissions, number of hospital admissions and days, proportion with skilled nursing facility (SNF) admissions and SNF days, proportion with home health visits and number of visits, and death rates. These outcomes will be measured over a period of time after the date when enrollment would be assumed to have taken place (such as after a hospital stay), based on the program's proposed targeting approach.[4]

Ideally, we will have multiple suitable comparison sites for each demonstration site requiring a comparison group. If so, we will select the two best choices. The comparison site that provides the closest predemonstration match to the demonstration site will be the "designated" comparison site; the other site will be the "alternate" comparison site. The survey sample will be drawn from the designated site. Both sites will be used to compare outcomes drawn from Medicare claims, as a test of the robustness of the estimates.

[3] The eligibility criteria applied in these calculations will be limited to those that can be ascertained from Medicare claims and enrollment files (for example, a hospital admission for a specific diagnosis, age, place of residence, not terminally ill, no claims for certain excluded comorbidities, and so on). The amount of error in these eligibility determinations will vary across sites, depending on the number of criteria that are not observable from (or approximated by) claims data.

[4] We will not rely on Medicare costs for these comparisons. Costs could differ even if service use patterns were very similar, due to differences in the unit cost of services. Unit cost differences can be adjusted for in our analysis of impacts on costs.

Selecting Comparison Group Cases. After the comparison sites have been chosen, we will select the comparison group cases. The task will require us to match the set of individuals who enroll as nearly as is possible and in a timely enough way so that they can be interviewed approximately six months after they would have been enrolled had they been in the demonstration site and were willing to participate. It will likely be somewhat difficult to select cases, because sites may take one year or longer to enroll their patients. This time frame creates difficulties, because we will be interviewing treatment group members six months after enrollment and have to interview comparison group cases at a comparable point in time.

We will adapt our approach based on how the site identifies potential participants. In sites that enroll patients gradually over the year, it will be necessary each month to obtain Medicare claims data on beneficiaries who enroll in the demonstration that month, and use these data to create a profile of prior service use and comorbidities for this cohort of enrollees. We will then identify the eligible (according to claims data) cases from the comparison sites, extract their claims data for the prior year, and select those that best match the enrollees on prior service use and Medicare variables. If claims data from Medicare files must be used to identify the cases, the data will not be reasonably complete until about three months after service use occurs, so the time frame available to select comparison group members before they must be surveyed is fairly tight. (See Section B of this chapter for a detailed discussion of the timing.)

We will not be able to use the increasingly popular approach of propensity scoring to draw the comparison group for the survey sample. Propensity scoring has been used effectively to replicate random assignment results in one study (Dehija and Wahba 1999); however, it requires estimating a logit model on the treatment group and a pool of potential candidates for the comparison group. The estimated model is then used to generate propensity scores for all cases, and the comparison site cases that best match the treatment group on propensity score are selected as the comparison sample. This approach, which yields a comparison group that is well matched on observable characteristics, is feasible to use when the timing of sample selection is not crucial but would be impractical in cases such as this evaluation. Lags in the claims data further exacerbate the difficulty of gathering data on comparison site cases, estimating a propensity scoring model, and selecting comparison cases, all within six months after enrollment. Given the small number of cases expected each month, the need to select cases on a bimonthly basis makes this approach impractical and costly.[5]

We defer further discussion of how we would select the comparison group to the site-specific analysis plans. Only when the details of the demonstration site's targeting become clear will it be possible to specify precisely what options are available for selecting a comparison sample for the survey. In most sites, enrollment will include a mixture of beneficiaries from various referral sources. Further details are provided in Section III.B.

Estimating Program Impacts with a Comparison Group Design. Use of a comparison group design will change how program impacts are estimated. Estimates are relatively straightforward to produce for studies with an experimental design (see Section D of this chapter); however, in sites in which a comparison group strategy is used, it will be necessary to account for the possibility that the comparison group may differ systematically from the treatment group on preenrollment characteristics. The likelihood and extent of these biases will depend on the eligibility and targeting criteria of the site and on how well eligibility can be modeled with claims data. However, some common issues must be addressed in any case.

[5] However, we will conduct some sensitivity tests using the propensity score approach on claims-based outcomes measures.

We will use three basic approaches (and a few variants) when estimating impacts in sites in which a comparison group approach is necessary. One approach is simply to assume that the match, however it is done, yields a comparison group similar enough to enable a simple regression model to control for any inherent differences between the two groups. However, that naïve approach is likely to yield biased estimates if the program imposes additional eligibility restrictions that are not observable from claims data, or if those who choose to enroll would have had different outcomes from other eligibles even in the absence of the intervention. The second approach is to compare (regression-adjusted) outcomes for *all* individuals in the demonstration site area who meet the claims-based eligibility criteria, regardless of whether they enrolled or not, with outcomes for a sample of eligibles who did not have the opportunity to participate because they lived in the comparison area. Dividing this estimated difference by the proportion of the eligibles enrolled in the demonstration yields estimates of program impacts on participants (because all impacts must be concentrated in the participants if there is no provider-based contamination of nonparticipants). Finally, we will estimate econometric models to control for selection bias. (Selection bias occurs when unmeasured differences between the treatment and comparison groups affect outcomes and therefore produce biased estimates of program effects.)

Each of these three approaches requires a different survey sample design and selection of cases. Under the first approach, we need samples of participants and comparison cases. For the second approach, we need a sample of both participating and nonparticipating eligibles from the demonstration site, plus a sample of eligibles from the comparison site. The third approach requires only participants and nonparticipants from the demonstration site.

We will sample all three groups—participants, eligible nonparticipants, and comparison site cases—in order to use all three approaches to evaluate program sites where a comparison site approach is used. This strategy will enable us to test the robustness of the estimates. By selecting eligible nonparticipants and comparison site cases that match the enrollees as closely as possible on prior service use and comorbidities, we hope to both minimize the preexisting differences among the three groups and increase the precision of our estimates. The sample sizes for the three groups are given in Section A.2. Section D describes the statistical models in greater detail.

c. Selection of External Comparison Groups for Sites with Random Assignment

Although we will have random assignment in the great majority of the sites, we will also select an external comparison group for each site in order to provide evidence on the validity of the estimates obtained for sites in which randomization was not possible. These comparison group samples will not be surveyed—only claims data will be available. For all outcome measures drawn from the claims data, we will estimate impacts using the comparison design approach. These estimates will be obtained by estimating the differences between eligibles in the treatment area (including actual treatment group cases) and eligibles in the comparison area and dividing this difference by the participation rate. We will then compare these estimates with the estimates from the randomized design.

A match in most sites between estimates from the randomized design and estimates from the comparison design will provide additional assurance that the comparison group strategy produced valid impact estimates in the sites in which random assignment was not conducted. However, substantial differences between the two sets of impact estimates might be due to either the imprecision of the comparison group estimates or violation of one or more of the following assumptions under which the comparison strategy will produce unbiased estimates:

- Only the treatment group's outcomes are affected by the demonstration.
- Outcomes for eligibles in the two areas would be equal, in the absence of the demonstration.

- The proportion of eligibles who would participate in the demonstration if it were offered in both areas is equivalent in the two areas.

The first assumption requires that no intervention effects "spill over" to the control group (also referred to as "contamination" of the control group), as that would bias the estimates from the randomized design toward zero. We will draw some inferences about the likelihood of spillover effects from our assessment of the demonstration designs and site visits, and from some empirical analyses (see Section III.D on estimation procedures). In general, we expect very little contamination in sites that conduct random assignment. (Sites with a strong likelihood of significant contamination will be requested to modify their demonstration design or to switch from randomization to a comparison group methodology). If this assumption is violated, we expect to observe larger impact estimates from the comparison group approach than from the randomized design.

The second assumption could be violated if practice patterns differ in the two market areas. We will guard against this possibility by selecting comparison areas for which the target population in the demonstration site and the comparison site have similar outcomes in the predemonstration period. Although practices in the two areas could change in somewhat different ways from year to year, any differences are likely to be moderate over such a short time span.

The third assumption requires that the two areas not differ markedly in (measured and unmeasured) beneficiary characteristics associated with participating in demonstration programs such as these. This assumption seems likely to be satisfied if the comparison areas are chosen carefully to match on observable socioeconomic variables and geography. Regression models used in estimation should control adequately for any remaining differences.

The major reason that impact estimates from the two methods are likely to differ is the large variance that accompanies the comparison site approach. Detecting an impact of 10 percentage points in the probability of hospital admission (when the mean is .50) with adequate precision requires samples of roughly 309 treatment group cases and 309 controls when using a randomized design. To obtain the same level of precision in estimating impacts with the comparison group approach (in which we compare eligibles with eligibles) requires samples that are $(1/p)^2$ times larger, where p is the proportion of eligibles who participate. For example, if the participation rate among eligibles were 30 percent, obtaining a comparable level of precision would require 3,433 eligibles (1,030 participants and 2,403 nonparticipating eligibles) and 3,433 comparison site cases.[6]

Our objective in selecting comparison cases for these analyses will therefore be to define the eligible population from which we will draw the sample in such a way as to maximize the participation rate. To do so, we may have to define the target population for the purposes of this estimation in a manner that excludes a portion of the actual target population. For example, suppose a program enrolls 80 percent of its participants from those who are hospitalized for CHF in a particular hospital, but also enrolls CHF patients referred to the program from other area hospitals or physicians. Suppose that the hospital-affiliated physicians are very supportive of the program, so the participation rate within the core hospital is 70 percent, but that participation rate among other eligibles in the area is only 5 percent. In this case, it will be very difficult to

[6] When comparing eligibles in the demonstration site with eligibles in the comparison site, the impact on *participants* is equal to $p*b$, where b is the regression-adjusted difference in outcomes between eligibles in the two sites. So, to detect an effect of (say) 10 percentage points on participants, we must be able to detect a difference between *eligibles* of $.10p$. The sample size required to detect a given detectable difference is inversely proportional to the square of the detectable difference. Thus, detecting a difference of $.10p$ requires a sample size that is $(1/p^2)$ times larger than the sample needed to detect a difference of .10.

estimate program impacts reliably for the full target population, whose overall participation rate is about 20 percent. However, it should be possible to estimate impacts reasonably well for the subset of the sample that was drawn from the hospital, by identifying another hospital or group of hospitals in the area whose patients had similar outcomes in the year preceding the demonstration startup.[7]

The method for selecting the comparison sample for the random assignment sites will differ somewhat from the method used to select a comparison group in sites without random assignment. The method of selecting the comparison *site* will be the same, but we will select *all* eligible beneficiaries in both the demonstration site and the comparison site. We will be able to use this approach because we use only claims data (no survey data) for this segment of the analysis. Reliance on only the claims data also means we do not need to identify the comparison cases for the random assignment sites monthly over the demonstration site's intake period; rather, we will make the identification only after the target sample size has been reached in the corresponding demonstration site. Even though we will have all the eligibles in each site, we will also select a subset of these cases, using a sampling approach that essentially replicates the process required to select comparison group cases in the demonstration sites that do not have

[7] The goal will be to define "eligibility" from the claims data in ways that will maximize the participation rate among beneficiaries meeting the eligibility criteria, without excluding too many actual participations under this definition. That is, we want to define eligibility to maximize *A/(A+B)* while keeping *A/(A+C)* as close to 1.0 as possible.

	Meets "Eligibility" Criteria	
	Yes	No
Enrolled	*A*	*C*
Nonenrolled	*B*	*D*

random assignment. The sample size will match the number of comparison cases drawn in those sites.

Both the full population of eligibles and the sample will be used to generate impact estimates for comparison to the impact estimates from the randomized design, to illustrate the importance of sample size for the comparison group approach. Our expectation is that the estimates obtained from comparing the full population of eligibles in the demonstration and comparison areas will more closely match the impact estimates from random assignment than the estimates obtained from the smaller survey samples of eligibles. We will also use the propensity score approach and selection-bias-correction models to estimate program effects with the comparison group and will contrast the estimates with the impact estimates from the randomized design.[8]

d. Site-Specific Analysis Plans

Our proposed approach for developing the samples, while providing some generic guidelines, will be modified for individual demonstration sites, based on the particular characteristics of the sites. The size and composition of the target population, additional eligibility criteria, projected sample size and flow, expected participation rate, intake procedures, and distribution of cases across referral sources will be important factors in defining the most appropriate research design for each site.

Before preparing the analysis plans, we will conduct a site-specific assessment of the research design. That assessment will identify the likelihood of reaching the target sample sizes

[8] The propensity score approach will be feasible for this analysis, because the comparison groups for the random assignment sites will be chosen at one time, and only claims data are required for the analysis.

within the available time frame, ambiguities in the definition of the target population or recruitment strategy, inconsistencies between previous years' Medicare data and the site's proposed estimates of the size or average costs of the target population, possible sources of contamination of the control group, and any other factor that may have consequences for the evaluation. This assessment is expected to require a substantial amount of interaction with the sites to resolve questions about their designs. (We have completed this assessment for the Georgetown program and have identified a number of issues that will need to be resolved.) We will also calculate waiver cost estimates for each site, which will provide the estimates of sizes and anticipated average cost of the target population in the absence of the demonstration.

After the site-specific assessments of the research designs have been completed, we will prepare site-specific analysis plans. These plans will provide the details of how random assignment (or comparison group selection) will be implemented, the time frame for intake, how sampling will be done (if needed), and how the comparison "site" for each demonstration site will be selected. They will identify the most logical comparison sites to be investigated, the criteria to be used to assess the suitability of alternative sites, the definition of eligibility to be used in identifying the target population in the demonstration and comparison areas, and how the comparison sample will be drawn from the population of eligibles.

In addition to these issues related to the target population, randomized design, and comparison group selection, the site-specific analysis plans will identify any data that might be used for the site in question but not for other sites. These data could include claims data for measuring outcomes or control variables that are relevant for that site but not for others. They may also include data from patient intake forms that would be useful as control variables. The site may even have some data on outcome measures that are not available from claims data. However, because collecting this information on the control group raises concerns about contamination, we will generally discourage sites from collecting data on control group members.

An example might clarify the data issues that will be addressed in the site-specific analysis plans. Suppose a program is targeted at diabetic beneficiaries. In this case, we might include as outcome measures whether a beneficiary has Medicare claims for the recommended examinations for glaucoma and periodic blood tests or for other indicators that the patient is receiving certain types of preventive care or tests. We may also use as control variables whether the beneficiary received those tests in the year preceding enrollment and whether the beneficiary had any hospitalizations with diabetes as the primary diagnosis, or how long prior to enrollment the most recent such hospitalization occurred. The intake form may provide information on the stage of the patient's diabetes and other factors relevant to the severity of his or her condition at enrollment. The site-specific analysis plans will investigate all these opportunities for enhancing the evaluation of impacts at that site, and for making the evaluation as sensitive as possible to key issues for the target population.

2. Sample Sizes

The minimum number of cases that demonstration sites are required to enroll in the study was dictated by the Request for Proposals for the demonstration. This recommendation was based on the research design prepared by MPR for HCFA (Brown 2000). In practice, some of the demonstration sites selected by HCFA have proposed larger sample sizes, raising the possibility of having greater precision in these sites. Furthermore, in order to achieve the target sample sizes for analyses based on survey data, a larger number of cases will have to be enrolled (to account for survey nonresponse). Here, we first discuss the statistical precision that will be obtainable for the analysis of survey-based outcomes and for claims-based outcomes in the sites with random assignment. We then discuss the sample sizes and precision for the impact estimates that will be obtained from the comparison group approach described in the previous section. The final subsection describes sample sizes for the physician survey.

a. Patient Sample Sizes for Sites with Random Assignment

The minimum sample size HCFA requires for the demonstration sites that are doing randomization (309 treatment group cases and 309 control cases) should be sufficient to ensure that most policy-relevant impacts will be detected. This minimum sample size yields 80 percent power of detecting an impact of 10 percentage points on a binary variable with a mean of .50, using one-tailed *t*-tests at the .05 significance level. Half or more of Medicare beneficiaries with CHF, chronic obstructive pulmonary disease (COPD), and a number of other chronic illnesses are hospitalized at least once during a 12-month period (Schore et al. 1997). Thus, this standard implies that sizable impacts (20 percent or larger) on the probability of a hospital admission are likely to result in statistically significant treatment-control group differences. Furthermore, the coefficient of variation for number of hospitalizations is approximately the same (1.0) as for the probability of a hospitalization, so the sample should also be adequate to detect a 20 percent reduction in number of hospitalizations.

A sample of the minimum size will have less power to detect impacts smaller than 10 percentage points (a 20 percent reduction from a mean of .50). Therefore, some programs, especially relatively less expensive interventions, that do not need such large effects on hospitalizations to generate net savings for the Medicare program may not exhibit statistically significant treatment-control differences in hospital admissions in the sample. However, detecting smaller effects would require substantially larger samples and would be markedly more expensive. For example, to have the same 80 percent power to detect a 10 percent (five percentage point) reduction in hospitalization would require a sample four times larger.

Although there is some risk that these small samples will fail to detest some cost-effective interventions, our review of the literature and the proposals suggests that many programs have achieved reductions this large or greater. Thus, the minimum sample size should be adequate.

The precision of estimates will also be less for survey-based outcome measures than for claims-based outcomes, due to survey nonresponse. If only the minimum number of 618 beneficiaries enroll in the demonstration, the minimum detectable difference for binary outcomes obtained from the survey, with a mean of .50, will be 10.5 percentage points, assuming a 90 percent response rate. To ensure that a difference of 10 percentage points will be detectable, programs will need to enroll a total of 686 individuals, 343 each in the treatment and control groups.

We refer to the number of cases included in the study as the "sample size." In fact, however, these cases will not be a sample but rather, the entire population of study participants in many of the demonstration sites. Six of the 15 demonstration sites propose to enroll substantially more than the minimum number of cases. In the case of sites intending to enroll approximately the minimum number of cases, we will survey all enrolled cases. Although surveying all enrolled cases implies that there is no sampling error and no need for statistical tests, we will conduct the analysis as though the cases available for analysis were a random sample from a much larger population. Because this commonly used approach treats the observations as though they are a random sample of all beneficiaries who would enroll in the study if it were an ongoing program over many years or were replicated in other areas, it provides a basis for generalizing the results.

We have budgeted only 618 completed surveys per site for sites using random assignment, assuming a 90 percent response. Therefore, for the six sites that expect to enroll substantially more than the minimum number of cases, we will need to select a sample of the cases to be included in the survey. The six sites expect to enroll in the treatment group anywhere from 500 to 5,500 Medicare beneficiaries. Table III.1 shows the sample sizes these programs specified in their proposals.

One major sampling concern is that sites will not be able to achieve the targeted sample size. If the sites do not do so, drawing a sample of cases to be surveyed based on the program's false projections would leave the survey sample short of the required number of cases. Our experience in numerous other studies suggests that sites rarely enroll the expected number of cases; they often fall far short of those totals. Especially germane to this evaluation was the Medicare Case Management demonstration, in which all three sites fell far short of the targeted number of cases (although one of the sites was eventually able to reach the target number).

One way to address this problem would be to simply survey all cases enrolled in the study until 309 surveys are completed for the treatment group, and 309 for the control group. However, this approach may yield an unrepresentative sample of cases enrolled over the first year of the study, and impacts might well be lower for that cohort than for those enrolled after the program has acquired some experience.

We will be able to balance these two concerns to some extent because we will observe enrollment for the first six months before having to conduct the first interview. At the end of the first five months of operation, for each of the six sites, we will check whether the site has enrolled beneficiaries at the expected rate and appears to have a steady flow of enrollees each month. If so, we will select a random sample from each month's enrollees that is sufficient to generate the desired survey sample size over the 12-month intake period. If enrollment has lagged behind expectations, we will calculate the sampling rate required to generate the target survey sample size and sample at that rate, if we are confident that enrollment will continue at that pace for the remainder of the year. If the calculations suggest a very high sampling rate is

TABLE III.1

SAMPLE SIZES AND DETECTIBLE DIFFERENCES

	Target Number	Minimum Detectable	Minimum Detectable
	of Treatment	Difference	Differences
Samples	Group Cases[a]	(Binary, p = 0.5)	(cost, CV = 2.5)
Survey-Based Outcomes	309	.10 (20%)	50%
Claims-Based Outcomes			
Typical Site[b]	309-350	.10 (20%)	50%
Larger Sites			
CenVa Net	500	.079 (16%)	39%
Core Solutions Medical	947	.064 (13%)	32%
Carle	1,198	.051 (10%)	25%
Quality Oncology	1,908	.040 (8%)	20%
Medical Care Development[c]	4,370	.027 (5%)	13%
Washington University	5,500	.024 (5%)	12%

NOTE: The minimum detectable difference presented is the difference between the two populations being compared for which we have 80 percent power when conducting one-tailed tests at the .05 significance level, using the sample sizes specified. That is, in 100 trials with random samples of this size, we would expect to find a statistically significant difference between the two groups 80 times only if the true effect on a binary variable with a mean of .50 were .10 or larger.

$MDD = 2.487 s \sqrt{1/n_T + 1/n_c}$, where s is the standard deviation of the outcome measure and n_T and n_C are the sample sizes for the two groups. In percentage terms, $MDD = 2.487 CV \sqrt{1/n_T + 1/n_c}$, where CV is the coefficient of variation (standard deviation divided by the mean) of the outcome variable.

[a] Core Solutions plans to randomly assign 947 beneficiaries to the treatment group and 615 to the control group. All other sites plan to assign an equal number of patients to the two groups.

[b] For these sites, we assume a sample size of 309 per group. The difference in precision between samples of 309 versus 350 is small.

[c] Medical Care Development is the only site that intends to pursue a comparison group design. The calculations here assume that the simple regression model will eliminate any biases, so the detectable difference is the same as from a randomized design. Actual precision is likely to be substantially smaller for this site.

required (for example, 90 percent or more), we will interview all enrollees until the target sample size is reached. This approach simplifies the survey process and provides protection against the possibility of enrollment falling below expectations during the remaining months of the year.

We will continue to monitor enrollment levels each month throughout the remainder of the year. If enrollment begins to drop, it may be necessary to increase the sampling rate for those enrolling in the later months of the intake period. In this case, we will have to weight the data by the inverse of the sampling rate when conducting the analysis, in order to appropriately represent the first-year enrollees.

In order to reach the target sample sizes of completed survey interviews (309 per group), we will have to select approximately 343 enrollees in each group for the survey sample. This calculation assumes that 90 percent of the enrollees selected for interview (or their proxies) will complete the six-month interview. We should be able to achieve this high rate of completion because (1) we will have excellent contact information from the enrollment/consent form, (2) beneficiaries will have agreed to be interviewed as a condition of having a chance to receive the intervention, and (3) the enrollees are not likely to be highly mobile. However, given their chronic conditions, a substantial fraction are likely to die over the six-month period (an average of 1.5 percent per month is expected for this population). We will attempt to interview proxy respondents for sample members who die (provided the members survive for at least three months after enrolling), but the response rate is likely to be lower for this group. Refusals are expected to be rare.

One option that we will explore with the sites and with HCFA is allowing the sites to assign 60 percent of eligible applicants to the treatment group and 40 percent to the control group. Although a 50:50 split of the eligible applicants yields the greatest precision for a given total sample size, a 60:40 split yields estimates that are nearly as precise while enabling a site to serve a higher proportion of the interested, eligible applicants. The level of precision achieved with 618 total cases (309 cases in each group) can be replicated by enrolling 644 beneficiaries in the study (386 treatment group cases and 258 controls). This split may appeal to the program staff, who understandably dislike telling people recruited for the study that they are not going to receive the intervention. This design may be worthwhile because it does not increase recruitment goals substantially. If sites choose to have a 60:40 assignment ratio, the target number of cases to be selected for the survey sample, after accounting for the 90 percent response rate, increases from 687 to 716 (429 treatments and 287 controls). Our current budget should be adequate to expand the survey sample size by this modest amount because it assumes that several of the sites would not have random assignment and would therefore require larger survey samples (see below). However, the current budget also covers evaluation of only nine of the demonstration sites (plus the two Lovelace sites).

We do not expect much survey nonresponse; therefore, little bias is likely to be present in our estimates from the survey data. Nonetheless, response rates for the treatment and control groups could differ if control group members are disgruntled about their group assignment, which could introduce a bias. The claims data, which are available for the entire sample, will be used to assess how the impact estimates for claims-based outcome measures calculated on the responding sample differ from those calculated on the full sample.

We will use all the observations when conducting the claims-based analysis.[9] As seen from Table III.1, in the six sites that intend to enroll more than the minimum number of cases, we will be able to detect substantially smaller impacts on claims-based outcomes than what will be

[9] However, we will not be able to use control variables obtained from the survey for the regression analysis that generates the impacts.

detectable in the other sites. The detectable effect of 10 percentage points on a binary variable with a mean of .50 (for example, the probability of hospitalization) drops to about 8 percentage points when the sample is increased to 500 cases in each group. It drops substantially more for the sites expecting to enroll several times the minimum standard.

The impact estimates are likely to be much less precise for estimating effects on Medicare costs. The standard deviation of Medicare costs is about 2.5 times the mean (CV = 2.5), whereas for a binary variable with a mean of .50, the standard deviation is equal to the mean (CV = 1.0). As Table III.1 shows, with the minimum sample size, we can be confident of observing a statistically significant effect on cost only if the true impact of the program is to cut median costs in half. We can be confident of detecting even a 20 percent true reduction in total Medicare cost only when the sample size rises to nearly 2,000 cases (as it does for Quality Oncology, in the table). Even though large reductions in hospitalizations are highly likely to reduce total Medicare costs, it seems likely that we will observe inconsistencies between the statistical significance of effects on hospitalizations, for which the coefficient of variation is about 1.0, and that of effects on costs. Section III.C.2 describes an alternative way to assess the effects on costs that may help explain the inconsistencies between effects on hospitalizations and the effects on costs that we are likely to observe in the sites with small sample sizes.

b. Patient Sample Sizes for the Comparison Site Approach

We had budgeted twice the sample size for the survey for the sites requiring a comparison group approach as we had for the sites with random assignment. The sample in the demonstration site is to be split evenly between participants and eligible nonparticipants. Thus, for a site with the minimum sample size, we still will have survey data on 309 participants, along with 309 eligible nonparticipants in the demonstration area and 618 cases in a comparison area.

To obtain the desired sample sizes of 309 completed interviews of program participants, programs will need to enroll at least 343 individuals (assuming a 90 percent response rate). For eligible nonparticipants and comparison site members, we expect a response rate of only about 70 percent, because we will not have an intake from the beneficiary's phone number and address. Thus, to obtain 309 completed interviews with eligible nonparticipants in the demonstration site we will select a sample of 441 cases. To obtain 618 interviews with comparison site beneficiaries, we will draw a sample of 883 cases.

If impacts are estimated by comparing eligibles in the demonstration area with eligibles in the comparison area, this sample size would not be sufficient to yield impact estimates as precise as those obtained from the randomized design unless about 70 percent of demonstration site eligibles actually enroll. Table III.2 shows the sample sizes that would be necessary for this level of precision under various participation rate assumptions. As the table shows, the sample sizes are prohibitively large unless the participation rate is about 50 percent. The budgeted sample size is what was considered feasible.

Fortunately, only one site proposes to use a comparison group design, and it has set its target enrollment at 4,370. Although this large sample size would be adequate to generate fairly precise estimates of impacts on claims-based outcomes, the precision of impact estimates for survey-based outcome measures will be limited by the much smaller survey sample. Furthermore, this target enrollment level seems wildly optimistic for this demonstration's rural Maine setting, so the estimated effects on claims-based measures may also be less precise for this site than for sites with randomization. Our site-specific analysis plan for the Maine site will explore the situation in more detail.

Our planned approach of using comparison groups to obtain a separate impact estimate in the sites that do use random assignment is likely to illustrate the difficulty of obtaining a reliable

TABLE III.2

SAMPLE SIZE NEEDED TO MATCH PRECISION OF RANDOMIZED DESIGN

Participation Rate	Treatment Group	Eligible Nonparticipants	Comparison Group
.1	3,090	27,810	30,900
.2	1,545	6,180	7,725
.3	1,030	2,403	3,433
.5	618	618	1,236
.7	441	189	631
1.0	309	0	309

NOTE: This table provides estimates of the sample sizes needed to yield impact estimates from a comparison group design that are as accurate as those from a random assignment design with 309 cases each in the treatment and control groups. The estimates are based on the assumption that the method used to estimate impacts is to compare mean outcomes for all eligibles (participants and nonparticipants) in the demonstration program’s service area with the means for eligibles in the comparison area, and to divide the difference by the participation rate observed in the demonstration area. Other methods of estimating impacts would require different assumptions and sample sizes.

impact estimate without random assignment. We expect that the impact estimates we obtain by comparing outcomes for all eligibles in the demonstration and comparison areas will be considerably closer to the impact estimates obtained from comparing treatment and control groups in the sites with the largest enrollments. However, that may not be the case; participation rates may vary widely, and some programs may attract a fairly random group of eligibles whereas others may attract primarily eligibles who would have better outcomes than other eligibles even in the absence of the demonstration. Discussions with site staff will help inform this assessment.

c. Physician Sample Sizes

We will interview a sample of physicians serving patients in each demonstration program to assess their satisfaction with the program, and to obtain information on any changes that they have made in their practices as a result of the program. We have budgeted interviews with 50 physicians in each site, but it is likely that a number of the sites will have fewer than 50 physicians who refer patients to the program. Other sites may have many physicians referring patients, but many physicians may have only a single patient who receives the intervention. If 50 physicians refer patients to the program, they would have to average referring about one eligible and willing patient per month in order for the site to enroll the minimum number of patients in the study over the course of one year.

B. DATA SOURCES

The impact analysis will use data from four sources: (1) a six-month patient survey, (2) a physician survey, (3) Medicare claims and enrollment data, and (4) site-specific data. We will administer a telephone survey of sample members six months after their enrollment. This survey will measure treatment and control group members’ well-being, satisfaction with care, health-related behaviors, adherence to medication, and knowledge of their condition. Information on providers' satisfaction with the intervention will be obtained by conducting a telephone survey of a sample of physicians fielded 9 and 22 months after each site starts enrollment. We will obtain service use and reimbursement data from Medicare claims files, and demographic and eligibility information from the Medicare Enrollment Database (EDB). County-level environmental descriptors will be taken from the Area Resource File (ARF). We will work closely with the implementation contractor and the demonstration sites to use any site-specific data that would enhance the evaluation, including intake information about a patient's severity of illness, cognitive ability, or demographics; any data the site tracks on the types and amount of care coordination services they provided to individual patients; and records of program costs.

1. Patient Survey

The patient survey will measure patient demographics, primary language, well-being, health status, satisfaction with care, health-related behaviors, adherence to medication, and knowledge of condition. The survey will be conducted by telephone and will be limited to about 20 minutes to minimize burden placed on patients.

a. Sample

In sites using random assignment, we will complete surveys for 309 treatments and 309 controls. In sites using a comparison design, we will complete surveys for 309 treatments, 309 eligible nonparticipants using the same network of providers, and 618 external comparisons who meet eligibility criteria but who use a different network of care.

We will begin the interview by ascertaining whether a patient is available to respond. If the patient has died less than three months after enrollment, we will not complete the interview with a proxy, as it is unlikely that the demonstration would have affected the outcomes of interest for the patient. However, if the patient died three or month months after enrollment, we will attempt to complete an interview with the proxy, typically a next of kin. Interviewing proxies of patients who were enrolled for at least three months limits recall to three months, at the most. It also ensures that the patient was enrolled in the demonstration for at least three months, which is enough time for the intervention to possibly have had an impact. We will remove questions about functioning or health status to limit the proxy survey to questions about the patient's experiences with the care coordination program. If a patient is hospitalized, the interviewer will ascertain when he or she will return home and will reschedule the interview accordingly. If the patient is seriously impaired (for example, in a coma) the interviewer will ask the proxy to complete the full survey.

b. Timing

The patient survey will be administered to each sample member six months after enrollment. Sample members will be told about the survey on their consent forms and will be reminded again by a letter sent two weeks before the interview. The six-month time frame for followup seems a reasonable compromise between (1) minimizing recall problems on measures of program participants' satisfaction with the program, and (2) ensuring that our estimates reflect only program effects that are reasonably long-lasting.

The survey must provide data for estimating both satisfaction with the program and program effects; the former pushes us in the direction of a shorter follow-up period, whereas the latter suggests a longer follow-up interval (two separate surveys would be substantially more expensive). It is important to obtain feedback on program satisfaction relatively soon after enrollment. We expect that the most intense period of clients' interaction with the programs will occur during the first few months after enrollment, even for interventions that are ongoing or that continue to intervene with the patient for a fairly long period of time.

Although patients' recollections may fade by month 6 after enrollment, influential programs are likely to have made a lasting favorable impression. Conversely, we would expect a successful program to have some effect on survey-based outcome measures within six months after enrollment. For example, if patients have not improved their self-care habits within the first six months after enrollment, they are less likely to ever do so in response to the intervention. Whereas a program impact on self-care may dissipate over time, we will at least know whether it existed at month 6—a nontrivial length of time after enrollment. Similarly, if the impact on health status and the other survey-based outcome measures does not exist six months after enrollment, they are not likely to be meaningful for patients. We will use a standard six-month followup for all sites to attain outcomes that are comparable across sites.

For sites that use a comparison design, we will need to carefully time the selection of the comparison group members so they can be surveyed at the appropriate time. Although only one site has proposed a comparison design, sites that have proposed random assignment may have to adopt a comparison design if we identify substantial contamination issues. To ensure comparability with sites using random assignment, we should collect survey data on comparison group members and on eligible nonparticipants from the demonstration area at similar time points. However, a challenge arises because there are no enrollment dates for the comparison group or for eligible nonparticipants, and demonstration program participants will be enrolled over the course of a year or longer.

To address this problem while preserving the desired six-month recall period, we will select the sample of eligible nonparticipants in the demonstration area and eligibles in the comparison areas in six waves. This approach would be operationalized as follows (all times are measured in months since program startup):

Wave	Sample Selected During Month	Based on Service Use During Months[10]	Interview Conducted During Months
1	5	1-2	7-8
2	8	3-4	9-10
3	10	5-6	11-12
4	12	7-8	13-14
5	14	9-10	15-16
6	16	11-12	17-18

This schedule allows three months for the service use that we need to identify the samples (primarily hospitalizations or treatments for specific diagnoses) to appear in the Medicare claims data.[11] Although not all claims will be posted by three months after the date of service, a sufficient number should be available to enable us to choose the sample. With this time period, the interview will be conducted close to six months after the service use that identifies the sample; thus, sample members in the comparison site(s) will be interviewed six months after

[10] The time period over which service use will be examined to identify sample members will depend on the procedures the program uses to recruit patients. The time period given in the table (service use during a recent two-month interval) is appropriate for identifying comparison sample cases when the demonstration site targets patients who have recently been admitted to the hospital with a particular diagnosis. However, for sites recruiting patients from multiple settings, the samples of eligibles must be defined to include all such cases in the same proportions as the participants. For example, if a program draws 40 percent of its patients being discharged from the hospital, and 60 percent from patients who were not recently in the hospital but who had one or more physician visits for the diagnosis in the past year, we will draw the samples of eligibles in the demonstration and comparison sites to match this distribution. Thus, for 60 percent of the cases to be selected, the service use criteria for the sample drawn in month 8 would be one or more physician visits in the period ranging from eight months prior to program startup to four months after startup (with no recent hospitalization).

[11] Standard Analytic Files (SAFs) containing cleaned claims data are posted in June/July for all data HCFA received from January through June. Data are then updated quarterly for the rest of the year. Unfortunately, using National Claims History Files (the uncleaned files used to construct the SAFs) does not reduce the lag time considerably relative to using the SAFs. This timing would make it difficult for us to choose the comparison group in “real time” using the approach we describe here. We will discuss with HCFA the feasibility of obtaining claims data on an accelerated schedule or of using an alternative approach.

enrollment, as they are in random assignment sites. We note that while drawing the samples of eligible nonparticipants and comparison group members in “real time” in the comparison site(s) is critical to completing timely surveys, we will be able to wait four months to obtain claims data for the impact analyses. Based on previous work, we expect that allowing an extra month for processing time will slightly increase the percentage of claims processed for the relevant dates of service.

In each cohort of eligibles, we would first exclude the program participants from the sample frame of eligible nonparticipants. Given our target sample sizes (for completed interviews) of approximately 309 participants, 309 nonparticipating eligibles, and 618 external comparison cases (for completed interviews) for each site using a comparison design, the allocation of sample over the six periods will be determined by the timing of participants’ enrollment. We will interview participants in each wave until the 309 total is achieved (unless the program expects to enroll many more participants than required in the first year, in which case sampling will be used, as described above).

A small number of the beneficiaries whom we survey as eligible nonparticipants may later enroll in the demonstration, thus becoming participants. Although we do not expect this to happen often, we will address it by treating these beneficiaries as eligible nonparticipants until the time that they enroll in the demonstration; from that point on, we will treat them as participants. We will not have full followup for these eligible nonparticipants (or for anyone who later joins managed care or who dies). The eligible nonparticipants we survey before they “cross over” and become participants will actually increase the precision of our impact estimates slightly, due to the positive covariance between their observation as an eligible nonparticipant and their observation as a participant.

c. Instrument Development

The patient survey instruments will contain a set of core questions that will be asked of all sample members regardless of diagnosis or condition, and a series of condition- or disease-specific modules to be administered only to sample members with those specific conditions. Whenever possible, we will draw questions for both the core and condition-specific modules from tested, valid, and reliable instruments. Section III.C provides our initial thoughts on how these variables will be measured. Here, we focus on the survey procedures we will use to collect the data.

The core questions will measure respondents' health and functional status; adherence; health-related quality of life; access to health care; satisfaction with health care; satisfaction with the care coordination program (for program participants); and receipt of certain elements of preventive care that are recommended for all seniors with chronic illnesses, such as immunizations and counseling on smoking. Information for control variables will include respondents' sociodemographic characteristics, comorbid conditions, attitudes toward seeking health care, preenrollment drinking and smoking patterns, and living arrangements.

We will develop modules for five of the specific conditions the demonstration sites plan to target: diabetes, CHF, other heart disease/stroke/vascular disease, cancer, and COPD. These conditions are some of the most common and costly that Medicare beneficiaries experience, and condition-specific instruments are readily available. The advantage of condition-specific modules is that outcomes particularly relevant to patients with a condition (for example, low blood sugar in diabetes or breathlessness in chronic lung disease) may be more responsive to the intervention than more general outcome measures. We also wish to include condition-specific questions on whether sample members received recommended care for that condition (for example, eye or foot exams for patients with diabetes), and whether sample members report desired behaviors (for example, compliance with medications or dietary recommendations).

If necessary, we will develop new items and will use cognitive testing to refine their wording. We will conduct the testing with Medicare beneficiaries with chronic diseases to determine whether they comprehend the questions. The survey director will conduct any required cognitive tests by telephone. We will use a mix of techniques, such as concurrent or retrospective think-aloud methods, to assess how respondents arrive at their answers. We will also test probes and memory cues to assess the effectiveness of techniques to reduce misunderstanding and response error.

We will develop the final survey instrument by taking the following steps. First, we will list as broad categories the patient-level outcomes and control variables we wish to capture and will review the list with HCFA staff. The list will include categories that are *not* available within claims data, such as disease- and non-disease-specific functional status and health-related quality of life, self-rated health, disease-specific behavioral risk factors, satisfaction with care, educational status, living arrangements, and income. Because the questionnaire could include more information than time will allow, we will work closely with HCFA staff to define the boundaries and priorities at this stage.

Second, we will group under each category questions and scales from existing instruments relevant to that category. Some categories will present us with more than one option for individual questions. For example, under the category of non-disease-specific functional status, there is more than one version of questions for restricted activity days or activities of daily living. Similarly, under the category of disease-specific quality of life, there is more than one instrument for quality of life in COPD. In these cases, we will favor options that meet as many of the following criteria as possible: used in large-scale, widely cited surveys; comparable to versions used in the care coordination literature; accepted among practitioners of care coordination; and suitable for CATI survey.

Third, we will make modifications that facilitate interviewing an elderly, chronically ill population or their proxies by telephone to enable people being interviewed to overcome communications, stamina, and cognitive challenges. One way to remove communications barriers is to reduce the use of high- frequency sound, as high-frequency hearing loss is common in the elderly. Certain high-frequency sounds (*s, z, t, f*, and *g*) are particularly difficult to hear over the telephone. Substituting words with low-frequency sounds will improve sample members' ability to self-respond reliably. To overcome stamina barriers, we will include checkpoints that give the sample member an opportunity to take a break. In a recent 44-minute interview of people with disabilities, we offered three opportunities for respondents to stop the interview and to be called back later. Eleven percent of the 1,500 respondents requested at least one break. These respondents tended to tire quickly or to have difficulty using the telephone for prolonged periods. Every respondent who needed a break honored his or her commitment to complete the interview (Ciemnecki and Cybulski 2000). To overcome cognitive challenges, we will develop selection rules and question wording for proxy respondents.

After we have developed solid draft survey instruments (that is, the core plus the modules), the fourth and final step will be to pretest each of the five instruments on beneficiaries with the relevant conditions. We will perform formal telephone pretesting through MPR's telephone center to evaluate interview length, flow, format, item nonresponse, and the CATI program. We will perform the pretests on nine beneficiaries for each disease, as allowed by OMB (a separate instrument will be submitted to OMB for each disease). We will test both English and Spanish versions of the instrument. Finally, we will finalize the questionnaires with HCFA and will prepare the OMB clearance package.

For the evaluation of programs that target a specific condition, we will administer the core and the module appropriate for that condition to all sample members. In programs that do not target specific conditions, all sample members will receive the core component of the survey. In addition, we will explore using intake forms and Medicare claims data to test whether we will be able to determine if some sample members in sites that do not target a particular disease have a principal disease or condition. If we are able to make this determination, then we will also be able to administer the disease-specific modules to sample members whose principal condition can be identified. If we cannot do this, we will ask patients to identify their principal condition during the survey and will conduct the appropriate module.

d. Response Rates

We estimate a 90 percent response rate in random assignment sites for the six-month follow-up patient survey. We base this estimate on our previous experience with similar studies of Medicare beneficiaries enrolled in HCFA-sponsored demonstrations that collected patient contact information (including telephone numbers). Recent examples of such surveys, where we have achieved response rates of 90 percent or higher, are the Home Health Prospective Payment Evaluation's per visit and per episode surveys and the surveys conducted on the Evaluation of the Medicare Case Management Demonstration. In contrast, other surveys of Medicare beneficiaries that have had response rates of only 65 to 70 percent had to rely exclusively on contact information from the EDB or Group Health Plan (GHP) files, which contain good names and reasonably accurate addresses but no telephone numbers.

We expect to be able to obtain telephone numbers and up-to-date addresses for participants (and their next of kin) in this demonstration from the intake forms completed at the time of enrollment; because we do not expect much movement in the six-month period after intake, we should achieve a 90 percent response rate. In comparison design sites, we will have this contact information only for participants; in random assignment sites, we will have contact information for all members of the treatment and control groups. If we are unable to reach a beneficiary by telephone at the six-month interview point, we will contact the next of kin. We also expect to be able to work with the care coordination programs to locate the few sample members whom we cannot find after intensive telephone and electronic searching. Because we will use HCFA’s contact data (which does not contain telephone numbers) for comparison group members and eligible nonparticipants in the sites using comparison group designs, we expect a lower response rate (70 percent) for those beneficiaries.

2. Physician Survey

The purpose of the physician survey is to provide detailed descriptions of physicians’ reactions to, and satisfaction with, the different care coordination programs. Physician acceptance of care coordination will be critical to its success, so these descriptive analyses will be important in assessing the viability of both care coordination in general and of each specific model tested. To facilitate a high response rate, the survey length will be limited to 10 minutes—comparable to the time a physician would spend for an office visit.

a. Timing

We will conduct two rounds of the physician survey, the first one 9 months after programs begin enrolling patients, and the second about one a year later (22 months after program startup).[12] This approach will enable us to assess differences in physicians' reactions as the programs gain experience.

b. Sampling

Because we are interested in physicians' satisfaction with specific programs, surveys will be conducted on the physicians of treatment group patients. A key issue is whether to sample only what we call primary care physicians (PCPs) or to include other physicians (usually specialists) who provide the treatment group patients with care. By PCP, we mean the physician whom patients see most often for their routine care and who is likely to be the primary point of contact for the care coordinator. The PCPs of Medicare beneficiaries in the fee-for-service program may be generalist physicians (family practitioners, general internists, or geriatricians), physicians specializing in the condition of interest (for example, diabetologists caring for patients with diabetes or cardiologists caring for patients with CHF), or specialist physicians whose specialty is unrelated to the condition under focus (gastroenterologists caring for patients with CHF or pulmonologists caring for patients with diabetes).

For programs in which the care coordinator will have contact only with their patients' PCP, we will survey only PCPs. In these sites, specialists will have had little or no contact with the demonstration and therefore are unlikely to provide much useful information for the evaluation. In sites that *do* plan to coordinate care among different physicians, we will survey both PCPs and

[12] We refer to the time when a program begins enrolling patients as "program startup."

specialists who treat conditions targeted by the program in order to obtain some information on how well the program improves communication with these physicians.

We will ask programs to record on the intake form the names, addresses, and telephone numbers of the patient's PCP and up to two other physicians whom the patient sees frequently. Some programs require physician's consent, and others will notify the physicians, so programs are likely to need this information in any case.

We may not be able to identify physicians readily in demonstration sites in which the structure of the program's organization or the nature of its intervention precludes the easy identification of physicians at intake. To identify PCPs in those sites, we will use the patient surveys, which we will administer on a rolling basis to patients six months after they enroll. These surveys will ask patients to name their "personal" or "regular" physician whom they "usually see when they are sick or need advice about their health" relating to the target condition (Spiegel 1983; and Center for Studying Health System Change 1997). Even if patients are unable to provide complete or accurate addresses or telephone numbers, we should have no difficulty locating a small number of physicians in an area by name. We will ask a patient who identifies more than one primary physician to identify the one seen most recently; we will then interview that physician.

The sample will be selected in two waves, the first at the end of the first six months of enrollment (or perhaps somewhat later if substantially less than half the target sample size has not been reached), and the second at the end of patient sample intake. The two waves will enable us to assess whether and how physicians' reactions to the program change over time. These changes may occur as both the program and the physician gain more experience, and as the set of physicians whose patients have enrolled changes.

We do not plan to limit the survey to the same PCPs in the two waves, as the questions are "snapshots" of satisfaction, rather than perceptions of change over time. Furthermore, restricting the sample to the same physicians would exclude physicians whose patients enrolled later on, as well as create difficulties if physicians relocate or drop out. However, physicians will be eligible for selection in both waves. We will not exclude physicians if all their enrolled patients have died or dropped out of the program. We want to understand the physicians' perspectives regardless of whether their patients were more or less severely ill, and whether their patients were satisfied with the program or not.

In each wave, we will examine the list of physicians identified for patients enrolled at roughly 3 months prior to each physician survey (that is, approximately 6 months and 19 months after program startup) and will select a sample of approximately 36 physicians. In sites in which we will survey both the PCPs and the specialists, the sample will include approximately 25 PCPs and 11 specialists. If fewer than 36 physicians are represented, we will include all of them in the survey. These physicians will be approached for interview, with the expectation that approximately 70 percent will complete surveys, yielding data from a total of 25 physicians in each site.

If sampling is necessary, we will consider sampling the physicians with probability proportional to the number of patients in the treatment group or stratifying the sample by this number of patients. Either approach ensures that the sample will not be dominated by physicians each of whom has one or two patients in the program and therefore very little experience with it. At this point, we lean toward stratifying by the number of patients, with some modest oversampling of physicians who have relatively many patients. This approach will balance the goals of (1) not wasting many observations on physicians with very little program experience, and (2) obtaining an unbiased assessment of the program. The concern about bias is that physicians with the greatest number of patients in the program are likely to be the ones with the most favorable impression of the program.

c. Instrument Development

The physician survey will collect basic background information and information on satisfaction with aspects of the care coordination program. As in the patient survey, we will try to base the physician survey on similar physician questionnaires that have been successfully fielded in the past (Dixon et al. 2000; Beck et al. 1997; Boult et al. 1998; and Center for Studying Health System Change 1997). The survey instrument will ask questions general enough to cover a diversity of programs and target diseases. We will pretest the physician survey with as many as nine physicians before submitting it for OMB clearance.

3. OMB Clearance for Patient and Physician Survey Instruments

We plan to seek only one round of OMB approval for the six-month patient follow-up survey and the physician survey instruments. We will develop one physician survey and five patient surveys. As discussed, each of the five patient surveys will contain a common core of general questions and a disease-specific module tailored to the patient's condition. Our approach to developing and pretesting the surveys takes into consideration the range of target diseases and interventions that demonstration programs have proposed and OMB clearance requirements to minimize the time and expense required. We will work closely with HCFA and the demonstration sites, with two goals in mind. First, we want to ensure that the core and modular survey questions are appropriate to the diversity of programs and diseases. Second, we will submit the clearance package to OMB as soon as possible so we have clearance in time to field surveys for early enrollees. We are currently planning to have a draft OMB package to HCFA by March 15, 2001.

4. Medicare Eligibility and Claims Files

We will use Medicare HIC numbers and other identifying information from the demonstration projects to develop a finders file, or list of beneficiaries for whom Medicare data will be requested from HCFA. Our current plan is to extract claims data for patients from the SAF. We will assume a four-month lag between the receipt of a Medicare-covered service and its appearance on these files for the impact analyses. For example, claims data drawn in month 47 for the final synthesis and the Reports to Congress will cover patients' experiences through month 43 of the project. When drawing the comparison group for sites using a comparison design, we will assume a shorter, three-month lag time to ensure we can draw a comparison group in time to conduct the six-month patient survey.[13]

Medicare eligibility data will be extracted from the Health Insurance Skeleton Eligibility Write-off (HISKEW) file. The EDB will provide beneficiaries' demographic characteristics (age, sex, race), dates of death, Medicare entitlement, HMO enrollment, reason for Medicare entitlement, and dual eligibility status. 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, and physician and other Part B providers) both before and after enrollment.

Unless a beneficiary was hospitalized at the time of enrollment, reference periods for Medicare claims-based measures of cost and service use for treatment and control group members will be defined by the date the beneficiary was randomly assigned to treatment or control status. If the beneficiary was hospitalized on the day of random assignment, the

[13] As discussed in footnote 2 of this chapter, we will discuss with HCFA whether we can obtain access to claims data on this schedule.

measures will be defined by the day after hospital discharge. We will assign costs and service use to the postenrollment period in this way because, in practice, care coordination would not be able to alter outcomes until the stay that identified a potential client to the project was over. Thus, the costs of the identifying hospitalization, which may be substantial, will be counted as preenrollment costs.

5. Site-Specific Data

We will use three types of data that sites may collect in the impact analyses: information collected on intake forms, patient-level data on coordinated care services used, and program-level cost data. We will work closely with the sites to determine whether any other site-specific data will be useful to the evaluation. Any available site-specific data will be used to *enhance* the analyses for that site. In general, we expect that data availability will vary by site. To ensure comparability of impact estimates across sites, our synthesis reports will present estimates for each site that do not use any site-specific data. These estimates will rely only on variables constructed from Medicare claims data, the EDB, and the six-month patient follow-up survey, which are available for sample members for all sites.

a. Intake Forms

All sites must obtain consent from beneficiaries before the beneficiaries can participate in the demonstration. For sites using random assignment, these consent forms will be sent to MPR, to randomize patients to the treatment or control group. These consent forms should collect the Medicare beneficiary HIC number; name, address, and telephone number of study participants; and the name, address, and telephone number of the participants' next of kin. In addition, they should ask patients to identify the physician they will see the most regarding their condition.

Each beneficiary would sign the form to indicate consent to participate in the study; consent would include a commitment to respond to a telephone interview six months after enrollment.

Collecting these data on the consent form will be important to the evaluation. As we have noted, obtaining the patient's telephone number from the intake form will enable us to obtain a high response rate on the six-month follow-up survey. Having a correct Medicare HIC number is critical to obtaining Medicare claims data, to estimate impacts on cost and service use. The name of the beneficiary's physician will be used to identify the physician sample to be surveyed.

Although our main impact analyses will use survey and claims data available uniformly for all sites, any site collecting additional information on an intake form would offer an *opportunity* for an enhanced analysis on that site. We will encourage sites to collect intake data about the study participants at the time of enrollment that we would not be able to obtain from the EDB or from the six-month patient survey but will not encourage providers to deliver extra services to the control group. MPR will propose a list of such characteristics to be included by all sites, which might include information on the severity of the patient's illness; measures of functioning or health status; and, perhaps, the patient's knowledge of his or her illness, self-care behavior, and attitude toward seeking care and following physicians' orders. In addition, some sites may collect data used to determine eligibility or to tailor their interventions. These data on patient characteristics obtained from the intake form are *not* essential for the analysis, but they offer the opportunity in random assignment sites to improve the precision of our estimates and to identify subgroups for which a site may be more or less effective. For example, if the subgroups are statistically large, we might be able to test whether impacts were larger for beneficiaries with more severe conditions than for other sample members. We successfully used the approach of obtaining data from intake forms in our Evaluation of Medicare Case Management Demonstrations, conducted for HCFA between 1993 and 1997.

The intake data for sites using random assignment have more analytic value than do similar data for sites using comparison designs; in the former case, the data are available for both treatment and control groups. The treatment and control groups in sites using random assignment should be very similar at the time of enrollment, so there is no real need to control for preexisting differences between them. If intake data on additional characteristics and attitudes *were* available, however, we would include them in the regression equations for that site, to increase the precision of the estimates. Sites that use a comparison design would be able to collect intake data on participants only. Because there would be no such data for either the nonparticipating eligibles or the comparison group, we cannot use them as control variables or to define subgroups in our analyses. However, the contact information will be helpful for conducting the follow-up survey. Contact information will hold down search costs and will increase response rates for the participant portion of the sample. We can also assess how participant *outcomes* (but not impacts) vary with patient characteristics at the time of enrollment.

b. Services Provided

With the help of the implementation contractor, some sites may track the amount and type of intervention services provided to each patient. For sites that collect high-quality data, we will analyze the effect of service receipt on outcomes among treatment group members. This information would not be available for the control group, so we will not be able to use it to estimate impacts. As discussed in the section on statistical methodology (Section III.D), endogeneity problems linking high service use to high medical needs will require that we use care when interpreting models estimating the effect of service provision on outcomes.

c. Cost Data

The final set of site-specific data we will collect for use in the impact analyses includes data on total program costs. These data will be collected as part of the implementation analysis, using invoices to HCFA and information collected from program-specific documents during site visits. Cost data will be used in the cost-effectiveness analysis. We will disaggregate program costs in two ways, into start-up versus ongoing costs, and into the costs of specific components of the intervention. Disaggregation in this way will help us to extrapolate what an ongoing program would cost HCFA, based on different packages of care coordination services.

C. OUTCOME MEASURES

1. Quality of Care

This section describes the outcome measures we plan to use for the analysis of the programs' effects on quality of care. Decision makers need to know program impacts both on Medicare costs and on patients' quality of care. Obviously, programs that simultaneously decrease Medicare costs and improve, or at least maintain, quality of care are attractive policy options. So, too, are programs that are cost neutral and improve quality of care. However, some programs may substantially improve quality while modestly increasing Medicare costs or, conversely, may substantially decrease Medicare costs with a small decrement in quality of care. Decisions on the wide implementation of programs like these are harder to make.

We organize the following discussion around the well-known process and outcome framework for assessing the quality of care developed by Donabedian (1980).[14] Process measures include data on the care provided to patients (for example, whether certain assessments

[14] Donabedian's framework also encompassed structure, which includes such features as the composition and training of staff and patient to provider ratios. Some of these features will be studied as part of the implementation analysis, discussed in Chapter II.

were conducted, tests or medications ordered, or patient education provided). Outcome measures of care describe the results of care, such as patients' health-related behavior or patients' health status. We also categorize quality indicators as "generic" or disease-specific. Generic measures, such as receipt of influenza vaccination or self-perceived health status, apply to all patients regardless of their diagnoses or conditions. Disease-specific measures, such as the performance of certain blood tests in diabetes, the presence of specific symptoms in heart failure, or the occurrence of particular complications in coronary disease, are appropriate only for patients with those conditions. Generic measures require the development of only a single set of measures, allow uniformity in data collection and analysis, and permit comparisons across sites regardless of target populations or diagnoses. However, in evaluations of programs that focus on specific diseases, generic measures may be less sensitive than disease-specific conditions to impacts on quality. Moreover, in disease-specific programs, program staff and patients are likely to find disease-specific survey questions more relevant and meaningful than generic ones (Patrick and Deyo 1989). The limitation of disease-specific measures lies in their very specificity. A separate group of measures must be developed for each disease or condition, and, in a demonstration like this one, the number of such groups of measures could be quite large. For these reasons, then, we plan to collect a set of generic measures from all patients, with additional disease-specific measures for CHF, diabetes, coronary artery disease (CAD), COPD, cancer, and stroke. These conditions are among the most common ones targeted by the awarded sites, as well as the ones with the best developed quality measures in the literature. For patients with less common diagnoses, we will rely on the generic measures to assess program impacts.

Including both generic and disease-specific questions to collect the survey-based measures may increase the length of the patient interviews unacceptably. In developing the survey instrument, we will therefore (1) collect a large pool of candidate questions, (2) prioritize the topics on such criteria as policy relevance and anticipated responsiveness to program effects, and (3) eliminate lower-priority items to reach a final set of questions that can be covered in an interview of reasonable length.

a. Potential Program Impacts on the Quality of Care

The essential premise of coordinated care programs is that the systematic delivery of certain key services (processes of care) to people with chronic illness will lead to both improved health outcomes and decreased health care utilization. Categorized into three larger steps, these processes include (Chen et al. 2000; Case Management Society of America 1995; and American HealthWays 1999):

1. ***Thorough Assessment and Planning.*** Identifying and addressing all important problems, choosing a clear set of goals, and developing a practical plan of care
2. ***Implementation and Delivery.*** Building relationships with patients, families, and primary care providers; providing support; arranging services; delivering evidence-based clinical interventions; and educating patients
3. ***Reassessment and Adjustment.*** Performing periodic reassessments, ensuring accessibility, and promptly making needed adjustments to the plan of care

Failure to address these processes of care is believed to increase patients' risk of adverse outcomes: treatment nonadherence; poor health status; dissatisfaction; repeated preventable hospitalizations; and ultimately, death.

The systematic approach of care coordination programs toward these essential processes of care contrasts with the haphazard approach characteristic of the current health care system. As researchers have repeatedly noted, traditional health care, with its emphasis on acute care, is too rushed, fragmented, and dependent on patient-initiated followup to render appropriate coordinated care for people with chronic illness (Wagner et al. 1996; Holman and Lorig 2000; Manian 1999; and Clark and Gong 2000).

Processes of Care. Our hypotheses of program impacts on process measures follow logically from the preceding discussion. We anticipate that, compared with the "usual care" the control group will receive, treatment group members are more likely to receive:

- ***Thorough, Systematic Assessments.*** Treatment group members' assessments are more likely to address such problems as patients' functioning, emotional distress, and health behaviors.
- ***Care Planning.*** Members of the treatment group are more likely to be aware of and to participate in formulating a set of goals and a clear plan to achieve those goals.
- ***Patient Education.*** Treatment group patients are more likely to receive education on both generic issues, such as diet, exercise, smoking, and medication adherence, and disease-specific issues, such as monitoring of symptoms, self-management of conditions, and handling of emergencies. Treatment group patients also are more likely to receive training in coping skills to manage the stresses of chronic illness, and in assertiveness and communication skills to deal with their physicians and the health care system.
- ***Service Arrangement.*** Treatment group patients are more likely to receive services they feel they need.
- ***Clinical Interventions.*** Treatment group patients are more likely to receive evidence-based interventions, both generic (such as influenza vaccinations) and disease-specific (such as specific medications for heart failure or tests for diabetes).
- ***Followup on Interventions.*** Treatment group patients are more likely to receive followup to ensure interventions are delivered as planned.
- ***Communication Across Providers.*** Treatment group patients are more likely to have providers who are informed about important facts about their cases.
- ***Periodic Reassessment.*** Treatment group patients are more likely to undergo periodic reassessment and monitoring of their condition.
- ***Ready Access to Answers to Health Questions and Concerns.*** Treatment group patients are more likely to be able to easily contact a health professional to answer urgent and nonurgent questions about self-care or symptoms.

In this list, we have tried to exclude hypotheses that are not amenable to treatment and control comparisons. For example, the building of strong relationships with patients, families, and PCPs by care coordinators is an important process in care coordination but obviously has meaning only for patients who *have* care coordinators (that is, patients in the treatment group).[15] However, most of the other processes are relevant for both treatment and control group members. Among control group members, the processes can and are (infrequently) performed by preexisting resources, such as PCPs, disease-specific support groups, Area Agencies on Aging, family members, and even patients themselves. As discussed below, we intend to survey treatment group members (and their physicians) in order to perform a descriptive analysis of their experiences with the programs.

We have also tried to exclude hypotheses involving process measures that are difficult or impossible to measure. The data available for the evaluation are Medicare claims and patient survey data. Medicare claims data can provide information only on services that can be billed to Medicare, such as laboratory tests, x-rays, eye exams, and visits to physicians. The patient survey data can provide details on processes of care not included in claims data, such as the provision of patient education, but they are still limited to information that patients are able to observe and recall. For example, checking for medication interactions may be an important process of care for chronically ill seniors, but it is unclear whether patients will be able to report on it. It is also unclear whether patients will be able to accurately recall care they received several months previously (for example, whether a specific issue was addressed during a health assessment). We will rely on the pretest of the survey to clarify these questions.

Outcomes of Care. Our hypotheses about program effects on patient outcomes also follow from our discussion of the goals of coordinated care programs. We thus expect treatment group patients to experience, relative to control group patients, positive impacts on the following outcomes:

[15] Enrollees in 1 of the 15 awarded programs, Qmed, Inc., will not have care coordinators.

- ***Health-Related Behaviors.*** Treatment group patients are more likely to be successful in generic behaviors, such as taking medications, quitting smoking, and increasing exercise, as well as in disease-specific behaviors, such as modifying diet and monitoring symptoms. They will also do a better job of managing stress, communicating with their physicians, and interacting with the health care system.

- ***Health and Functional Status.*** Treatment group patients will experience less functional impairment and activity restriction due to health problems. They will also enjoy an improved physical and emotional health-related quality of life.

- ***Patient Ratings of Care.*** Treatment group patients will rate their health care more favorably on a variety of dimensions, such as access, coordination, chronic illness support, service arrangement and unmet needs, and satisfaction. They will also rate their relationships with their PCPs more highly as a result of the programs' effects on communication and coordination in patients' care, patients' skills in interacting with physicians and the health care system, and improved health outcomes.

- ***Preventable Hospitalizations and Mortality.*** Treatment group patients are less likely to need hospitalizations for preventable acute exacerbations or complications of chronic illness. These preventable hospitalizations may be either generic or disease-specific. Their decreased morbidity may translate into decreased mortality.

Descriptive Analyses. Finally, we plan to conduct descriptive analyses using data collected only from patients in the treatment group (and from their physicians). We will ask treatment group patients to provide reports and ratings of their experiences with their care coordinators. These data will help us gauge whether care coordinators successfully established rapport with patients, provided support to patients, and arranged services, all key steps in care coordination (Chen 2000; and Case Management Society of America 1995). We will survey the physicians of program enrollees on their experiences and satisfaction with the program. Physicians' perceptions and acceptance of the programs will have important implications for efforts to widen implementation of such programs.

b. Measures of the Process of Care

Most of the data on processes of care will come from the six-month patient survey, with some obtained from Medicare claims data.

Content of Patient Assessments and Care Planning. It is important that chronically ill, elderly patients undergo periodic, thorough assessments to identify incipient or overt problems that threaten their health, and that clear goals and a plan to achieve these assessments are established. Some of the important areas to be assessed are generic: medical issues (review of medications), functional issues (ability to perform basic and instrumental activities of daily living), emotional issues (depression, coping skills), social situation (living arrangement, social support) and behavioral issues (medication adherence, smoking, alcohol use, exercise, diet, weight loss) (Fleming et al. 1995). We will ask patients whether their physician or any other health care provider discussed each of these issues over some fixed recent period of time. Table III.3 summarizes the generic topics we plan to cover.

Patients with certain conditions should also be assessed to determine how much they know about relevant self-care skills. For example, patients with heart failure should be asked whether they know how to monitor changes in their weight. Similarly, if they have diabetes, they should be questioned about their understanding of foot care; if they have COPD, their inhaler technique should be checked. Thus, we will ask patients with these conditions whether a health care provider discussed these skills with them. Recommended patient education topics are less straightforward in other diseases; however, we will incorporate any existing guidelines into the disease-specific modules. (Tables III.4 through III.9 provide preliminary lists of the disease-specific measures.)

Treatment group patients may have been assessed more recently than control group patients. Treatment group patients presumably will have received their initial assessments from program staff shortly after enrollment, and the survey will be conducted roughly six months after

TABLE III.3

SUMMARY OF QUALITY-OF-CARE MEASURES:
GENERIC

Items	Source of Items	Data Collection Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked about following issues at last health assessment[a]: Diet Review of all medications Medication adherence Physical activity Smoking Alcohol intake Functional or sensory limitations Symptoms of depression Amount of social support Advanced directives	BRFSS and draft	Patient survey
Patient Education		
Whether respondent reports receiving education or a referral for education on any of the preceding topics or on any other health promotion/health maintenance topic	Draft	Patient survey
Whether respondent reports receiving explanation on what symptoms or problems to look out for in his/her health conditions, and what to do if they appeared	Picker	
Care Planning		
Whether respondent reports setting goals for health with a health care provider, and developing a plan to meet those goals	Draft	Patient survey
Whether respondent reports being involved as much he/she wanted to be in setting these goals and making the plan	Draft	Patient survey
Service Arrangement and Follow-Up		
Whether the respondent reports in the past six months needing: Prescription medications Equipment Therapy Home health care Other assistance (such as in housecleaning, yard work, meals, personal hygiene, home repairs, errands, or transportation) Emotional support or counseling	SPEC, CAHPS	Patient survey
If so, for each, whether respondent received help (other than from friends and family) in obtaining these services, and whether reports a problem obtaining these services	Draft	Patient survey
If received the service for each, whether respondent reports someone other than friends and family checked to make sure the service met the need	Draft	Patient survey
Communication Across Providers		
Whether respondent reports instances in which health care professionals involved in his or her care had not spoken to each other, or did not have information he/she thought they should	Picker	Patient survey
Periodic Assessment		
Whether respondent reports times when a physician or nurse checked on him/her just to see how he/she was doing	Draft	Patient survey
How many times in the past 6 months respondent has spoken on the telephone or received a visit from or gone to an appointment with a nurse or physician	Draft	Patient survey

TABLE III.3 (*continued*)

Items	Source of Items	Data Collection Method
Whether respondent reports times when a health problem could have been avoided through more frequent contact with his/her physician or nurse	Picker	Patient survey
Ready Access to Answers and Advice		
Whether respondent reports being able to talk to someone to get help or advice related to his/her health as soon as he/she needed to	Picker	Patient survey
When respondent needed health-related help or advice in past 6 months, how often received that help or medical advice.	CAHPS	Patient survey
Clinical Interventions		
Whether and when respondent reports having:	BRFSS	Patient survey
Measurement of blood pressure, height, weight		
Vaccinations for influenza and pneumococcal pneumonia		
Screenings for colon and breast cancer		
Efforts to reduce number of medications		
Measures of the Outcomes of Care		
Health-Related Behaviors		
Whether, on advice of physician or other health professional, respondent tried to:	Taira	Patient survey
Stop smoking		
Lower alcohol intake		
Increase physical activity		
Respondent's current levels of:	BRFSS	Patient survey
Smoking and alcohol intake		
Physical activity levels		
Respondent's use of :		
Behaviors for cognitive symptom management[b]	SPEC	Patient survey
Effective behaviors for communicating with physicians and the health care system		
Community services for "tangible help" (such as personal hygiene, meals, transportation, and so on) or for emotional support		
Community health education programs or community support groups for diseases or health problems.		
Respondent's self-rated knowledge of what to be aware of with his/her health condition	Picker	Patient survey
Respondent's self-rated knowledge of what to do if his/her health problem didn't get better or got worse	Picker	Patient survey
Respondent's self-rated ability to manage his/her health problems	DSCA	Patient survey
Health and Functional Status		
Basic and instrumental activities of daily living	SHMO	Patient survey
Number of bed days in past two weeks	HH2	Patient survey
In past 30 days, number of:	BRFSS	Patient survey
Physically, mentally, or overall unhealthy days		
Days with activity limitation		
Days with pain causing difficulty in usual activities		
Days of feeling depressed, anxious, or not energetic, or days with inadequate rest or sleep		
Generic physical and emotional function, and self-perceived health	SF-12	Patient survey
Patient Rating of Care		
Respondent's perceptions of ease of:	MCM	Patient survey
Getting prescriptions filled		
Arranging for transportation to medical care		
Obtaining other needed services		
Respondent's reports of unmet service or care needs	MCM and HH2	Patient survey
Respondent's ratings of:	MCM	Patient survey
Advice on ways to prevent illness and promote health		
Reminders to make or keep appointments for medical care		
Overall quality of care received during the past six months		

TABLE III.3 (*continued*)

Items	Source of Items	Data Collection Method
Respondent’s ratings of primary care physician in the following areas[c]:	PCAS	Patient survey
Accessibility, responsiveness, continuity, and attentiveness		
Familiarity with respondent’s medical and social situation, involvement in care, and coordination of care		
Knowledge of respondent’s wishes and goals, explanations of medical problems, guidance in health matters, and trustworthiness		
Preventable Hospitalizations	Culler	Medicare claims data
Pneumonia		
Falls or hip fracture		
Dehydration		
Gastroenteritis		
Cellulitis		
Pyelonephritis		
Mortality	Stewart	Medicare enrollment data

NOTE: We will be collecting these measures on all patients, regardless of diagnosis or condition.

[a] We will make clear that “health assessment” means a routine appointment, not an appointment for a specific problem. As discussed in the text, we will asses in the patient survey pretest whether patients will be able to recall details of a health assessment that may have occurred during some fixed period of time, such as the past 6 months or 12 months. The “last health assessment” includes general physical exams, routine checkups, and telephone interviews by a physician, nurse, or other health professional.

[b] Cognitive symptom management is a set of techniques to deal with such problems as frustration, fatigue, pain, and isolation that frequently afflict individuals with chronic illness.

[c] As discussed in Section C.1.a of this chapter, one of our hypotheses is that the programs, by empowering them and improving their health outcomes, will increase the treatment group patient’s satisfaction with their primary care physician relative to control group patients.

CAHPS = Consumer Assessment of Health Plans Survey (Agency for Health Research and Quality 1998); Picker = Picker Ambulatory Care Patient Interview, from Lorig et al. (1996); BRFSS = CDC’s Behavioral Risk Factor Surveillance Survey (Centers for Disease Control and Prevention 2001); Draft = questions that will be drafted for this survey; Taira = Taira et al. (1999); SPEC = Stanford Patient Education Center (Lorig et al. 1996); DSCA = Diabetes Self Care Activities (Toobert and Glasgow 1994; and American Diabetes Association (2000), modified for generic chronic illness; SHMO = patient survey developed by MPR for analysis of the Social HMO II; MCM = patient survey developed by MPR for the Medicare Case Management Evaluation; SF-12 = Short Form-12 (Ware et al. 1996); PCAS = Primary Care Assessment Survey (Safran et al. 1998); HH2 = patient survey developed by MPR for the evaluation of the Medicare Home Health Prospective Payment Demonstration, Phase II; Culler = Culler et al. (1998); Stewart = Stewart et al. (1999).

TABLE III.4

SUMMARY OF DISEASE-SPECIFIC QUALITY-OF-CARE MEASURES:
CONGESTIVE HEART FAILURE

Items	Source of Items	Data Collection Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked at last health assessment about his or her dietary salt intake, frequency of self-weighing, knowledge of what to do with weight information, emotional coping with CHF, effects on family	Draft	Patient survey
Patient Education		
Whether respondent reports receiving education or a referral for education on CHF, symptoms to be monitored, how to respond to symptoms, and dietary salt intake	Draft	Patient survey
Clinical Interventions		
Whether and when respondent had an examination of the lungs and heart with a stethoscope	Draft	Patient survey
Whether or not the respondent is currently taking ACE inhibitors, AR blockers, spironolactone, beta-blockers.[a] If not, did physician tell not to take?	Draft	Patient survey
Measures of the Outcomes of Care		
Health-Related Behaviors		
Whether, on advice of physician or other health professional, respondent has tried to reduce dietary salt	Draft	Patient survey
Respondent's current dietary salt intake	Draft or Block/NCI	Patient survey
Respondent's adherence to medications	MCM	Patient survey
Respondent's current practice in weighing self		
Respondent's self-rated understanding of what to do about weight fluctuations or symptoms, and self-rated ability to take care of him/herself		
Health and Functional Status		
Physical and emotional impacts of CHF	LIHFE	Patient survey
Preventable hospitalizations	Culler	Medicare claims data
CHF		
Hypokalemia (potassium deficiency)		
Hyponatremia (water overload)		

NOTE: We will collect these measures, in addition to the generic measures in Table III.3 on all CHF patients.

[a] ACE inhibitors, AR blockers, spironolactone, and beta-blockers are all medications shown or believed to be beneficial for patients with CHF.

ACE = (angiotensin converting enzyme, AR = angiotensin recepter; draft = questions that will be drafted for the purpose of this survey; Block/NCI = dietary questionnaire developed at National Cancer Institute (Block et al. 1986); MCM = patient survey developed by MPR for the Medicare Case Management Evaluation; LIHFE = Living with Heart Failure health status instrument (Rector and Cohn 1992); Culler = Culler et al. (1998).

TABLE III.5

SUMMARY OF DISEASE-SPECIFIC QUALITY-OF-CARE MEASURES:
DIABETES

Items	Source of Items	Data Collection Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked at last health assessment about his/her:	Draft	Patient survey
Timing of meals and diabetes medications		
Foot care		
Home blood sugar testing		
Knowledge of what to do with test results		
Emotional coping with diabetes		
Sexual functioning		
Effects on family		
Patient Education		
Whether respondent reports receiving education or a referral for education on nutrition and exercise for people with diabetes, or any diabetes education at all, or meeting with a Certified Diabetes Educator	Draft	Patient survey
Clinical Interventions		
Whether and when respondent has had a foot exam (and whether with a special monofilament device to test sensation)	DQIP	Patient survey
Whether and when respondent had:	DQIP	Medicare claims data
Dilated retinal exam		
Blood test for hemoglobin A1c		
Urinalysis for microalbumin		
Blood test for cholesterol or lipids		
Measures of the Outcomes of Care		
Health-Related Behaviors and Knowledge		
Respondent's current adherence with diet, blood sugar testing, foot self-examination, and medications	DSCA	Patient survey
Respondent's rating of his/her understanding of:	DSCA	Patient survey
Foot care		
Nutrition		
Exercise		
Blood sugar testing		
Blood sugar target levels		
Management of diabetic symptoms		
Health and Functional Status		
Physical and emotional impacts of diabetes	PAID, DH, DQOL	Patient survey
Preventable Hospitalizations	Culler	Medicare claims data
Diabetes out of control or diabetic coma		
Gangrene		
Surgical debridement (removal) of infected tissue		
Lower extremity amputation		
Diabetic foot infection		

TABLE III.5 (*continued*)

NOTE: We will collect these measures, in addition to the generic measures in Table III.3, on all patients with diabetes.

Draft = questions that will be drafted for this survey; DQIP = survey developed for the Diabetes Quality Improvement Project, a coalition consisting of the American Diabetes Association, the Foundation for Acccountability, HCFA, the National Committee for Quality Assurance, the American Academy of Physicians, the American College of Physicians, and the Veterans Administration (American Diabetes Association 2000); DSCA = Diabetes Self-Care Activities (Toobert and Glasgow 1994; and DQIP 1998); DQOL = Diabetes Quality of Life, a 28 item instrument measuring diabetes-related quality of life (DCCT Research Group 1988); PAID = Problem Areas in Diabetes, a 21-item instrument measuring diabetes-related quality of life (Welch et al. 1997); DH = Diabetes Hassles, four questions on activity or lifestyle restrictions associated with diabetes (Greenfield et al. 1994; Culler = Culler et al (1998).

TABLE III.6

SUMMARY OF DISEASE-SPECIFIC QUALITY-OF-CARE MEASURES:
CORONARY ARTERY DISEASE

	Source of	Data Collection
Items	Items	Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked at last health assessment about his/her:	Draft	Patient survey
Adherence to cardiac medication regimen		
Cardiac symptoms		
Knowledge of what to do about symptoms		
Emotional coping with coronary disease		
Sexual functioning		
Effects on family		
Patient Education		
Whether respondent reports receiving education or a referral for education on:	Draft	Patient survey
What to expect with CAD		
Nutrition and exercise for people with CAD		
Referral for cardiac rehabilitation		
Clinical interventions		
Whether and when respondent had an examination of the lungs and heart with a stethoscope	Draft	Patient survey
Whether respondent is currently taking aspirin or other antiplatelet drug, cholesterol-lowering medication, or beta-blocker medications. If not, did physician tell not to take?	Draft	Patient survey
Fasting blood test for lipids		Medicare claims data
Measures of the Outcomes of Care		
Health-Related Behaviors and Knowledge		
Respondent's rating of his/her understanding of nutrition and exercise, and rating of ability to self-manage CAD	DSCA	Patient survey
Adherence to medications, diet, and exercise		
Health and Functional Status		
Physical and emotional impacts of CAD	SAQ or QLMI	Patient survey
Preventable hospitalizations	Culler	Medicare claims data
Unstable angina, myocardial infarction, cardiogenic shock		
Coronary angiography		
Coronary angioplasty		
Coronary artery bypass surgery		

NOTE: We will collect these measures, in addition to the generic measures in Table III.3, on all patients with CAD.

Draft = questions that will be drafted for this survey; DSCA = Diabetes Self-Care Activities modified for CAD (Toobert and Glasgow 1994; and DQIP 1998); SAQ = Seattle Angina Questionnaire (Spertus et al. 1995); QLMI = Quality of Life After Myocardial Infarction instrument (Oldridge et al. 1991); Culler = Culler et al. (1998).

TABLE III.7

SUMMARY OF DISEASE-SPECIFIC QUALITY-OF-CARE MEASURES:
CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Items	Source of Items	Data Collection Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked at last health assessment about:	Draft	Patient survey
Adherence to COPD medication regimen		
Knowledge of inhaler use		
COPD symptoms		
Knowledge of what to do about symptoms		
Emotional coping with COPD		
Effects on family		
Patient Education		
Whether respondent reports receiving education or a referral for education on inhaler use, exercise and breathing for people with COPD, and energy conservation techniques, or a referral for pulmonary rehabilitation	Draft	Patient survey
Clinical Interventions		
Whether and when respondent had an examination of the lungs and heart with a stethoscope, spirometry, or peak flow testing	Draft	Patient survey
Whether the respondent is currently taking ipratropium bromide alone or in combination. If not, did physician tell not to take?	Draft	Patient survey
Measures of the Outcomes of Care		
Health-Related Behaviors and Knowledge		
Respondent's rating of his/her understanding of COPD self-management	DSCA	Patient survey
Adherence to medications, exercise, breathing and energy conservation techniques		
Health and Functional Status		
Physical and emotional impacts of COPD	SOLQ or CRDQ	Patient survey
Preventable Hospitalizations	Culler	Medicare claims data
Exacerbation of COPD or acute bronchitis		
Acute respiratory failure		
Hypercapnea or CO_2 retention		

NOTE: We will collect these measures, in addition to the generic measures in Table III.3 on all patients with COPD.

Draft = questions that will be drafted for this survey; DSCA = Diabetes Self-Care Activities (Toobert and Glasgow 1994 and DQIP 1998), modified for COPD; SOLQ = Seattle Obstructive Lung Disease Questionnaire (Tu et al. 1997; CRDQ = Chronic Respiratory Disease Questionnaire (Guyatt et al. 1987); Culler = Culler et al. (1998).

TABLE III.8

SUMMARY OF DISEASE-SPECIFIC QUALITY-OF-CARE MEASURES:
CANCER

Items	Source of Items	Data Collection Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked at last health assessment about:	Draft	Patient survey
Adherence to medication regimen		
Pain		
Fatigue		
Loss of appetite		
Other symptoms		
Knowledge of what to do about symptoms		
Personal and family coping with diagnosis and treatment		
Patient Education		
Whether respondent reports receiving education, or a referral for education, on what to expect with treatment and tests and how to manage symptoms	Draft	Patient survey
Measures of the Outcomes of Care		
Health-Related Behaviors and Knowledge		
Respondent's rating of his/her understanding of cancer symptom self-management	Draft	Patient survey
Health and Functional Status		
Pain, fatigue, and nausea	Draft and MCM	Patient survey
Bodily self image		

NOTE: We will collect these measures, in addition to the generic measures in Table III.3, on all patients with cancer.

Draft = Questions that will be drafted for this survey; MCM = Patient survey developed by MPR for the Medicare Case Management Evaluation.

TABLE III.9

SUMMARY OF DISEASE-SPECIFIC QUALITY-OF-CARE MEASURES: STROKE

Items	Source of Items	Data Collection Method
Measures of the Processes of Care		
Patient Assessments		
Whether respondent reports being asked at last health assessment about:	Draft	Patient survey
Adherence to medication regimen		
Pain		
Whether stroke affects physical or emotional functioning or bladder and bowel functioning; if so, how severely		
Effects on family		
Patient Education		
Whether respondent reports receiving education or a referral for education on how best to take care of self after a stroke, signs and symptoms of recurrent stroke, and the importance of monitoring and control of cholesterol and blood pressure	Draft	Patient survey
Clinical Interventions		
If functioning impaired, whether respondent has been referred for rehabilitative therapy (physical, occupational, or speech therapy), and whether has been referred for any needed adaptive equipment[a]	Draft	Patient survey
If nonhemorrhagic stroke, whether respondent is currently taking anticoagulant medication (if in atrial fibrillation) or antiplatelet agent(s) (if not in atrial fibrillation)[a]	Draft	Patient survey
If nonhemorrhagic stroke, blood test for cholesterol or lipids	Draft	Patient survey
If elevated cholesterol, whether respondent is taking cholesterol-lowering medication	Draft	Patient survey
Measures of the Outcomes of Care		
Health-Related Behaviors		
Adherence to diet and medications	Draft	Patient survey
Self-rated ability to manage health problems, understanding of health problems	DSCA	Patient survey
Other Outcomes		
If blood pressure elevated, whether blood pressure was acceptable at last measurement by physician	Draft	Patient survey
Preventable Hospitalizations		Medicare claims data
Stroke or transient ischemic attack		
Overanticoagulation or overanticoagulation complicated by hemorrhage		

NOTE: We will collect these measures, in addition to the generic measures in Table III.3, on all patients with stroke.

[a] We will assess whether it is feasible to ask questions in the survey to ascertain this information; if so, we will incorporate appropriate skip logic into the survey instrument.

Draft =questions that will be drafted for this survey; DSCA = Diabetes Self-Care Assessment Activity (Toobert and Glasgow 1994; and DQIP 1998), modified for stroke.

enrollment.[16] Thus, we will ask patients in both the treatment and control groups to focus either on any assessments over some fixed period, such as the past year, or on their "last general health assessment or physical not related to a specific problem" (a modification of a question from the National Health Interview Survey; U.S. Department of Health and Human Services 1984). As discussed, during the survey pretest, we will assess whether patient recall will permit such questions.

Patient Education. Patient education, which enables patients to improve their health-related behaviors, monitor their illnesses, self-manage themselves appropriately, and maintain their health, is another key process in chronic illness care. Some types of education, such as quitting smoking or increasing exercise, are generic. Others, such as education on specific preventive measures or on how to monitor oneself for symptoms and respond appropriately, pertain to particular diseases. Thus, we will survey patients on whether they ever received education on important generic and disease-specific topics.[17]

Clinical Interventions. 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 pneumococcal vaccinations, and periodic measurements of blood pressure, height, and weight). The rest, such as periodic foot exams in diabetes, are disease-specific. As shown in Tables III.4 through III.9, we will be able to identify some of these preventive measures, such as the performance of hemoglobin A1c tests or dilated eye exams in patients with diabetes, using Medicare claims data,

[16] Whether the programs will, in fact, be successful in performing their initial assessments in a timely fashion remains to be seen.

[17] Measuring the *outcomes* of patient education will be discussed in the next section.

but information on others, such as administration of influenza vaccinations or foot exams, will have to be collected in the survey.[18] Because the prescription of appropriate medications is a frequently used process measure of the quality of care (Havranek et al. 1996; and Krumholz et al. 2000), and improving physician prescribing practices is thus often the focus of disease management programs (Aubert et al. 1998; Wells et al. 2000; and Monane et al. 1998), we will test the feasibility of collecting information on prescribed medications from the patient survey (Tables III.4, III.6, III.7, and III.9).

c. Measures of the Outcomes of Care

Under this heading we include a wide variety of measures that, broadly speaking, are *results* of the care patients receive. We include here health-related behaviors (adherence to medications, diet, lifestyle, self-monitoring for signs of illness exacerbation; dealing with health care providers; coping with stress), health and functional status, and satisfaction with health care. We also include patient mortality and hospitalizations that should be preventable if proper care is received (for example, hospitalizations for diabetic coma in patients with diabetes).

Health-Related Behaviors. Lack of adherence to prescribed medication regimens and recommended lifestyle changes is a major problem among persons with chronic illness (Chin and Goldman 1997; and Stewart et al. 1999). These behaviors then lead to acute exacerbation of illness, increased health care use, and poor health outcomes. The Best Practices project showed that many successful programs devote considerable efforts toward increasing patient adherence (Chen et al. 2000)

[18] Although Medicare now reimburses physicians for administering influenza vaccinations, claims data may still provide an incomplete picture of these shots, as many beneficiaries receive them at such locations as senior health centers or supermarket pharmacies.

We will ask all patients about their adherence with several generic lifestyle changes and behaviors, such as smoking cessation, moderation of alcohol intake, and increased activity levels. As shown in Table III.3, well-tested items from a wide variety of surveys are available for all these topics.

Other behaviors are more disease-specific. For example, decreased dietary fat intake would be especially important for those with coronary disease or diabetes, whereas decreased salt intake would be relevant to those with CHF or hypertension. Patients with diabetes should inspect their feet periodically, and patients with CHF should weigh themselves daily. It is also important that patients know how to recognize and act on symptoms, and how to manage emergencies. Where possible, we will develop questions for these areas from existing surveys or from instruments created for clinical studies (Tables III.4 through III.8).

Lorig et al. (1999) identified two additional skills that everyone with chronic illness should master—how to reduce stress, and how to interact effectively with physicians and the health care system. We will measure these behaviors using scales or questionnaires developed by Lorig and colleagues at the Stanford Patient Education Center (Lorig et al. 1996).

Patients' Health and Functional Status. Through all their presumed effects on the processes of care, care coordination programs should ultimately have positive impacts on patients' health and functional status. We will draw our measures of health status from published, well-tested, and psychometrically sound generic and disease-specific health status assessment instruments. These instruments reflect the fairly broad concept of health status that researchers who study health-related quality of life have developed; in addition to physical functioning, this concept encompasses emotional health, sense of well-being, and, sometimes, social functioning. This broad view seems appropriate, given the wide-ranging impacts of chronic illness on individuals' lives.

Although the patients will be surveyed only six months after enrollment, it is conceivable that the programs will have detectable impacts on measures of health status. The programs are supposed to lower hospitalization rates, and hospitalization is generally associated with a severe worsening in health status (Creditor 1993; Sager and Rudberg 1993; and Landefelt et al. 1995). If programs truly are able to avert a substantial number of hospitalizations, and the remaining enrollees experience some mild improvements in health status, there may possibly be measurable impacts on health status.

The potentially wide range of patients enrolled by the various programs also creates possible problems of "floor" and "ceiling" effects in the measurement of health status. Questions about basic activities of daily living, such as the ability to get out of bed without assistance, may be appropriate in severely impaired populations but are unlikely to detect important changes in a less impaired population, because the population's members will be at the "ceiling" for such questions. Likewise, questions that ask about more vigorous activities (vacuuming or bowling, for example) may work well in less impaired populations but will be meaningless to highly disabled persons, all of whom will be at the "floor." We thus anticipate including overlapping measures that cover the same general areas or topics (such as physical function in the above example), but that are tailored for populations with differing degrees of impairment. Tables III.3 through III.8 summarize both the generic and disease-specific measures we are considering.

Patient Ratings of Care. Care coordination programs should have positive impacts on both patients' perceptions about the specific areas of traditional health care that the chronically ill generally do not receive and their overall satisfaction with their health care. Thus, we plan a series of questions on patients' evaluation of how well health care providers know them "as a person," the amount of support they receive to help them cope with their illness, the quality of the patient education received, the presence and severity of unmet needs for service, and the degree of coordination of their care. In addition, we will ask a series of questions focusing on patients' perceptions of their relationships with their PCPs. We hypothesize that the programs, through increased communication, coordination, and empowerment of patients, will enhance this relationship. Examples of the items are listed in Table III.3.

Preventable Hospitalizations and Patient Mortality. Finally, if care coordination programs indeed function as they should, they should be able to help patients avert the health crises that often lead to emergency hospitalizations or even death. Hospitalizations themselves often cause still further declines in function. Prevention of hospitalizations also is essential if programs are to be considered cost effective.

Preventable hospitalizations may be either disease-specific or generic. Examples of disease-specific preventable hospitalizations include hospitalizations for CHF in patients with CHF or for lower extremity gangrene in patients with diabetes. Examples of generic preventable hospitalizations include hospital admissions for hip fractures or dehydration. We will ascertain the occurrence of both categories of preventable hospitalizations during the study's follow-up period, using the ICD-9 principal diagnosis codes in Medicare Part A hospital claims. We will ascertain patient mortality during the study's follow-up period, using Medicare beneficiary enrollment data.

d. Measures for Descriptive Analyses of Treatment Group Patients

Treatment group patients and their physicians can provide us with detailed information on programs' performance on elements of care coordination that are believed to be key for success. Thus, we will survey the treatment group patients about their perceptions of care coordinators' initial assessments, development of specific care plans with concrete goals, coordination of care, rapport-building, and patient support and advocacy. We will ask their physicians to compare their experiences with patients enrolled in the programs with those with their typical Medicare fee-for-service patients (Table III.10).

2. Costs and Service Use

Medicare costs and service use are among the most critical outcomes for the evaluation. Unless the need for expensive services is reduced, the cost of the intervention will result in a net increase in costs to HCFA. 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 source of any 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. In addition, as explained in Section III.B, our estimates of impacts on hospital use will be much more precise than our estimates of impacts on costs.

Coordinated care may also affect the use and cost of other services. Although we would expect coordinated care to reduce the use of other expensive services, the use of some services could increase if they replace or prevent the need for hospital care. For example, case managers may identify situations in which patients should see a physician, thereby increasing the average number of physician visits and Part B costs. We will estimate impacts on the use and cost of all major Medicare-covered services (hospital, home health care, 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 cost that the evaluation will examine include:

- The probability of receiving various Medicare services
- The amount of Medicare services received

TABLE III.10

MEASURES FOR DESCRIPTIVE ANALYSES OF TREATMENT GROUP PATIENTS AND THEIR PHYSICIANS

Items	Source of Items
Patient Survey	
How respondent heard about the program	Draft
Which program services (if program offers additional services) respondent used (check all that apply)	Draft
Whether respondent knows care coordinator’s name, and how to contact care coordinator during working hours and after hours[a]	Draft
Whether respondent and care coordinator have decided on goals and developed a plan to achieve goals	Draft
Respondent’s reports of other ways in which program was helpful (check all that apply)	Draft
Which issues, of those discussed/addressed in Table III.3 and Tables III.4-9 (such as diet, medication adherence, social support, education, service arrangement, health related behaviors, and so on), did care coordinator discuss or provide help on	Draft
Respondent’s ratings or perceptions of care coordinator’s[a]:	PCAS
Accessibility and responsiveness	
Knowledge of respondent’s medical, emotional, and social problems; values; and goals	
Supportiveness in improving self-care, adhering to plan, and reaching goals	
Helpfulness in arranging needed services and appointments	
Involvement in respondent’s care, thoroughness of monitoring, time spent with respondent, and attention to respondent’s opinion	
Communication with physicians or other health care providers	
Explanations of health problems or treatments symptoms to report and when to seek further care	
Advice and help in making decisions about respondent’s care	
Friendliness, warmth, caring, concern, and respect	
Trustworthiness, honesty, and role as respondent’s advocate	
Respondent’s rating of whether program increased his/her ability to obtain needed care and to take better care of self	Draft
Whether there were elements of the program respondent liked or *dis*liked; if so, which ones	Draft
Whether respondent was able to complain to program staff about problems in the program	Draft
Whether respondent would recommend the program to a friend or family member	HH2

TABLE III.10 *(continued)*

Items	Source of Items
Physician Survey	
Whether respondent:	Draft
Was aware of the program's existence and activities	
Believed the program reduced the burden of caring for enrolled patients and was worth the effort of working with the program	
Believed it improved patients' knowledge of and compliance with medications, diet	
Believed it improved communication between providers and with the patient and family	
Believed it helped respondent keep "on top" of enrolled patients, improved timeliness of patients' followup, and kept them from developing acute exacerbations or complications of their chronic illnesses	
Believed it did not encroach on patients' relationships with their physicians or interfere with respondent's relationships with other physicians, and even felt the program enhanced the patient-physician relationship	
Would recommend program to respondent's other patients, or to respondent's own family and friends	
Especially liked or *dis*liked any features or facets of the program; if so, which ones	

NOTE: We will collect these measures only from treatment group patients and their physicians. All measures will be collected through surveys.

[a] We use "care coordinator" as a generic term. If the program has another name for care coordinator, we will use that name. For programs, such as Q-Med, that do not have care coordinators, we will eliminate inapplicable questions and will substitute the word "program" for "care coordinator" in the remaining questions.

Draft = questions that will be drafted for this survey; PCAS = modified from the Primary Care Assessment Survey (Safran et al. 1998); HH2 = modified from an item in patient survey developed by MPR for the evaluation of the Medicare Home Health Prospective Payment Demonstration, Phase II.

- The cost to Medicare for those services
- The costs of running the intervention
- The net savings to Medicare

We will measure the amount of services used as the number of visits for home health care, physician care, emergency room care, and outpatient services; and the number of admissions and total days of care for hospital, hospice, and SNF care. In addition to measuring impacts on the costs of each type of service, the analysis will also estimate Medicare Part A, Part B, and total costs. All costs will be reported per Medicare-covered month, to control for people who were not covered by Medicare fee-for-service for the full 12-month follow-up period. Table III.11 summarizes these measures of service use and cost, which will be taken from Medicare claims data.

Diagnosis-Specific Measures. The evaluation will also test whether the intervention alters service use and reduces costs for services that are expressly for the target diagnoses. We expect that the interventions will be more likely to influence care related to the target diagnoses, and less likely to influence care related to other diagnoses. However, we will focus primarily on the use and costs for *all* diagnoses, because a true care coordination program should also address patient comorbidities. Furthermore, the diagnosis-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 diagnosis may help shed light on the sources of any cost savings.

Cost-Effectiveness. To assess the cost-effectiveness of coordinated care, the evaluation will measure each site’s net savings per client month. We will create a measure of the intervention cost based on project invoices to HCFA. Based on this variable, we will estimate the program cost per client month while in the program, and the cost per enrollee month over the 12-month

TABLE III.11

ANALYSIS OUTCOME MEASURES FOR MEDICARE-COVERED SERVICE USE AND COST:
ALL DIAGNOSES
(Site A)

	3 Months		6 Months		9 Months		12 Months	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Inpatient Hospital								
Any admission								
Number of admissions								
Number of days								
Reimbursement								
Skilled Nursing Facilities								
Any admission								
Number of admissions								
Number of days								
Reimbursement								
Home Health Care								
Any home health								
Number of visits								
Reimbursement								
Hospice								
Any hospice								
Number of days								
Reimbursement								
Outpatient Hospital								
Any outpatient use								
Reimbursement								
Any emergency room visits								
Number of visits								
Reimbursement								
Physician and Other Part B Services								
Any visits								
Number of visits								
Reimbursement								
Part A Medicare Reimbursement								
Part B Medicare Reimbursement								
Total Medicare Reimbursement								

follow-up period. We will compare these costs with the estimated savings to Medicare per client month over the follow-up period, to estimate the intervention's net savings per client month. Table III.12 provides a sample table shell for these measures of cost effectiveness.[19]

Alternative Estimates of Costs. Due to the high variance of Medicare expenditures across patients, 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 costs. For all services with statistically significant effects, we will develop a range of estimates of the savings associated with the estimated reductions. For example, if hospitalization is the only service use measure for which a statistically significant effect is observed, we will construct a price-weighted service use measure by multiplying an estimate of the average cost of a hospitalization by the number of hospitalizations saved (Table III.13). We will construct a range of estimates, using three estimates of the cost per hospitalization (the lowest 20th percentile, the mean, and the median cost of hospitalizations for the control group). These alternative estimates of savings will provide a range, allowing for the possibility that the hospitalizations "saved" by the intervention would have been relatively less expensive hospital stays than the average. We will compare these estimates of hospital savings with the estimates based on Medicare reimbursement amounts. If the estimates of cost based on service use and the estimates based on Medicare reimbursement amounts differ substantially, we will look for the presence of outliers. A single high-cost outlier could mask savings in a site that actually reduced costs for other beneficiaries.

[19] The distinction between cost per month in the program and cost per month during the follow-up period is important, because enrollees will be discharged or will disenroll. Costs may also vary with the enrollee's length of time in the program.

TABLE III.12

COMPARISON OF INTERVENTION COSTS AND SAVINGS

	Site			
	A	B	C	D
Intervention Costs				
Total Cost (Dollars)				
Case Manager Cost (Dollars)				
Case Manager Cost as Percentage of Total Cost				
Number of Clients Enrolled				
Client Enrollment as Percentage of Target				
Total Cost per Enrolled Client (Dollars)				
Means Months Enrolled per Client				
Total Client Months				
Total Cost per Client Month (Dollars)				
Intervention Savings				
Total Savings per Client Month (Dollars)				
Net Savings (Cost)				

TABLE III.13

SERVICE-WEIGHTED AND REIMBURSEMENT BASED ESTIMATES OF COST SAVINGS

	Estimated Impact on Service Use	Cost per Unit of Service	Service-Weighted Impact on Medicare Costs	Reimbursement-Based Estimated Impact on Medicare Costs
Inpatient Hospital				
Skilled Nursing Facilities				
Home Health Care				
Hospice				
Outpatient Hospital				
Emergency Room				
Physician and Other Part B Services				
Total Medicare Reimbursement				

Although reducing expenditures associated with the most expensive cases is a goal of the intervention, the presence of an outlier could be due to chance alone (for example, a heart transplant case). For this reason, we will estimate impacts by using data on all sample members and then trimming off outliers in the treatment and control groups.

Reconciling Impacts on the Various Outcome Measures. To understand whether a site produced cost savings, we will reconcile the various estimates of impacts on aggregate and service-specific costs and service use. This interpretative analysis will rely primarily on qualitative analysis. We will array the service impact, cost impact, cost impact without outliers, and cost-weighted service impact for each service category for all diagnoses and for the target diagnosis only, as shown in Table III.14. In some sites, estimates for all these outcome measures may provide evidence that the intervention reduced Medicare expenditures, or conversely, that the intervention increased Medicare expenditures. However, we also expect estimates for other sites to produce conflicting evidence. In these cases, we will analyze the array of estimates of the program impacts on costs and service use. We will focus on whether there were statistically significant impacts on service use for the most expensive Medicare-covered services—hospitalizations, SNF stays, and home health care. If the cost estimates are not statistically significant but are sizeable, we will consider the statistical power to detect an effect of the estimated size, and whether there were outliers. We will estimate costs with trimmed outliers, and using price-weighted service-use impact estimates. We will also examine diagnosis-specific measures of service use and cost. If the costs of care for the targeted condition constitute a small share of total Medicare costs, it is possible that we will observe statistically significant effects of the intervention on the disease-specific measures. As we reconcile the various impact estimates, we will draw on the insights gathered in the implementation analysis to assess the plausibility of

TABLE III.14

ANALYSIS OUTCOME MEASURES FOR MEDICARE-COVERED SERVICE USE AND COST:
ALL DIAGNOSES
(Site A)

	Probability of Service Use	Intensity of Service Use	Cost	Cost Without Outliers	Service-Weighted Cost
All Diagnoses					
Inpatient Hospital					
Skilled Nursing Facilities					
Home Health Care					
Hospice					
Outpatient Hospital					
Physician and Other Part B Services					
Part A Medicare Reimbursement					
Part B Medicare Reimbursement					
Total Medicare Reimbursement					
Diagnosis-Specific Care					
Inpatient Hospital					
Skilled Nursing Facilities					
Home Health Care					
Hospice					
Outpatient Hospital					
Physician and Other Part B Services					
Part A Medicare Reimbursement					
Part B Medicare Reimbursement					
Total Medicare Reimbursement					

the alternative estimates. In short, we will rely on various impact estimates and researchers' judgments to assess whether the intervention reduced Medicare costs in each site.

Impacts on the use and cost of non-Medicare services, such as home- and community-based services, Medicaid personal care services, adult day care, nursing home care, and prescription drugs paid for by the beneficiary or others, would indicate whether reductions in Medicare costs may be offset to some extent by increases in other costs. Although these shifts in financial burden may be socially desirable, it is important to identify them. Similarly, it would be useful to estimate impacts of care coordination on the use of free community services and unpaid care provided by enrollees' family members, friends, and community organizations. This information would have to be collected on the patient survey. We will attempt to balance the need for this information with the increased cost and respondent burden that a longer survey would create. We will discuss these tradeoffs with HCFA after drafting the survey instrument.

To detect short-term program effects and to assess the extent to which these impacts persist, we will measure use and cost outcomes over various intervals of time. Some care coordination programs plan to follow patients for a relatively short period, whereas others will monitor patients for the life of the program. Furthermore, even programs that follow patients indefinitely are likely to vary the intensity of their intervention. Thus, estimating the time path and persistence of impacts is important for drawing inferences about the cost effectiveness of the various interventions that the sites are testing.

Outcomes will be measured over the first 2 months after enrollment for use in the first interim report, and for 3, 6, 9, and 12 months after enrollment for the second interim and final reports.[20] Outcomes will be measured separately for each three-month interval, as well as cumulatively for each of the four quarters. For the interim reports, each outcome variable will be constructed only for sample members enrolled early enough to have claims data available for the observation period. For the final synthesis report, one-year follow-up data will be available for all sample members, and 18 months of follow-up data will be available for sample members enrolled during the first year of program intake for programs that start enrolling by October 2001 (month 13 of the evaluation).

D. STATISTICAL METHODOLOGY FOR ESTIMATING EFFECTS

This long section of the design report describes the statistical models that we will use to estimate program impacts and the sensitivity and robustness tests that we will conduct to increase our confidence that the estimates truly reflect program impacts. Throughout the analysis, impacts will be estimated separately for each demonstration site, because the interventions, types of clients, service areas, organizational settings, and practice styles will differ across the sites. Where sample sizes permit (200 or more beneficiaries), we will estimate impacts for subgroups defined by such factors as target diagnosis, prior Medicare expenditure or utilization (for example, top 50 percent of sample), and educational level. Measuring these differences in impacts across beneficiaries is important because (1) estimates of the average program impact over all enrollees could mask important impacts on subsets of the target population, and (2) our findings could suggest more efficient targeting strategies than are practiced by the demonstration sites.

[20] Note that the outcome measures for treatment group members are *not* restricted to those incurred during months enrolled in the program. That is, some treatment group members might voluntarily disenroll from the demonstration before the particular follow-up period under study is completed. These disenrollees are included in the sample because the intervention may continue to affect their service use and costs over time.

Most (if not all) demonstration sites will use random assignment. Although we can simply compare outcomes of the treatment and control groups in order to estimate program impacts in these sites, we will use regression models because regression analysis produces more precise impact estimates and eliminates any bias due to chance baseline differences between the two groups or to differential attrition or nonresponse. Regression analysis will also be used to estimate impacts in sites using a comparison design, and to test the comparability of random assignment and a comparison design in sites using random assignment.

We will conduct some analyses of claims-based outcome measures using only control variables constructed from claims data. These analyses will require different models, because the control variables will be limited to what is available from claims. Sample sizes will be larger for the claims-based analyses due to survey nonresponse, item nonresponse, and (in five sites) enrollment levels that exceed the survey sample sizes. In comparison sites, we will choose the survey sample on a rolling basis and expect to be able to draw a larger, better-matched sample for the claims analysis later in the evaluation.

The next three sections describe the regression and statistical models we will use in sites with random assignment to estimate impacts on basic outcome measures (Section D.1), expenditures (Section D.2), and subgroups of key interest (Section D.3). Section D.4 describes the models for estimating impacts for any demonstration programs that will not use random assignment, and that must rely instead on a comparison group design. These methods will also be used for the sensitivity tests we will conduct on the random assignment sites, replicating a comparison design approach there in order to test the ability of that approach to generate estimates similar to those from the randomized design. Section D.5 describes how we will try to estimate the portion of the overall impact on the final outcomes (costs, use of expensive services, and patient well-being) resulting from program impacts on patient behavior and other intermediate outcomes. Section D.6 discusses how we will conduct the statistical tests of the many hypotheses about program effects, and Section D.7 presents various sensitivity tests we will conduct on the comparison design estimates, as we are less confident of obtaining unbiased estimates with that design.

1. Regression Models

Regressions will be used to estimate the intervention's impact on various intermediate and final impacts. The appropriate method for estimating impact models depends on the form of the dependent variable.

a. Ordinary Least Squares

If the dependent variable is continuous (for example, Medicare cost per month), linear regression techniques can be used. The ordinary least squares (OLS) equation to be used for the regression models on continuous dependent variables is

(1) $Y_i = a_0 + a_1 T_i + \Sigma a_j X_{ji} + e_i$,

where Y_i is the outcome measure of interest for the *i*th individual, T_i is a treatment status indicator (a binary variable equal to one if patient *i* is a member of the treatment group and zero if he or she is a member of the control/comparison group, X_{ji} is a set of *j* individual background characteristics (such as age, gender, past Medicare service use and expenditures), and e_i is a random disturbance term. Under this simple specification, a_1 estimates the impact of the demonstration on outcome $Y_{i.}$

b. Logistic Regression

If the dependent variable is binary (such as whether or not a patient received a particular preventive care procedure or test), the model will be estimated using logistic regression. If the dependent variable is a count variable, such as the number of hospitalizations after random assignment, or an ordered scale, such as self-reported health status (poor, fair, good, excellent), then an ordered logit model will be used.[21] Logit models for binary dependent variables are based on an assumed distribution for the error term *e* and the following framework:

(2) $Y_{\mathrm{i}}^{*} = a_0 + a_1T_i + \Sigma a_jX_{ji} + e_i,$

(3) $Y_i = 1 \text{ if } Y_i^{*} > 0\ [e_i \geq (a_0 + a_1T_i + \Sigma a_jX_{ji})],\ Y_i = 0 \text{ if } Y_i^{*} \leq 0,$

where Y_i^{*} is an observed propensity to exhibit the behavior in question (for example, to have a hospitalization). The propensity is determined by observable factors (T_i, X_{ji}) and unobservable factors (e_i). When this propensity exceeds a threshold (arbitrarily set equal to zero here), the patient is observed to have $Y_i = 1$. The logit model is based on the assumption that e_i has an extreme value distribution, which leads to the following expression for the probability that $Y_i = 1$:

(4) $P\ (Y_i = 1|\ T_i, X_{ji}) = 1/[1 + exp(a_0 + a_1T_i + \Sigma a_jX_{ji})].$

In contrast to the OLS regression model shown in equation (1), the coefficient a_1 in equation (4) does not provide a direct estimate of the size of impact of coordinated care on the probability that $Y_i = 1$, but the statistical significance of this coefficient indicates whether the effect on the odds that $Y_i = 1$ is statistically significant. Overall impact estimates are obtained by computing, for each sample member, the difference between the predicted probability that $Y = 1$ when T is set equal to one and when T is set equal to zero. The average of these differences in predicted probabilities yields the overall impact estimate.

[21] Poisson models could be used to estimate effects on count variables, but they sometimes generate anomalous estimates.

c. Hazards Model

To account for the fact that sample members will be observed for different lengths of time, we will also use event-history or "hazard" models for binary outcome measures. These models provide unbiased estimates of program effects on binary outcomes when patients' data are truncated, limiting the length of follow-up period we observe. Truncation will arise because enrollees die, move out of the area, or enroll in Medicare + Choice managed care plans. In addition, data on some patients used for the interim analyses will be truncated because the patients will have enrolled later in the demonstration and we will not be able to observe their full followup. The Cox proportional hazards model will allow us to obtain unbiased estimates of program effects on the length of time until events occur, despite the presence of truncation.

The Cox proportional hazards model is written as follows:

(5) $log\ h_i(t) = a_0 t + a_1 T_i + \Sigma a_j X_{ji} + e_i$.

The data we use to estimate the hazard model consists of *one observation per individual per month*. The dependent variable $h_i(t)$ is a binary variable that equals zero in months when an individual is still in the sample and has not yet experienced the event, and that equals one in the month when the individual experienced the event. For example, suppose the event of interest is a hospitalization, and an individual had a hospitalization in month 4 after enrollment. This individual would contribute four observations to the data set—the dependent variable would be zero during the first three months and one during the fourth month. If an individual moved to managed care at the start of month 6 (and is thus right-censored) and had not experienced a hospitalization before then, he or she would contribute five observations to the data set. In this case, the dependent variable would be zero for all five observations. The control variable t is a measure of time (number of months since enrollment). Similar to the OLS specification shown in equation (1), a_1 here estimates the proportional impact of the demonstration on the probability that an event occurs in any month, given that it has not yet occurred. The model can also be used to derive an estimated effect on the time until an event occurs or the probability that it occurs within a given interval.

2. Two-Part Models for Estimating Impacts on Costs

The models for estimating impacts on costs deal with both the skewed distribution of Medicare expenditures and the presence of a large number of people with no expenditures for particular service categories. To account for skewness in the right tail of the distribution and the presence of outliers, we will transform dependent variables that measure Medicare costs, *Y*. Two types of transformations will be explored: (1) log *Y***,** and (2) square root (*Y*). These transformations have been shown to eliminate the undesirable skewness in the distribution, producing more efficient estimates of impacts. Inferences about whether a regressor has a significant impact on the dependent variable can be drawn directly from the estimated model based on the transformed (log-dollar or square-root-dollar) scale. However, predictions must be made on the actual expense (dollar) scale, rather than on the transformed scale. To obtain estimates on the dollar scale, we will reverse the transformation of the dependent variable, taking the exponential when the log transformation was used, and squaring when the square root transformation was used. Following Duan et al. (1982) and Manning (1998), we will adjust the re-transformed predicted outcome with a "smearing factor" to provide predicted costs that have smaller mean squared errors than do estimates from regression models.

The two-part model for estimating impacts on costs is designed to reflect the presence of beneficiaries with no expenditures in particular service categories, as well as skewness in the expenditure amount. For example, many beneficiaries will not have costs for hospitalizations, SNFs, home health, or outpatient care. In these cases, the dependent variable is truncated at zero, so OLS is not the most appropriate statistical model for these services. Instead, we will estimate a two-part model for the costs of every service that some sample members do not use (that is, for which they have no expenditures). We do not expect to use this model for total Medicare costs, because the demonstrations are likely to target beneficiaries all of whom will have at least some Medicare costs.

The two-part model estimates whether a patient has any Medicare expenditures and the cost of these expenditures (Duan et al. 1982). The first equation, sometimes referred to as the hurdle equation, predicts the probability that any expenditures occur, given the control variables and whether the beneficiary is a treatment group member, $[P(Y > 0)|X, T]$, using a probit or a logit model such as that specified in equation (4). The second equation, called the levels equation, then estimates the amount of the expenditure conditional on there being a positive expenditure, using the same control variables, $E[Y|Y>0,X,T]$. The second equation is the OLS equation (1), estimated using only the sample that has positive expenditures. Because of the skewness of expenditures, Y will be transformed in this equation into either the logarithm or the square-root-of costs, as described above.

To obtain the impact estimate, we first estimate each sample member's predicted expenditures by multiplying the predicted probability of having an expense by the estimated conditional expectation of expenditures for each individual (using the smearing factor described above). We then sum the predicted expenditures for all treatment group members and subtract from it the sum for all controls. The difference is the estimate of the impact of the intervention on expenditures.

3. Estimating Subgroup Impacts

If sample sizes permit, we will estimate the intervention's effects for a few key subgroups of beneficiaries, while bearing in mind that there will be substantially less power to detect impacts of any given size. Subgroup status will be represented by indicator variables. Interaction terms (the product of the treatment status indicator, *T,* and the variables defining the subgroups, X_j^ss will be added to the regression models and the augmented models will be estimated:

$$(6)\quad Y_i = a_0 + a_1 T_i + \sum_j a_{2j}\, T_i X_{ji}^s + \sum_j a_j\, X_{ji} + e_i\,.$$

The coefficient a_{2j} on the *j*th interaction term will measure the difference in program effects between those with the subgroup characteristic and those without it. This approach provides unbiased estimates of the actual effect of a given characteristic on program impacts. The estimated models will then be used to generate estimates of impacts for various subgroups, controlling for other characteristics. These estimates will be generated by calculating the mean for the interacted variables (X_j^ss) over the subgroup for whom impacts are bring estimated (those with $X_{ki}^s = 1$) and inserting them into the expression for the subgroup impacts:

$$(7)\ \textit{impact for subgroup}\ (X_j^S = 1) = a_1 + a_{2j} + \sum_{k \neq j} a_{2k} \overline{X_k^S},$$

where mean X_k^s is the mean value for interacted variable X_k^s calculated over the cases for which $X_{ki}^s = 1$.

We will most likely analyze impacts on subgroups defined by claims-based or intake-based variables in sites with larger enrollments, and in models estimated on data pooled across the sites. We will undertake little or no site-specific subgroup analyses in sites that have low enrollment, because our tests for these effects will have little power given the available sample sizes. Similarly, we probably will not be able to estimate subgroup impacts using survey-based measures. Key subgroups we will analyze when possible include:

- Whether discharged from the hospital within one week of enrollment
- Education
- Age
- Prior year costs
- Stage of disease

In another subgroup analysis, we will compare impacts for the early cohort of enrollees in each site with those for later enrollees in that site, to determine whether the program's effectiveness increased over time. The sample will be split into the half that enrolled earliest and the remaining half. If some sites experience any site-specific changes that we believe are likely to influence impacts, we may define alternative calendar points at which to define subgroups.

4. Models for Estimating Impacts with an External Comparison Group

The regressions we use to estimate impacts with a comparison design approach will differ somewhat from the regressions presented in the preceding section, which are appropriate for random assignment sites. The comparison design regressions allow for the possibility that the comparison group and the treatment group are not well matched on unobservable characteristics. Thus, a number of sensitivity tests will be conducted to assess the robustness of our estimates. These tests and samples are described in Table III.15. The remainder of this section discusses the basic approach used to estimate impacts with a comparison site approach and the sensitivity tests to be undertaken. The reader may wish to refer back to the table after reading this section.

TABLE III.15

COMPARISON STRATEGIES TO ESTIMATE IMPACTS AND TEST ROBUSTNESS

	Random Assignment		Comparison Group Design		
	Survey Measures	Claims Measures	Survey Measures	Claims Measures	Rationale
Basic Approach					
Survey Respondents	T–C	T–C	$(E_{DS} - E_{CS})/p$	$(E_{DS} - E_{CS})/p$	Basic estimates
Sensitivity Tests[a]					
Survey Sample	—	T–C	—	$(E_{DS} - E_{CS})/p$	Assess nonresponse bias
All Demonstration Cases[b]	—	T–C	—	$(E_{DS} - E_{CS})/p$	Assess survey representativeness
All Eligibles		$(E_{DS} - E_{CS})/p$	—	$(E_{DS} - E_{CS})/p$	Assess comparison methodology
Propensity-Score-Matched Cases		T–MC	—	T–MC	Assess comparison method
All Eligibles, Alternative Comparison Sites (E_{CS}^{*})	—	—	—	$(E_{DS} - E_{cs}^{*})/p$	Assess comparison site sensitivity
Pooled Pre-demonstration and Demonstration Eligibles	—	$[(E_{DS} - E_{CS}) - (E_{DS}^{O} - E_{CS}^{O})]/p$		$[(E_{DS} - E_{CS}) - (E_{DS}^{O} - E_{CS}^{O})]/p$	Assess bias in comparison design
All Eligibles in Demonstration Site Only (Heckman Model)	—	$T - E_{NP}/\lambda$		$T - E_{NP}/\lambda$	Assess validity of MC
Alternative Eligibility Group		$(\hat{E}_{DS} - \hat{E}_{CS})/\hat{p}$		$(\hat{E}_{DS} - \hat{E}_{CS})/\hat{p}$	Increase precision of comparison site estimates

NOTE: All estimates will be obtained using regression analyses. For outcome measures obtained from the survey, control variables will be drawn from the survey and intake form, as well as from Medicare files. No survey or intake form control variables will be available for claims-based outcomes unless the sample is restricted to the survey sample (intake form data available) or survey respondents.

[a] All sensitivity tests will be performed with only a limited set of key outcome measures from claims data.

[b] For sites with more demonstration enrollees than needed for survey sample.

T = treatment group; C = control group; E_{DS} = eligibles in demonstration sites; E_{CS} = eligibles in comparison sites; p = participation rate among eligibles; E_{CS}^{*} = eligibles in alternative comparison site; MC = matched comparison group (propensity score approach); $(E_{DS}^{O} - E_{CS}^{O})$ = predemonstration period eligibles in demonstration and comparison sites; E_{NP}/λ = eligible nonparticipants in demonstration site, adjusting for selection bias; $\hat{E}_{DS}, \hat{E}_{CS}$ = eligibles in demonstration and comparison sites using more restrictive eligibility definition; and $\hat{p}$ = participation rate among those meeting more restrictive eligibility definition.

a. Regression Models

The following equation predicts outcomes for all eligibles in a demonstration site (that is, participants and nonparticipants) and for all eligibles in the comparison site:

(8) $Y = a_0 + a_T T + a_{NP} NP + \Sigma a_j X_j + e,$

where T is equal to one when the individual is a treatment group member, NP is equal to one when the individual is a nonparticipating eligible, and T and NP equal zero otherwise (that is, in comparison cases, $T = NP = 0$). We can estimate the impact on all *eligibles* in the demonstration site by taking the weighted average of the coefficients on the participants and nonparticipants, where the weights are equal to the proportion of the total eligibles in the treatment site represented by each group:

(9) $\textit{Impact on eligibles} = a_T P_T + a_{NP}(1 - P_T)$.

Here, P_T is the participation rate, or the proportion of all eligibles in the demonstration site who actually joined the treatment group, and $(1 - P_T)$ represents the proportion of all eligibles in the demonstration site who did not participate.

Dividing the result obtained in equation (9) by the participation rate in the treatment site gives an estimate of the impact per participant:

(10) $\textit{Impact on participant} = [a_T P_T + a_{NP}(1 - P_T)]/P_T = a_T + a_{NP}(1 - P_T)/P_T$.

The coefficient a_{NP} should not be statistically significant if three conditions are met: (1) those who enroll are similar to eligible nonparticipants on unobserved characteristics that influence outcomes, (2) the comparison site is well-matched to the demonstration site, and (3) the demonstration has no spillover effects on nonparticipants. We will match the sites carefully and expect very little or no contamination; thus a finding that the coefficient a_{NP} is significant suggests that selection bias may be present. If a_{NP} is not statistically significant, the coefficient a_T measures the impact of the intervention. If a_{NP} is statistically significant, a_T alone will misestimate the effects of the program by $bias = a_{NP}(1 - P_T)/P_T$. We can see that the bias approaches zero as the proportion of eligibles participating approaches one, and that it increases as the participation rate approaches zero. We assume there is no program impact on the external comparison group or the eligible nonparticipants.[22]

We will test the sensitivity of these impact estimates to the choice of the comparison group and to the model used to estimate impacts. We will estimate models using three comparison groups: (1) the comparison group drawn for the survey sample; (2) a group drawn after enrollment has been completed, using updated claims data and a propensity-score approach; and (3) a group drawn using updated claims data and a propensity-score approach from a second external comparison area. Under the two propensity-score approaches, we will use models similar to those used for the random assignment sites; we will compare participants with the selected comparison group directly, rather than compare all eligibles in the demonstration area with all eligibles in the comparison area and dividing by the participation rate. For each sample,

[22] After this model has been estimated, if a_{NP} is sizable we will reestimate the model with *T* and *NP* variables combined into a single variable set equal to one for all eligibles in the demonstration site and set equal to zero for all comparison site cases. The coefficient on this variable, divided by the participation rate, yields an unbiased estimate of program effects, as does the approach described above, but with smaller variance (because only one parameter is being estimated, instead of two.) We estimate the separate coefficients model first, however, to assess the size of the bias that would exist by simply comparing the participants with the comparison group. If a_{NP} is essentially zero, we will reestimate the model with *NP* excluded (that is, comparing participants with the combined group of nonparticipants and the comparison group), to yield substantially more precise estimates than those available by comparing eligibles and eligibles and dividing by the participation rate.

we will describe the effect size we would expect to detect given the participation rate and the sample size.

b. Propensity Score Approach to Selecting a Comparison Group

We will use a propensity score approach to draw a comparison group that matches the participants (rather than all program site eligibles) as closely as possible. We expect this comparison group to be better matched than the survey-based comparison group because we will be able to use more up-to-date claims data and the more sophisticated propensity score approach to ensure that, on average, the treatment and comparison groups are similar.

The following overview describes the propensity score approach. First, we will estimate a model using participants and eligible nonparticipants to determine how each characteristic that affects outcomes also affects the decision to participate. Second, based on this information, we will assign to each actual participant and each eligible beneficiary in the comparison area a propensity score that summarizes how that individual's characteristics affect the decision to participate. Finally, for each participant, we will select a comparison group member with a similar propensity score.[23]

More specifically, selecting a comparison group of simulated participants using the propensity score method consists of the following steps:

- Collect demographic, health status, and preintervention outcome data for participants, eligible nonparticipants, and the pool of potential comparison group members.

[23] This is actually a modification of the usual propensity score approach. The typical approach tries to draw a comparison sample when no data on eligible nonparticipants are available. The "participation" logit mode is then estimated on participants and on the comparison group. The approach we propose is stronger, if the comparison area is well-matched.

- Code an indicator variable equal to one for each participant and equal to zero for each eligible nonparticipant. Call this indicator variable *P*.

- Define indicator and continuous variables that represent the demographics, health status, and preintervention outcomes of participants, eligible nonparticipants, and potential comparison group members. Call this collection of variables *X*.

- Using participants and eligible nonparticipants, estimate a probability model, such as a logit or probit model, where the dependent variable is *P* and the set of independent variables is *X*. Results from the probability model will include parameter estimates, or a collection of values that indicate how each respective *X* affects *P*. Call this collection of values *BETA*.

- For each participant and potential comparison group member, define a variable that equals the sum of each *BETA* value times each respective *X* value. Call this variable *P**. *P** indicates each individual's propensity score.

- For each participant, select a potential comparison group member with the closest absolute *P** value. This selection process should be done with replacement so that a comparison group member may be matched to more than one participant.

- Selected comparison group members define the comparison group of simulated participants.

A side-by-side comparison of the characteristics of each participant and respective simulated participant is likely to indicate that the two differ with respect to specific *X* values, or characteristics. These differences are acceptable as long as the *X* values, or characteristics of participants and the comparison group of simulated participants, are similar *on average*.

The next-to-last step in the list—to select the simulated participants from the pool of all eligibles in the comparison area *with* replacement—is worth emphasizing. Research has shown that impacts based on a comparison group selected with replacement can be similar to those random assignment would produce, whereas impacts based on a comparison group selected without replacement are likely to be different (Dehejia and Wahba 1998 and 1999). Selecting with replacement is especially useful in situations that have few similar potential comparison group members. We are unlikely to have such a problem in our analyses, because the comparison group (unlike the nonparticipating eligibles) should have a sizeable number of cases with high predicted probabilities of participation.

c. Selection Models

If the estimates across the different comparison groups are similar, and the estimates from the comparison design approach are similar to those from the randomized design in the sites where the two approaches were compared, we will have greater confidence that they reflect true impacts. If the estimates are not robust, however, we will explore various models for estimating impacts when the treatment group is self-selected. We will use the approach developed by Heckman (1976) to estimate models when individuals self-select into the program (or other behavior) being studied. Comparing outcomes for participants and outcomes for eligible nonparticipants yields a biased estimate of program effects, because those who choose to participate may differ from those who do not on unobserved characteristics that affect the outcome of interest. For example, it is likely that beneficiaries who are very interested in playing an active part in improving their health and are able to do so will have better outcomes than other beneficiaries with the same illness, whether they participate in the demonstration or not. Thus, comparing outcomes for participants and nonparticipants would confound this difference with any effects of the program, likely leading to overestimates of program effectiveness.

The model developed by Heckman involves estimating a participation equation using eligible beneficiaries in the demonstration site, then using the estimated model to construct a new variable, called “lambda” in the literature. The model assumes that the error terms in the participation and outcome equations are bivariate normal. Under this assumption, including this variable as an additional control variable in a model such as equation (1) that compares outcomes for participants and outcomes for eligible nonparticipants eliminates (asymptotically) the bias in the coefficient on *T*. The constructed variable is proportional to the conditional expectation of the error term in the outcome equation, given that the individual chose to participate (or to not participate) in the demonstration. Including this variable directly in the model eliminates the correlation between the participation variable *T* and the remaining error term, and, therefore, the bias in the estimated impact (the coefficient on *T*).

Estimating this model requires identification of one or more variables that are likely to influence the probability of participation but are not likely to directly influence the outcomes of interest. For example, we will construct a measure of each beneficiary's provider's exposure to the program, defined by the proportion of the physician's eligible patients who participate in the program. We will also try to eliminate bias by collecting in the survey variables that could be included in the model to eliminate possible correlation between the error terms. For example, we could include a survey variable asking beneficiaries how likely they would be to participate in an experimental cancer trial if they were diagnosed with that disease. If these models are necessary, we also will consider including the proportion of eligible beneficiaries in the sample members' county or ZIP code that participates in the demonstration. We will revisit the identification of these control variables after determining whether any of the sites will require a comparison group design.

d. Difference-in-Difference Models

Although we will select comparison sites to match the demonstration sites as well as possible on predemonstration service use outcomes for the target population, sizable predemonstration differences on some outcomes may remain. These differences can create biases in a comparison site design because observed differences between the two groups of eligibles during the demonstration period may merely reflect preexisting differences between the sites in practice patterns or other factors. To adjust for potential biases, the evaluation will use a difference-in-difference approach to estimate impacts on claims-based outcome measures. The difference-in-difference approach estimates the program impact by comparing the difference in the outcomes of all eligibles in the demonstration and comparison sites after the intervention with the difference in the outcomes of all eligibles in the demonstration and comparison sites before the intervention. Note that the group of eligibles in the two time periods will differ because eligibility is likely to be based on recent service use and on being in fee-for-service Medicare during the relevant time period.

(11) $Y_{\text{Eligible in Demonstration Area, Postintervention}} - Y_{\text{Eligible in Demonstration Area, Preintervention}}) -$
$(Y_{\text{Eligible in Comparison Area, Postintervention}} - Y_{\text{Eligible in Comparison Area, Preintervention}}).$

The regression model used to estimate the impact of the intervention is:

(12) $Y = a_0 + a_1E_D + a_2POST + a_3\, E_D\, POST + \Sigma a_jX_j + e,$

where E_D is equal to one if the eligible beneficiary resides in the demonstration site (and is equal to zero for comparison site eligibles), and *POST* is equal to one if the observation is for the postintervention period (and is zero for preintervention observations). In this specification, the impact on participants is a_3/P_T, where P_T is the participation rate among eligibles in the demonstration site.

e. Reconciliation of Estimates

One of the major challenges for the impact analysis in the site using a comparison group design will be to reconcile the various impact estimates. In addition to reconciling estimates on various related outcome measures, such as service use and costs, we will also compare the impact estimates generated for various alternative comparison groups and for different estimation techniques.

5. Linking Final and Intermediate Outcomes

We will want to better understand the likely source of favorable program impacts on key outcomes. Therefore, in both random assignment sites and comparison design sites, we will examine whether intermediate outcomes are linked to final outcomes, including patient well-being, service use, and costs. Intermediate outcomes include any outcomes that demonstration sites target as a mechanism to improve the final outcomes of cost and quality of care. Examples of intermediate outcomes sites may seek to include are patients' health-related behaviors (smoking, exercise, weight control, and adherence to medication and diet); patients' knowledge of their disease; unmet needs for social services; and receipt of preventive clinical interventions, such as influenza vaccinations and health screening tests relevant to their condition. This part of the analysis will answer three questions: (1) What intermediate outcomes does the site seek to achieve? (2) Do these intermediate outcomes actually affect final outcomes? and (3) Do treatment members experience better intermediate outcomes than control group members when the intermediate outcome is a goal of the demonstration? To test these hypotheses, we estimate equation (1) with Medicare expenditures as the dependent variable, and add in a set of *k* control variables measuring a patient's intermediate outcomes in areas that the site targets, and a set of *k* terms interacting treatment status and the various intermediate outcomes:

$$(13)\quad Y_i = a_0 + a_1T_i + \Sigma b_k INTERMEDIATE\ OUTCOME_{ki} + \Sigma a_jX_{ji} + e.$$

In this specification, each of the b_k coefficients on the vector of intermediate outcomes test whether the specific intermediate outcome affects the final outcome. If b_k is statistically significant and shows that intermediate outcomes are associated with more favorable final outcomes, and the program has improved intermediate outcomes, then the coefficient a_1 will be smaller (in absolute value) than it is in models that do not control for intermediate outcomes because some of the observed impact on Y is the result of the improvements in intermediate outcomes. From this equation, the overall impact of the program can be shown to equal $a_1 + \Sigma b_k \triangle INTERMEDIATE\ OUTCOME_k$, where $\triangle INTERMEDIATE\ OUTCOME_k$ is the program impact on the kth intermediate outcome. The coefficient a_1 captures any effects of the intervention that do not result from improvements in intermediate outcomes.

6. Testing Strategy

We will use a standard set of procedures and significance levels to test the numerous hypotheses considered in the evaluation. Most of the tests of hypotheses about the existence of overall program effects will be two-tailed tests of whether the coefficient on treatment status in our models is significantly different from zero, using a 0.10 significance level. We believe that care coordination most likely will reduce costs and improve quality, but impacts in the opposite direction are possible. For example, care coordination might encourage patients to obtain additional services or might reduce their satisfaction. Thus, we will conduct two-tailed tests for nearly all the hypotheses, limiting the one-tailed tests to the few outcomes for which the only possible impact is in one direction. The use of the 0.10 significance level corresponds to a 0.05 level for a one-tailed test and is used here instead of a more stringent 0.05 level because sample sizes are smaller, and we do not want to overlook important program effects. Our final assessment of whether a statistically significant difference is plausible evidence of a true program effect or a statistical anomaly will be based on examination of related outcomes. We will also indicate whether impact estimates would be statistically significant at even smaller significance levels.

If any sites decide to randomize interested physicians instead of patients (none have suggested doing so in their proposals), we will estimate standard errors for impacts that adjust for clustering of patients among physicians. Clustering of patient observations may lead to understatement of the standard errors because the variation due to the physician effect is not fully reflected when the sample of patients is treated as if it were randomly selected from the universe of patients. We will use a statistical program called SUDAAN to calculate approximate standard errors that take this design effect into account. If it is necessary to use SUDAAN, we will do so for a few key outcome measures in each substantive area of our analyses and will calculate the design effect for the other measures by computing the ratio of the estimated standard error from SUDAAN to the estimated standard error for the conventional regression models. We will use the average of these ratios to adjust our standard errors and *t*-statistics for all hypothesis tests.

7. Sensitivity Tests

We will perform tests of the robustness of our estimates. Because all but one of the sites propose to do random assignment, the results are expected to be very robust. Nonetheless, a few situations warrant sensitivity tests. For example, if nonresponse rates are high or differ markedly for the treatment and control groups, we will assess the effect of nonresponse bias on our estimates of impacts on survey-based outcome measures. We will do this by comparing estimated impacts on claims-based outcome measures (excluding from the models any control variables obtained from the survey) obtained on the full survey sample with those estimated on only the survey sample *respondents*. If estimates on these claims-based outcomes are similar for the full and censored samples, we can be reasonably confident that estimates of impacts on the survey-based measures are not biased by sample attrition. Similarly, we will test the representativeness of the survey samples in the few large sites by comparing impacts on claims-based outcomes estimated over all treatment and control group members with impacts estimated only on the survey sample.

Other tests of the robustness of our estimates will include examination of the effects of outliers on our impact estimates, checks for consistency between cost and utilization impact estimates (discussed in Section II.C.2), and comparison of a site's impact estimates from a random assignment design with those from a comparison group design. Examining the sensitivity of the findings to outliers for continuous dependent variables, such as costs, is essential with samples of this size. We will perform this assessment in a variety of ways, including "trimming" extreme values in both the treatment and comparison groups, using functional forms that minimize the effects of outliers (for example, square root transformations), and examining treatment-comparison group differences in the distributions as well as in the means of the outcomes.

The comparison of impacts measured by treatment-control differences in outcomes with impacts estimated from differences between eligibles in the program site and in an external comparison site will provide evidence on whether the estimates from the comparison group approach are as reliable as the estimates from a randomized design. This check will indicate how accurate the estimates for the site(s) that use a comparison design are. Furthermore, assessing the comparability of estimates from random assignment and a comparison design will be useful if HCFA wants to use a comparison design approach to monitor cost effectiveness after the demonstration ends. Differences will also provide valuable guidance on the plausibility of published estimates that use comparison group designs to estimate the effects of coordinated care programs, as well as on how such approaches can be improved.

Finally, we will estimate the extent of contamination of the control group, one of the biggest potential threats to the evaluation, by comparing outcomes for control group patients whose physicians have many treatment group patients with outcomes for control group patients whose physicians have few or no treatment group patients. The most likely source of contamination in the demonstration programs, if any, will be from physicians with patients in both groups who change their practices for *all* their patients as a result of the intervention. For example, programs may distribute protocols or guidelines (for example, on medication dosage) that these physicians adopt universally, or they may institute reminder systems on educating patients or increasing patient adherence that the physicians begin to use for all their patients. Alternatively, a physician may observe treatment group members receiving helpful community-based services arranged by the care coordinator and may then decide to refer his or her control group patients to local organizations for such services.

To test for contamination, we would include in the regression model used to estimate impacts a binary variable for whether the control group member had a physician with (say) five or more patients in the treatment groups.[24] A finding that the coefficient on this variable is significantly different from zero suggests that outcomes are different for these control group patients than for those whose physicians had little or no contact with the demonstration, suggesting possible contamination. We would also control in this model for the total number of patients that a patient's PCP had enrolled in the study. This variable would account for differences in patient outcomes that may be due to some physicians having more patients with the target condition than do other physicians; the former group of physicians may provide better care regardless of the intervention. It would also account for differences in outcomes occurring

[24] The cut-off point will be determined after examining the sample distribution of physicians by the number of patients in the two groups. A continuous measure, such as the number or percentage of the physician's patients in the evaluation sample who are assigned to the treatment group, could also be used as an indicator of the risk of contamination.

because physicians with more demonstration patients were better physicians than were those who refer fewer patients to the program. This estimate of the extent of contamination bias should be relatively unbiased itself, as the number of a given physician's patients assigned to the treatment groups will be determined by random assignment.

E. CONTROL VARIABLES FOR IMPACT ANALYSIS

The set of explanatory variables in the models used to control for preexisting differences between the treatment and comparison groups will depend on the sample used, the program examined, and, to a limited degree, the specific outcome measure being estimated. For analyses conducted on all eligibles in the treatment area, the set of independent variables is limited to the variables that we can construct from Medicare claims and enrollment data. Control variables will vary across program sites because different programs may collect different data on their intake forms. In general, control variables will measure patient demographics, prior Medicare use and expenditures, comorbidities, and complexity of illness. These factors may influence patients' Medicare service use and costs; therefore, we must control for them, because differences between the treatment and the control groups may arise by chance or due to differential patterns of survey nonresponse. Some of the variables may also be used to define subgroups, if they are deemed likely to influence patients' ability to benefit from the intervention.

Table III.16 lists the control variables and their sources. ***Demographic and socioeconomic characteristics,*** including age, sex, race, original reason for Medicare eligibility (age or disability), and whether Medicaid pays the beneficiary's Part B premium, will be taken from the Medicare EDB; education, income, living arrangements, smoking and drinking behaviors prior to the demonstration, and care-seeking attitudes will be drawn from the patient survey; and

TABLE III.16

CONTROL VARIABLES AND THEIR SOURCE
(Site A)

Medicare Enrollment Data Base
- Age
- Sex
- Race
- Original reason for Medicare entitlement (age or disability)
- Date of death
- HMO enrollment

Patient Survey
- Education
- Income
- Living arrangements
- Predemonstration smoking and drinking behavior
- Care-seeking attitudes
- First language other than English

Intake Form (When Available)
- Diagnosis
- Severity-of-illness measures
- Reading level
- Health behaviors
- Referral source

Medicare Claims (Standard Analytic Files)
- Diagnoses
- Use and cost during preenrollment year
- Number of different physicians billed in preenrollment year
- Rehospitalization rate for all eligible patients seen by the patient's provider in year prior to demonstration start

Medicare Health Insurance Skeleton Eligibility Write-Off (HISKEW) File
- Medicare eligibility

diagnosis will be drawn from the patient intake (consent) form. Any additional site-specific data collected on the intake form, such as severity-of-illness indicators, reading level, or health behaviors, will be considered for inclusion as control variables in site-specific analyses. We may include a set of control variables measuring a patient's referral source to account for the fact that sites may use different referral sources to enroll patients. This variable may be available only for random assignment sites.

We will also control for ***Medicare expenditures and utilization*** in the year prior to enrollment because they are good indicators of health status and are the best predictors of future Medicare costs and service use. These use and cost data will be drawn from Medicare claims data contained in the SAFs and will be broken down by service type (emergency and nonemergency hospitalization, SNF, hospice, physician, emergency room, other services, and home health care). In addition, measures of total Part A and Part B reimbursements and number of days since the last inpatient admission may be used. Variables measuring use and expenditures cannot be accurately constructed for beneficiaries who have not been entitled to Medicare for a full year preceding enrollment in the demonstration or for those who were in HMOs for a portion of that year. For these beneficiaries, we will calculate an annual equivalent by multiplying service use or expenditures per month of Medicare eligibility by 12.[25]

Regressions will control for patient comorbidities and the number of different physicians billed in the year preceding enrollment. These can be constructed from Medicare claims. They represent a rough proxy for the complexity of medical needs and the intervention's potential to

[25] We do not expect many beneficiaries to be newly enrolled in Medicare (and therefore to lack an expenditure history), because most programs will target beneficiaries based on past service use. We will include a dummy variable for any new beneficiaries to indicate that these individuals did not have prior expenditures.

have an effect on a patient. As a proxy for a beneficiary's ability to communicate with providers, regressions for the survey sample will control for whether the beneficiary reports English as a first language. If the treatment and comparison/control group members are drawn from different geographic areas or networks of providers, we will control for hospital and area characteristics that are likely to influence cost and service use patterns. For example, in regressions estimating the probability of rehospitalization during the year after enrollment, we will add a control variable measuring the probability of rehospitalization for *all* eligible patients who had the same provider as the sample member during the year prior to the demonstration. Area- and provider-level controls will be included only when members of the treatment or comparison group come from areas or had providers with different preintervention utilization and cost patterns before the demonstration began.

We will also test the sensitivity of the impact estimates to patient deaths during the time interval examined. Including an indicator for whether the patient died during that interval will control for treatment-control differences in Medicare costs that could arise from differences in mortality rates. Given the high average costs Medicare beneficiaries incur during their last months of life, differences in mortality rates could be associated with substantial differences in Medicare costs. Although one could argue that death is endogenous (that is, it may be influenced by the coordinated care intervention being tested), the potential bias resulting from failure to control for exogenous differences in deaths may be a greater concern than the potential simultaneity from including death as a control variable. We will first test for whether the program affects mortality. If it does not, we will include whether the patient died as a regressor and will examine the sensitivity of our estimates to the inclusion of this variable. If the difference in mortality rates *is* statistically significant, we will try to determine whether the estimated treatment-control difference in mortality rates is likely a result of the intervention.

This determination will help us assess whether controlling for mortality biases our estimates or corrects them for chance differences between the groups.

Our main emphasis will be to estimate impacts, but we may also examine the sources of variation in the treatment group's outcomes. For these analyses, we will control for the amount and type of demonstration services received by each participant. However, this analysis is subject to three limitations: (1) reliable data on demonstration service use may not be available because the programs are not required to collect these data, (2) the precision of the estimates will be limited by the small sample sizes, and (3) service receipt may be endogenously related to outcomes.

IV. SYNTHESIS ACROSS SITES

A. OVERVIEW OF THE SYNTHESIS

The ultimate goal of the evaluation will be to provide guidance to HCFA on whether care coordination interventions for chronic illness should be offered as a regular Medicare benefit, and if so, how this benefit might best be structured. Whether the benefit should be offered depends on whether the demonstrations lead to better outcomes for beneficiaries and on the net cost or savings to Medicare. If some programs do exhibit impacts, the structuring of the benefit becomes relevant. This structuring requires assessing (1) what types of organization should be allowed to receive reimbursement for providing care coordination to beneficiaries, (2) what types of beneficiaries should be eligible for the benefit, (3) what types of activities care coordination providers should be required to perform to merit reimbursement, and (4) how providers of the benefit should be reimbursed.

To address these goals, we will conduct two interim and one final synthesis of our findings. In these syntheses, we will pull together our findings from all the sites and outcome measures from both the implementation and impact analyses; we will use this information to draw inferences about the ability of care coordination programs to improve care for Medicare fee-for-service beneficiaries with chronic illnesses and about the most successful ways to implement these programs. The syntheses will entail determining how program effectiveness varies with program characteristics, and how it varies with patient characteristics.

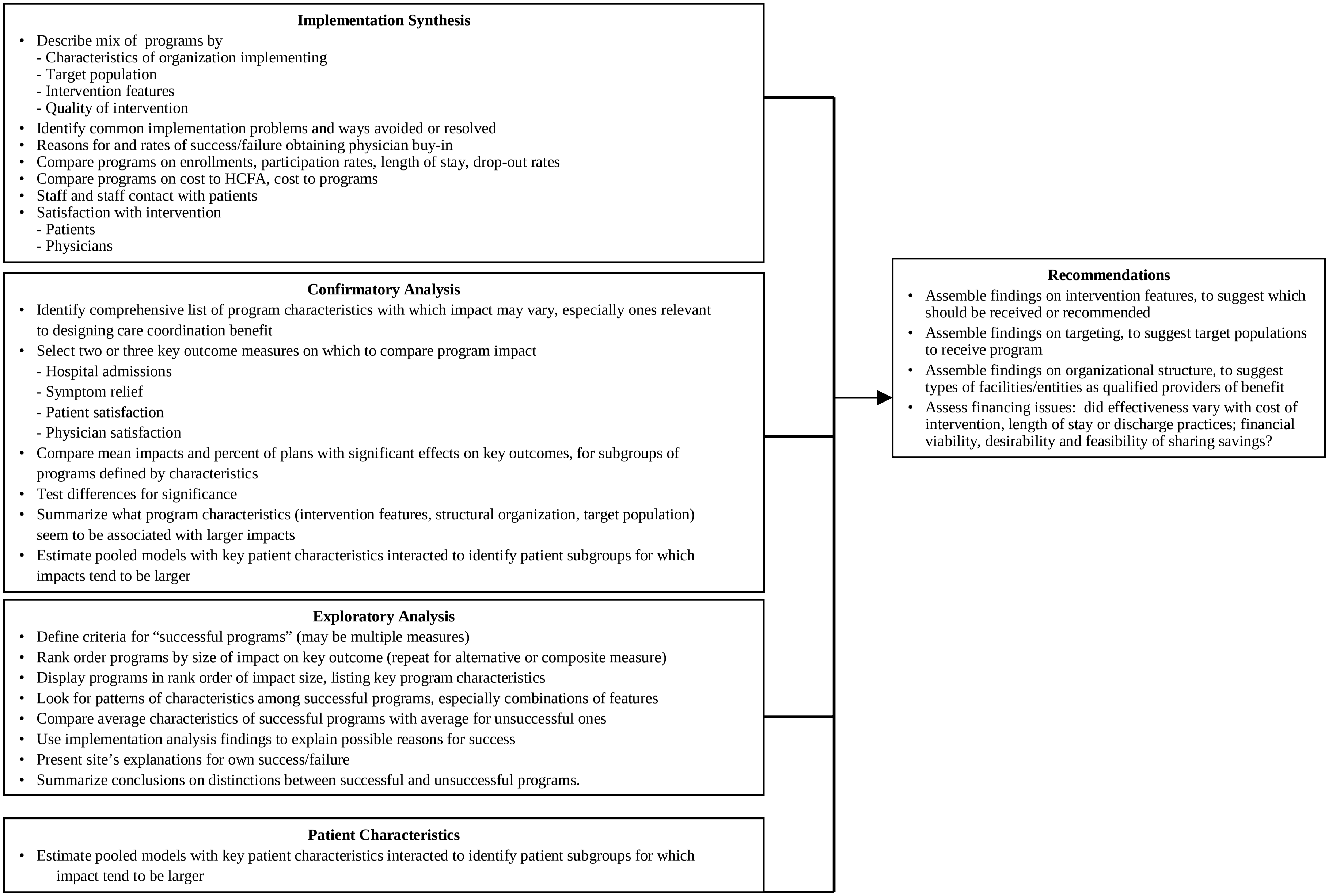
The first synthesis report will be submitted 16 months after the first care coordination site begins enrolling patients. It will focus primarily on synthesizing our findings on program implementation, as only very preliminary estimates of program impacts will be available at that time. The second and final synthesis reports, which will be due 40 months after first the site's startup and at month 57 of the evaluation, respectively, will use findings from both the impact and implementation analyses to draw inferences about what program features appear to work best, and for whom.

To accomplish the study's basic goals, we will draw on the 17 site-specific implementation and impact analyses to describe the range of interventions that were tested, and how impacts varied with the many program characteristics that could potentially influence program efficacy. Our approach to the synthesis will involve four components, all of which will feed into the final recommendations. (We summarize the synthesis in Figure IV.1). In the first component (implementation synthesis), we will summarize and describe in detail the range of interventions the programs implemented. In the second component ("confirmatory" analysis), we will test hypotheses about whether program impacts are greater for programs with certain features or characteristics than for programs that lack these characteristics. The third component ("exploratory" analysis) will entail rank ordering the programs by the size of the estimated impacts on a key outcome measure and visually examining the characteristics of these programs displayed in that ordering for evidence of relationships among combinations of characteristics and impact size. We will also compare the characteristics of "effective" programs with those of "ineffective" programs. Finally, in the fourth component, we will synthesize the findings on the types of patients for whom the programs were most effective.

In Section B of this chapter, we describe the framework for organizing the syntheses. In Sections C, D, and E, we describe how we will conduct the component parts of the synthesis. Section F concludes with a discussion of how we will pull together the findings from these components to make recommendations about whether offering a care coordination benefit appears to be warranted, how it might be structured and targeted, and what further information must be obtained to address remaining policy questions concerning a care coordination benefit for Medicare fee-for-service.

FIGURE IV.1

APPROACH TO SYNTHESIS



Implementation Synthesis

- Describe mix of programs by
  - Characteristics of organization implementing
  - Target population
  - Intervention features
  - Quality of intervention
- Identify common implementation problems and ways avoided or resolved
- Reasons for and rates of success/failure obtaining physician buy-in
- Compare programs on enrollments, participation rates, length of stay, drop-out rates
- Compare programs on cost to HCFA, cost to programs
- Staff and staff contact with patients
- Satisfaction with intervention
  - Patients
  - Physicians

Confirmatory Analysis

- Identify comprehensive list of program characteristics with which impact may vary, especially ones relevant to designing care coordination benefit
- Select two or three key outcome measures on which to compare program impact
  - Hospital admissions
  - Symptom relief
  - Patient satisfaction
  - Physician satisfaction
- Compare mean impacts and percent of plans with significant effects on key outcomes, for subgroups of programs defined by characteristics
- Test differences for significance
- Summarize what program characteristics (intervention features, structural organization, target population) seem to be associated with larger impacts
- Estimate pooled models with key patient characteristics interacted to identify patient subgroups for which impacts tend to be larger

Exploratory Analysis

- Define criteria for “successful programs” (may be multiple measures)
- Rank order programs by size of impact on key outcome (repeat for alternative or composite measure)
- Display programs in rank order of impact size, listing key program characteristics
- Look for patterns of characteristics among successful programs, especially combinations of features
- Compare average characteristics of successful programs with average for unsuccessful ones
- Use implementation analysis findings to explain possible reasons for success
- Present site’s explanations for own success/failure
- Summarize conclusions on distinctions between successful and unsuccessful programs.

Patient Characteristics

- Estimate pooled models with key patient characteristics interacted to identify patient subgroups for which impact tend to be larger

Recommendations

- Assemble findings on intervention features, to suggest which should be received or recommended
- Assemble findings on targeting, to suggest target populations to receive program
- Assemble findings on organizational structure, to suggest types of facilities/entities as qualified providers of benefit
- Assess financing issues: did effectiveness vary with cost of intervention, length of stay or discharge practices; financial viability, desirability and feasibility of sharing savings?

B. FRAMEWORK FOR SYNTHESIZING RESULTS

As a first step in conducting the syntheses, we will report on the number of programs that appear to have met the basic demonstration goal of either reducing the net costs to Medicare without reducing patient well-being or being cost neutral while improving patient well-being. The programs will be cross-classified by their effect on the cost of Medicare-covered services (increased, no effect, reduced but not enough to offset intervention costs, reduced by enough to offset intervention cost, or reduced by more than enough to offset intervention costs) and by their effect on patient well-being (improved, no effect, or reduced). Each assessment will require integrating findings from multiple outcome measures, with possibly conflicting evidence on the size and statistical significance of the effects. For example, a program's estimated impact on cost may not be statistically significant even as the estimate for hospital admissions shows significant reductions. Similarly, estimated impacts on some measures of patient well-being may be statistically significant, whereas others may not be. We will therefore base inferences on the preponderance of the evidence in each site on each dimension.

After this summary of the evidence has been compiled, we will use a unifying framework to synthesize the findings across the individual demonstrations; the goal of the synthesis will be to identify the many dimensions of care coordination and the wide range of program characteristics and features that might be related to program effectiveness. For both the implementation and impact syntheses, we will focus our discussion on the following questions:

- What was the nature of the interventions provided, and how did impacts vary with intervention features?
- What types of organizations provided these services, and how did impacts vary with these characteristics?
- Who received the interventions, and for which subgroups were the interventions most effective?

- How should care coordination providers be reimbursed?

These four basic questions encompass scores of subsidiary questions and measures. Our preliminary attempt to develop a framework (see Figure II.1 in Chapter II) was based on addressing the first three questions but describes only the most rudimentary measures on each of these dimensions. In the following four sections, we describe more fully the types of measures that will be used to address the four broad questions. The four categories of questions will be used to organize the syntheses of both the implementation and impact analyses, in order to facilitate the integration of findings from the two evaluation components. The results showing what program characteristics are most strongly associated with program impacts will guide our recommendations about whether and how to structure such a Medicare coordinated care benefit.

1. What Was the Nature of the Intervention, and How Did Impacts Vary with Intervention Features?

Perhaps the most critical factor to examine is the way that impacts vary with characteristics of the intervention. In our current thoughts about a program classification scheme (Chapter II), we describe a simple (and crude) way of categorizing programs according to where they place their emphasis:

- Improving patient self-care and adherence behavior
- Improving physician prescribing and treatment practices
- Improving communication and coordination among providers
- Improving the arrangement and provision of services

Programs may focus exclusively on one or two of these components or may have interventions that attempt to accomplish all four objectives. There are 15 possible combinations that can be defined with these four categories. The goal of our analysis here will be to assess whether programs that focus exclusively on (say) patient education can be just as successful as those that focus on all four components. We may find that a program does not seem to have to devote much attention to the arrangement of services (say) in order to achieve substantial improvements in patient outcomes.

As Chapter II clearly shows, even the 15 possible categories of program emphasis do not begin to cover the range of potentially important intervention features. Most important, the *quality* of the intervention in a particular area will be an important determinant of program success. Thus, we will want to describe and relate program effects not only to the combination of interventions a program attempts to provide, but also how well the interventions in each area are implemented. We will assess programs on various aspects of care coordination identified in Chen et al. (2000)—assessment and care planning, service delivery and monitoring, and reassessment and revision of care plans. We will also describe how the interventions were implemented and will relate that to outcomes. This dimension includes the professional backgrounds of the care coordinators, the amount of their contact with patients, caseloads, and many other program characteristics. Table II.2 in Chapter II presents a preliminary list.

2. What Type of Organizations Participated, and How Did Impacts Vary with These Characteristics?

Another key component of the synthesis will be our assessment of the types of entities that have the capability to provide effective care coordination services. For example, we have learned from the previous Case Management Demonstrations that a care coordination program lacking strong involvement of and support from primary care physicians is likely to fail. Thus, we will assess how program impacts vary with the extent to which the program is integrated with local hospitals and physicians—fully integrated, partially integrated, or fully independent. We will also investigate the extent to which impacts vary with the type of organization implementing the intervention, such as academic medical centers, health care systems, or commercial vendors, which may also affect how well the intervention is integrated with the medical care system.

3. Who Received the Interventions, and for Which Subgroups Were the Interventions Most Effective?

Another key issue for the synthesis will be determining whether the tested care coordination interventions appear to work better for some target populations than for others. For example, programs that serve only patients with a particular disease (disease management programs) may be more or less successful at improving care and/or reducing costs than programs that serve patients with a wide range of diseases, and that therefore take a more generic approach to care coordination. Similarly, programs serving patients with particular diseases (for example, CHF) may have greater impacts on costs or quality of care than programs serving patients with other diseases (for example, diabetes or COPD). It will also be important to ascertain whether other criteria that programs use in defining their target populations affect the programs' degree of success. For example, programs that restrict their enrollees to beneficiaries who have no severe cognitive impairments may be more likely to improve outcomes than programs without this restriction. Conversely, programs that do not exclude certain types of patients, such as those with many comorbid conditions, may be able to achieve larger effects on program costs by coordinating care across more providers.

For this part of the synthesis, we will also explore whether some effective programs appear to "cream" the potentially eligible population in order to enroll cases that are most likely to benefit from the intervention. Some degree of screening may be appropriate and desirable, but it will be important to ascertain the generalizability of the intervention, and the size of the true target population for which the intervention has been tested.

4. How Should Care Coordination Be Reimbursed?

In addition to assessing how impacts vary with the programs' structural organization, process of care features, and patient characteristics, we will discuss what we have learned about payment methods for the benefit. We will have somewhat limited information for this analysis, because all the demonstration programs will be paid in the same way (a capitated rate per person month enrolled in the program). However, a number of issues can be addressed. Specifically, we will examine the distribution of programs and how program impacts vary with the following financial characteristics:

- Monthly per beneficiary cost to HCFA
- Beneficiaries' average length of stay in the program
- Start-up costs
- Sharing of net cost savings to Medicare
- Financial viability (cost to the program relative to charge to HCFA)
- Payments to physicians for care coordination
- Any financial incentives to case managers

We must examine monthly cost to HCFA because the more expensive interventions may generate larger net savings to HCFA through larger impacts on the use of expensive services. Length of stay is important because programs may discharge patients when they deem care coordination unnecessary or after a fixed length of time; some programs may never discharge patients. Furthermore, some programs may provide different levels of care coordination intensity, depending on a patient's needs or length of time since enrollment, and the rate charged to HCFA may vary with this level. Start-up costs—those paid by HCFA and those borne by the program—are likely to vary considerably across programs and should be documented and related to program impacts. A comparison of operating costs the programs incur with the payments they receive from HCFA will enable us to assess whether providing care coordination at a cost acceptable to HCFA is likely to be financially viable for potential providers of the benefit. We will examine whether payments are made to physicians, and the size and nature of the payments, to determine whether these features lead to greater physician satisfaction, greater physician buy-in to the program and more cooperation with it, and differential impacts. Finally, we will determine whether the programs provide any other financial incentives to encourage better outcomes, and how impacts for programs that use incentives differ from those that do not use them.

C. THE IMPLEMENTATION SYNTHESIS

The first goal will be to synthesize the findings from the implementation analyses by tabulating the distribution of demonstration sites on both program features and program implementation successes or failures. Features include programs' enrollment levels, disenrollment rates, conditions targeted, types of interventions, program focus, quality of the interventions, and the many other characteristics described in Chapter II. These tabulations will also include combinations of the various characteristics, such as separate distributions of disease management and case management programs, by type of intervention, and the combination of interventions (see Figure II.1). Program successes or failures include whether the programs achieved their target enrollment levels, whether they reported difficulty obtaining physician cooperation, and the ratings we assign to them on the quality of their intervention components (for example, whether the patient education intervention is rated as strong).

The implementation synthesis will also summarize our findings from the site-specific analyses on the programs' financial characteristics, such as those identified in Section B.4, and the nature and extent of implementation difficulties encountered across the sites. The synthesis will also attempt to determine why some programs appeared to have more difficulty than others in implementing their interventions as planned, and what lessons organizations that contemplate establishing a Medicare coordinated care program in the future might draw from their experiences.

D. HOW DO IMPACTS VARY WITH INTERVENTION FEATURES AND PROGRAM ORGANIZATIONS?

Obviously, it will not be possible with 17 sites to sort out the combination of the many characteristics that explains why some programs have substantial impacts on the costs and quality of patient care and others have no (or smaller) effects. Our goal will be much more modest: to identify how program effects vary with these program characteristics, and to test for whether the differences we observe are statistically significant.

If at least some of the programs have significant impacts on key outcomes, we will conduct both confirmatory and exploratory assessment of the sources of these differences. The confirmatory analysis will be accomplished by defining a broad array of characteristics, based on the discussion in the previous section, and by comparing mean impacts on key outcomes for programs that have a given characteristic with the mean impacts for programs that lack that characteristic. Table IV.1 provides an example of the characteristics to be used in grouping programs for these comparisons. The exploratory assessment will be accomplished by distinguishing programs that successfully improve a given outcome from programs that do not, and by comparing the characteristics of the successful and unsuccessful programs. The exploratory analysis will therefore determine the extent to which program success appears to be specific to programs with a particular characteristic. The exploratory analysis will also be used to determine whether program success seems to be linked to combinations of measured characteristics or to any less tangible characteristics identified in the implementation analysis.

TABLE IV.1

PROGRAM CHARACTERISTICS FOR WHICH IMPACTS WILL BE COMPARED

	Number of Programs	Average Impact	Percent of Programs with Impacts
Structural			
Fully integrated with providers			
Partially integrated			
Not integrated			
Type of Organization			
Academic medical center or hospital			
Vendor			
Targeting			
Specific diseases only			
CHF			
Diabetes			
COPD			
Other			
Chronically ill			
Frail elderly			
Intervention Focus			
Improving patient self-care			
Yes			
No			
Improving physician practices			
Yes			
No			
Improving service arrangement			
Yes			
No			
Improving communication among providers			
Yes			
No			
Common combinations of intervention			
All four dimensions			
Improving self-care only			
Etc.			

TABLE IV.1 *(continued)*

	Number of Programs	Average Impact	Percent of Programs with Impacts
Quality of Intervention			
Patient education			
High			
Medium			
Low			
Patient Assessment			
High			
Medium			
Low			
Physician Education			
High			
Medium			
Low			
Communication Among Providers			
High			
Medium			
Low			
Patient Monitoring			
High			
Medium			
Low			
Feedback to Case Managers and Providers			
High			
Medium			
Low			
Physician Buy-in and Involvement			
High			
Medium			
Low			
Financial Incentives			
Yes			
No			
Physician Cost/Month			
>$300/month			
$200-300/month			
$100-200/month			

We will make the comparisons of impacts for key outcomes, including hospital admissions, total cost, satisfaction with care, symptom improvement, health status, functioning, and patients' ability to perform their normal activities. In addition, we will construct a composite categorical measure of program effects—programs that have significant sizable effects on expensive Medicare services and patient outcomes, ones that improve patient well-being but not costs, and ones that affect neither patient outcomes nor costs. It is likely that evidence for this measure will be mixed, with some measures showing improvement and others showing none. Therefore, we will have to consider all measures to establish a sense of whether patient well-being is affected. This assessment may be based on the number of patient outcome measures for which the program has significant effects or on the significance of a key selected measure that appears to be representative of other measures. We will also construct a summary cost-effectiveness measure for each program, defined as the ratio of estimated savings in Medicare cost per month divided by the estimated intervention cost per beneficiary per month, both calculated over the 12-month period after enrollment. These cost-effectiveness ratios will provide an indication of which interventions generate the greatest savings per dollar invested by HCFA.

1. Confirmatory Analysis: Testing for Differences in Impacts Between Groups of Programs

In drawing these comparisons we will present tables showing how program features are associated with the mean impact and with the proportion of plans with a statistically significant impact in the desired direction (or the proportion with a point estimate larger than a certain level). Use of mean impacts allows us to assess whether impacts tend to be larger for programs with a given feature than for those without the feature and ensures that we identify situations in which impacts are consistently larger for certain types of programs but not statistically significant. However, comparison of mean outcomes can mask important relationships if some impact estimates are negative or extremely large. Comparing the proportion of programs with significant effects prevents this problem but fails to capture any differences in the magnitude of the impacts.

These comparisons are further complicated by the different sample sizes expected across sites. Although we expect all the random assignment sites to have the same-sized survey sample, for outcomes obtained from claims data, we expect to have substantially larger sample sizes for five of the sites. We will rely on the survey sample estimates in the syntheses to eliminate this source of disparity but will note any cases in which the claims sample estimates indicate statistically significant differences that do not appear in the survey sample estimates.

We will test the differences in mean outcomes for groups of programs defined by characteristics (and differences in the proportions of programs with significant effects) to determine whether they are statistically significant. Because the programs and samples are independent of each other, the variance of the difference in mean impacts between (say) 5 programs that have a particular characteristic and 12 that do not have it is simply

$$(1) \quad var = \frac{\sum_{i=1}^{5} s_i^2}{25} + \frac{\sum_{i=6}^{17} s_i^2}{14},$$

where s_i^2 is the variance of the impact estimate for the *i*th site.

If the variances for the site-specific impact estimates are equal across sites (we expect them to be similar, given the similar sample sizes), this reduces to $s^2(1/5 + 1/12) = s^2\,(17/60)$, or about 28 percent of the size of the variance of an individual site-specific estimate. Because minimum detectable differences are proportional to the standard error of the estimate, this implies that, in comparing impacts across sites, we should be able to detect differences that are about half ($\sqrt{.28}$) of the impact size detectable at the site level. For example, we can detect effects of 10 percentage points on the probability of hospital admission in the site-specific estimates. Thus, we should be able to detect (with the same 80 percent power) differences of about 5 percentage points between the average impact for a group of five sites and the average impact for the other 12 sites. Differences when the sites are more evenly split between the two categories will have smaller variances and smaller detectable differences (that is, more precision).

2. Exploratory Analysis

The exploratory analysis will be useful for identifying combinations of characteristics that seem to be associated with program success (assuming that at least some of the programs have favorable impacts). As noted, there are far too many potentially important characteristics to determine the relative importance of each one, so it will be impossible to conduct a rigorous analysis to determine what *combination* of characteristics is most influential in producing desirable program impacts. Nonetheless, much can often be learned by comparing the characteristics of the successful and unsuccessful programs.

We will make this comparison in two ways—first, by arranging the data on program characteristics in a manner that facilitates visual identification of patterns, and then by comparing the mean characteristics of successful and unsuccessful programs. We will first order the programs by the size of their impact on hospital admissions, because this outcome probably is the most important one for which programs should have impacts if they are to be at least cost neutral. A reduced need for hospitalizations among this high-risk population is also an indicator of improved quality of care. We will then create a large table, in landscape format, to display all the characteristics we believe are most likely to influence program impacts. Each row of the table will represent a different program site, and each column will represent a characteristic of the programs. Because the programs will be listed in descending order by size of impact on hospital admissions, programs characteristics that cluster in the top portion of the table will tend to be associated with successful programs. Examining other characteristics that these programs share may enable us to distinguish patterns suggesting combinations of characteristics that are important for success. For example, we might observe that programs focusing on CHF as a diagnosis tend to be heavily represented among the most successful programs, but that CHF programs lacking a strong perceived buy-in by primary care physicians were not among the successful ones, or had noticeably smaller impacts. We will repeat this process with programs arranged by size of impacts on one or more key patient outcome measures, such as self-reported health status.

The second exploratory approach will be to define some subset of the programs as "successful," based on their impacts on some combination of key service use/cost and patient outcomes, and to compare mean characteristics of the two groups. In addition to those listed in the landscape table, the characteristics we examine will include others identified in the implementation analyses. The characteristics listed in Table IV.1 provide an illustrative list of some of the characteristics that we expect to use in these comparisons. We will use several alternative definitions of "successful" programs to ensure that our inferences are robust to the definition used, as it is somewhat arbitrary. For example, we may define programs as successful only if they show statistically significant impacts on hospital admissions. Alternatively, the definition could include any program if its average monthly Medicare cost for the treatment group was more than one standard deviation below that of the control group.

E. RELATING PATIENT CHARACTERISTICS TO IMPACTS

In addition to determining what types of interventions and organizations seem to yield the best outcomes, the syntheses will also assess whether care coordination appears to work better for some types of beneficiaries than for others. In most sites, sample sizes will be too small to yield highly reliable estimates of impacts for subgroups of patients. However, we may be able to identify patterns in the results if we compare these findings across sites. We will also pool data from multiple sites to assess whether impacts appear to be greater for some subgroups of beneficiaries than for others.

The hypotheses that we will test include whether care coordination programs appear to have greater impacts for patients who (1) have certain diagnoses, (2) are younger, (3) have relatively higher education levels, and (4) are at a stage of illness that is neither too severe nor too mild to be affected by the interventions. Clearly, care coordination programs may be able to make greater improvements in outcomes for some conditions than for others, especially given the relatively short follow-up period (one-year). Beneficiaries who are younger may have fewer cognitive difficulties than do older beneficiaries, and those who are more highly educated may be better able to adhere to recommended self-care, and to appreciate the importance of doing so. Stage of illness may be important if programs enroll some individuals who are relatively healthy and at fairly low risk of adverse outcomes, leaving little for the intervention to improve. Conversely, programs may not be able to affect outcomes for beneficiaries whose condition is too severe for any intervention to reverse or halt the decline. We may also examine other subgroups, such as the number of comorbidities the beneficiary has, whether the beneficiary lives alone, the number of hospitalizations or total Medicare cost in the prior year, and characteristics that emerge as important predictors of outcomes. We will also elicit program staff's opinions on the types of beneficiaries whom they believe will likely derive the greatest benefit from the intervention. We will use these measures to define subgroups of interest.

We will use two methods to determine whether any of these beneficiary subgroups did indeed experience greater impacts than others: (1) compute the average subgroup effects estimated for the individual program sites, and (2) pool the data from multiple sites to estimate a single model with interactions. Subgroup estimates obtained on individual sites will be averaged to determine whether beneficiaries with a particular characteristic experience greater impacts than beneficiaries without the characteristics. We will also compile the distribution of sites by the relative size of the estimated impact on the subgroup of interest and the other enrollees. For example, we may calculate the proportion of sites with impacts for the subgroup of interest that were (1) at least one standard deviation larger than those for the other enrollees, (2) within +/- one standard deviation, or (3) at least one standard deviation smaller. We will not restrict these comparisons to statistically significant differences, because the available sample sizes will typically be too small to yield this level of precision, even if the impact differences are quite large.

To estimate differences across beneficiary subgroups by pooling the data, we will combine all the data from certain sites. For example, we expect to be able to pool the data from the eight demonstration sites that will limit their intervention to CHF, to test whether impacts differ with the stage of illness at admission (using the NY Heart Association scale) and with other patient characteristics. We will also investigate the possibility of pooling data from all sites, but some of the beneficiary subgroups of interest (such as stage of illness) will differ across sites in ways that would make pooling impractical. This approach has the advantage of increasing the sample sizes substantially for the subgroups being compared but runs the risk of creating biases by pooling beneficiaries whose outcomes are determined by very different processes. Thus, careful modeling with the pooled data set will be necessary to ensure that the distinctions among programs are taken into account.

F. REPORTING AND RECOMMENDATIONS

We will conduct two interim syntheses, which will form the bases for the two scheduled Reports to Congress on Coordinated Care, and one final synthesis. The first synthesis report will be completed (in draft) in month 16 after the first site begins operations (that is, November 2002, assuming that the first demonstration site begins enrolling in July 2001). This report will synthesize primarily the findings from the site-specific case studies and implementation analyses, as relatively little data on patient outcomes will be available. However, we will incorporate estimated impacts from the first interim site-specific analyses that have been completed by that date (that is, sites beginning enrollment within three months of the first site). This interim analysis will be based on short-term outcomes (hospital admissions during the first two months after enrollment) for an early cohort of enrollees (those enrolling during the first four months of operation). The first Report to Congress will be due two months after the draft first synthesis report.

The (draft) second synthesis report will be delivered 40 months after the first demonstration program begins enrolling (November 2004) and will incorporate findings from the second round of site-specific analyses. These analyses will be based on virtually the entire analysis samples for the set of sites for which the second site-specific report has been completed, provided that sites are able to enroll the targeted number of sample members within one year from startup. The second synthesis report also will update implementation findings from the first synthesis report; we will obtain the updated information in telephone discussions with program staff conducted one year after program startup. All the issues described in this chapter will be addressed in the report. The second Report to Congress is due two months after the second interim synthesis report and will be pulled directly from the revised version of it.

The final synthesis report will be based on impact estimates for the full samples from all the programs and will be completed (in draft form) in month 57 after the start of the evaluation (June 2005). It will update the second interim synthesis report. The results may differ substantially fromr the ones presented in that report; the second interim report will not include results from late-starting programs, and sample sizes for included sites may differ. Chapter V discusses our intention to revisit the timing and contents of the interim site-specific and synthesis reports after site start-up dates have been determined, so that the second Report to Congress reflects the experiences of more demonstration sites.

To make recommendations about the value and structure of a Medicare coordinated care benefit, we will use the findings from the four components of the synthesis to address the basic questions about what interventions appear to be worth offering to beneficiaries nationally, what types of organizations should be considered eligible for Medicare reimbursement for providing the service (and what qualifications they should exhibit to quality), who the benefit should be offered to, and what reimbursement and financial issues HCFA should consider. This assessment will also indicate the issues about which we are unable to make recommendations, due to insufficient evidence (because too few programs exhibited a particular characteristic to enable us to assess its relationship to impacts, or because the evidence about whether the characteristic was associated with cost-neutrality was mixed). It is likely that additional demonstrations may have to be conducted before some questions about the optimal structure of a benefit can be answered. However, we hope to identify some types of programs for which the evidence strongly suggests that a care coordination benefit meeting those criteria would be likely to benefit Medicare beneficiaries and save money for the Medicare program.

V. REPORTING OF DEMONSTRATION FINDINGS

The demonstration evaluation will produce several types of reports, including site-specific analysis plans and case studies of individual demonstration sites, as well as interim and final site-specific reports. We will also produce reports that synthesize findings across all the sites. The synthesis reports will be adapted to develop two reports 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	Calendar Month
Design report	5	2/01
Site methodologic evaluation	6	3/01
Draft site-specific analysis plans	8	5/01
Site case studies	6 months after site enrollment begins	1/02-7/02
First interim site-specific evaluation	12 months after site enrollment begins	7/02–1/03
Second interim site-specific evaluation	33 months after site enrollment begins	4/04-10/04
First interim synthesis	26*	11/02
First report to Congress	28*	1/03
Second interim synthesis	50*	11/04
Second report to Congress	52*	1/05
Final synthesis	57*	6/05

*Assumes first BBA program starts enrolling in July 2001 (month 10).

A. SITE-SPECIFIC DESIGN ASSESSMENTS AND ANALYSIS PLANS

The site-specific analysis plans will assess the applicability of the basic research design to each site’s particular circumstances. Draft analysis plans will be delivered to HCFA within three months after the demonstration award date (January 19, 2001). This schedule is slightly longer than the one presented in our proposal because more sites have been funded than were anticipated. Final analysis plans will be completed after the first case study calls (approximately two months after sites begin patient enrollment).

Sites will differ with respect to patient referral sources; expected total enrollment; availability, content, and format of data on non-Medicare services provided; expected impacts on patient behavior; and many other areas. The site-specific analysis plans will address these issues and will present a plan for dealing with each site's unique circumstances within the design framework of the evaluation. To develop the plan, we will prepare a site-specific assessment of each site's research design within two months after award. We will then contact the programs by telephone to discuss any potential problems resulting from the proposed sample sizes, experimental design, intake and randomization procedures, eligibility rules, possible contamination of the control group, or any other program features or assumptions that could adversely affect the evaluation. After obtaining agreement on feasible ways of adapting a site's approach to overcome these problems, we will share these potential adaptations with the site, elicit its comments, and agree on a final research design. We will then use the input from each site and from our own assessment to prepare the site-specific analysis plan. We will deliver it to HCFA by mid-March of this year.

In particular, we will have to adapt the basic design by determining whether sampling will be necessary to select the cases to be interviewed and, if so, how it should be conducted; what methodology we will use to select the comparison site or to conduct randomization; and what data are available. We do not expect to conduct sampling in most sites; rather, we will likely have to interview all their enrollees to meet the minimum sample size (only 5 of the 15 BBA sites expect to enroll more than the minimum number of cases; see Section III.A). In the case of sites that expect to enroll far more than the required number of cases during the intake period, which may last as long as one year, we will determine a sampling rate and establish a process for monitoring enrollments relative to expectations. In this way, we will ensure that proposed survey sample size targets are met. For any sites without random assignment, the site-specific modification will be more complex, because we will have to identify the criteria for selecting the full comparison sample, as well as the subset of cases to be interviewed.

Other site-specific adaptations to the analysis plan will be necessary if the intake information varies across sites, cases are identified in different ways, or different site-level data are available for analysis. For example, some sites may have useful intake information on the severity of illness for both treatment and control cases. Some sites may have research-quality data on additional services they provide beyond those covered by Medicare, which we will be able to use to determine the proportion and types of cases receiving these services. We can then test for whether impacts are greater for the individuals most likely to receive the additional services.

B. CASE STUDIES

The case study reports will describe the program goals of the site, the nature of the intervention the site plans to deliver, and its start-up and early implementation experiences. We will submit each case study report to HCFA within six months of patient enrollment at each site.

The information in the case study reports will be derived from protocol-guided telephone interviews with key site staff and through the review of site documents (proposals, protocols, and data collection forms). The telephone calls with the sites will take place approximately two months after each site begins patient enrollment.

The case study reports are site-specific evaluations that will monitor each site’s early progress in implementing its proposed intervention and evaluation. We will describe the structure of each program, including (1) how the interventions are targeted; (2) how the program is organized, including the degree to which the program is integrated with providers (fully integrated, independent, or mixed); and (3) the goals of the program interventions (improving patient education/compliance, improving provider practice, improving service arrangement, and improving patient and provider communication). We will identify potential problems, including enrollment shortfalls, changes in the proposed intervention or target population, contamination of random assignment, staffing difficulties, physician opposition, and poor data quality. We will also describe how the site assesses participants' needs, arranges or delivers care, and reassesses or adjusts care or coordination. These studies will provide early feedback to HCFA and valuable input for our site-specific reports.

C. FIRST INTERIM SITE-SPECIFIC REPORTS

The first interim site-specific reports will describe program operations over the first nine months and will use short-term outcomes on an early cohort of sample members to provide very preliminary estimates of program effects in each site. These reports will be sent to HCFA 12 months after the start of enrollment in each site. Given that we expect the earliest sites to begin enrolling in July 2001, the first interim reports will be submitted in July 2002, with the remainder due over the subsequent six months or so.

Data for the first interim site-specific reports will combine the findings of the implementation analysis, based on in-person site visits and review of site documents, with descriptive data about enrollment levels and site project costs.[1] Estimates of demonstration impacts will come both from patient telephone surveys and Medicare eligibility and claims data.

The interim site-specific report will contain a description of the progress of participant recruitment and will provide detailed descriptions of the intervention. Analyses of recruitment

[1] We have assumed that the implementation contractor for the demonstration will collect enrollment and cost data and will make them available to us for analysis in this report.

will identify the number of beneficiaries enrolled over the first nine months of operations and their characteristics, using enrollment files and intake data. In addition, we will calculate participation rates and disenrollment rates and will present reasons for enrollment and disenrollment. We will use claims data for the 12 months prior to enrollment to describe the preenrollment service use and cost patterns of beneficiaries who enroll during the first 4 months of operations. Data on prior use and costs and intake data will be used to compare (1) the treatment group with the control group, (2) dropouts with stayers, and (3) participants with nonparticipants.

The evaluation will also report findings from our visit to the site, which will have been completed just before the first interim report's due date. This discussion will describe the intervention and its components in considerable detail. The intervention will be compared with what had originally been planned and will have been reported in the site's Case Study Report (completed six months after the site begins enrollment). We will provide explanations for any deviation from the original design or early months of operation. Site visit staff ratings of the site on each key component of care coordination will be provided as well. The report will also present statistics on services provided to the early cohort during the first few months of enrollment, using data provided by the site. The types of data available for analysis from the site and the quality of these data will be described.

In order to produce estimates of program impacts by month 12 after the start of enrollment, we will have to limit our data to cases enrolled during the first four months of operations and, for these cases, we will be able to estimate impacts on claims-based outcomes over only the first two months after enrollment. Under this schedule, we will have four months to obtain reasonably complete claims data, and two months for analysis and report preparation. Outcome measures obtained from the patient survey also will be available for the four-month cohort. Given the limited amount of time available, we will limit both claims and survey analyses to a few key outcome measures. If fewer than 200 beneficiaries (100 treatments and 100 controls) enroll during a site's first four months of intake, we may decide to forego conducting these preliminary impact analyses in that site, because the results obtained with data from such small samples could be highly misleading.

D. SECOND INTERIM SITE-SPECIFIC REPORTS

The second interim site-specific reports will be much more comprehensive than the first ones. In the second interim reports, we will perform all the analyses on this partial sample needed for the final estimates, to facilitate rapid preparation of the final synthesis report. These reports will be sent to HCFA approximately 33 months after the start of enrollment in each site.

Data for the second interim reports will come from (1) a second protocol-guided telephone interview with key site staff, (2) a review of site documents containing descriptive data about enrollment levels and site project costs, and (3) findings of site-specific impact analyses.

Each report will include an implementation analysis that updates the information provided in the first report; the update will be based on the second round of telephone calls with site staff.[2] The same topics will be covered, but with an emphasis on how the program has evolved since the first interim report was completed 21 months previously. Enrollment statistics and comparisons will be updated, using all beneficiaries enrolled during the first 12 months of intake. We will construct models to determine the effect of various beneficiary characteristics, such as prior service use and cost, demographic characteristics, and place of residence, on probabilities of enrollment (using all eligibles) and of dropping out (for treatment group members).

[2] The second round of calls will take place approximately 24 months after the site begins patient enrollment and 18 months after the site visit.

If programs enroll beneficiaries for more than 12 months (as most intend to do), the estimates of program impacts will be based on two different cohorts: (1) those enrolled during the first 12 months, and (2) those enrolled during the first 18 months. For those enrolled during the first 12 months, we will have data on claims-based outcome measures for the full 12-month follow-up period (allowing time for data lags, followup, and analysis).[3] For those enrolled during the first 18 months, we will have data on claims-based outcomes for the first 6 months of enrollment. The report will compare six-month impact estimates for those enrolled in months 1 through 12 and for those enrolled during months 13 through 18, to assess whether impacts changed as the program matured. Impacts will be estimated on outcomes measured over the first 3, 6, 9, and 12 months for the early cohort, to assess the persistence of program effects. Survey outcomes measures will be available only for the survey sample, which we expect to enroll during the first 12 months. However, if more time is required to enroll the sample, we will be able to include survey data on everyone who enrolled during the first 18 months. In addition to patient survey data, the full physician survey sample will also be available for the second interim analysis.

We will compare impacts across outcome measures to draw inferences about the source of any impacts on patient service use or health outcomes. We will also ascertain whether programs improve both costs and quality of care, only costs or only quality of care, or neither costs nor quality of care. Because the cost estimates for the 12-month followup should be reliable by the time we produce this report, we will present preliminary estimates of the cost effectiveness of the

[3] We would have 12 months of intake, plus 12 months of followup, 4 months of lag for the data to become complete, 1 month to update the files, 2 months to conduct the analyses, and 3 months to draft the report.

intervention. We will compare estimated savings in Medicare costs with program costs to determine net savings per month at risk for the year after enrollment.

E. SYNTHESIS REPORTS

One of the most critical components of the evaluation will be the synthesis of the findings from site-specific analyses to determine whether some types of interventions appear to have greater impacts on patient outcomes and savings than others. This component will be one of the most difficult parts of the analysis to conduct because there are only 17 sites, each of which is likely to differ from the other sites on numerous potentially important dimensions.

We will conduct a first interim synthesis (draft due 16 months after the first program begins enrollment), a second interim synthesis (draft due 40 months after the first program begins enrollment), and a final synthesis (draft due 57 months after the award of our contract).

The synthesis reports are cross-site reports. The first and second interim synthesis reports will be based on the first and second interim site-specific reports and will not require additional data collection. We expect that additional site data and Medicare data will be available for the final synthesis report, which will require additional analysis.

Note that the synthesis reports will incorporate only the findings from the site-specific reports that have been completed in time for inclusion in the synthesis. Because the syntheses will form the basis for the two Reports to Congress, their timing is determined by the mandated schedule for the Reports to Congress. If the first coordinated care site begins operations in July 2001, the first synthesis report (due in November 2002) can include findings from the first interim site-specific reports only for sites that begin by October 2001. If this aspect of the reports presents problems for HCFA, an alternative schedule must be developed.

1. First Interim Synthesis

The first interim synthesis report, due 16 months after the first care coordination site begins operation, will focus on comparing the implementation experiences of the sites, as the impact estimates available from the first interim site report are likely to be based on very few observations. Although we will compare impact estimates across sites, we will not attempt to draw inferences from them at this early stage of the evaluation. We will also compare sites' success in enrolling and retaining beneficiaries in their programs, as well as the characteristics of these enrollees. In addition, we will compare sites on findings from the physician surveys. Results for the 2 Lovelace programs and the results for the 15 care coordination demonstrations will be described in two separate sections. A third section will compare and contrast the programs, using the program classifications described in Chapter II of this report and any other important characteristics that emerge from the implementation analysis. The report will also present estimates of impacts across patient subgroups.

2. Second Interim Synthesis

This report, due in draft form 40 months after the start of the first demonstration program, will draw on the second interim site-specific reports to compare impacts on enrollees' service use, costs, quality of care, and all other outcomes examined in those reports. We will attempt to identify factors associated with successful or unsuccessful interventions, with success measured in terms of both cost savings and patient outcomes. Programs will be grouped in eight possible categories, defined by the interaction of their effects on cost (net savings, cost neutral, reduction in Medicare cost but less than intervention costs, or no reduction in Medicare cost) with their effects on quality of care (significant improvement or no significant improvement). We will seek to identify common features of programs within a given cell or outcome category, relying on the characteristics identified at the outset of the project, as well as on characteristics that emerge from the implementation analysis.

3. Final Synthesis

The final synthesis report, due in draft form in month 57, will provide final site-specific impact estimates for all the programs and will compare findings across sites, using the approaches described above. We will be particularly interested in comparing the cost effectiveness of the different programs. The report will compare findings across programs but will also include a summary description of each demonstration program and its effects.

The final report will discuss the feasibility and desirability of making effective models a permanent benefit under the Medicare fee-for-service program. In that report, our challenge will be to define the intervention in such a way as to ensure that the success experienced by the demonstration sites can be replicated in an ongoing program. We will therefore have to carefully specify (1) what types of organizations should be allowed to receive reimbursement for care coordination, (2) the patient screening criteria allowed, (3) the intervention, (4) monitoring procedures, and (5) financial arrangements. Programs wishing to be reimbursed for providing care coordination services should be required to meet criteria ensuring their ability to implement successful care coordination. Patient screening criteria will be necessary to ensure that an ongoing program continues to target the group for whom the intervention has been shown to be beneficial, and to prevent cream-skimming. New programs should attempt to replicate the successful interventions as closely as possible, but a balance must be struck between rigid adherence to a successful intervention and stifling creative new approaches that could improve outcomes or increase savings. Monitoring is necessary to ensure programs do not implement ineffective interventions that will increase costs. As an example of one possible way to conduct monitoring, we might suggest tracking hospitalization rates and costs for beneficiaries under an ongoing program and comparing them with the results achieved in the demonstration.

F. REPORTS TO CONGRESS

We will produce two reports to Congress based on our evaluation. The reports will be due approximately 18 months and 42 months, respectively, after the start of patient enrollment at the first demonstration site. These reports will analyze implementation experiences and findings to date of the 15 BBA-funded demonstration sites, as described in the first and second interim synthesis reports. We will write the reports for an audience of high-level policy makers and decision makers who may not be familiar with the demonstration project or evaluation methodologies.

VI. PROJECT MANAGEMENT AND TIMELINE

This evaluation project requires MPR and S.A. Squared, our subcontractor, to organize and coordinate many simultaneous tasks. MPR's core team to manage this project consists of Randall Brown, Jennifer Schore, and Arnold Chen. Dr. Brown will direct the project and will serve as the point of contact for communications between HCFA and the project team. He will lead the impact analysis of program effects on service use and costs and will be responsible for all major design decisions and coordination with other project staff. He also will be responsible for monitoring the project timeline and budget. He will direct the impact analyses and will oversee the writing of all the project reports. Ms. Schore and Dr. Chen will serve as co-principal investigators. Ms. Schore will be the task leader for the case study analysis and for the collection and analysis of all qualitative data. Dr. Chen will be the task co-leader for the development of the patient and provider surveys and will lead the impact analysis of program effects on quality of care and satisfaction.

A. CHANGES TO THE EVALUATION

In the time since MPR submitted its proposal for the evaluation, the demonstration sites have been selected, leading to several changes in the evaluation's schedule and budget. First, the number of BBA sites was increased from 9 to 15. This change means that MPR will require additional funds to include six more sites in the evaluation. Second, the award date for the demonstration sites was approximately three months later than we had anticipated in our proposal, so that we had to recalculate the project timeline. Third, the start dates of the sites will be staggered. For planning purposes, we have assumed that three sites will begin enrollment in July 2001 (Georgetown and the two Lovelace sites), half the BBA sites will begin in October 2001, and the remaining BBA sites will begin in January 2002. These staggered starts will cause an even greater overlap in the number of tasks occurring simultaneously. In addition, if the sites (particularly late-starting sites that begin enrollment in January 2002) are not able to recruit all the patients they need within 12 months, we will not have 18 months of complete follow-up data for all their patients. If HCFA wants us to include these data in our analysis, we may have to have an extension of the end date of our contract. Finally, our proposal (based on 11 demonstration sites) assumed that 7 sites would require random assignment and 4 would require the development of a comparison group. It now appears that 16 sites will require random assignment and 1 will require the development of a comparison group.

B. TIMELINE

To develop the project timeline (Figure VI.1), we have used the following dates and assumptions:

- The start date for the evaluation contract was September 29, 2000. Most dates in the project timeline (Figure VI.1) and in the schedule of deliverables (Table VI.1) are in terms of evaluation months (for example, October 2000 is month 1). Site-specific deliverables are linked to the date when enrollment begins in that site.

- The site contracts were awarded in January 2001. We assume patient enrollment will begin in July 2001 for the Georgetown and Lovelace sites, October 2001 for half the BBA sites, and January 2002 for the remaining BBA sites.

- Patient enrollment (for the survey sample) will last 12 months.

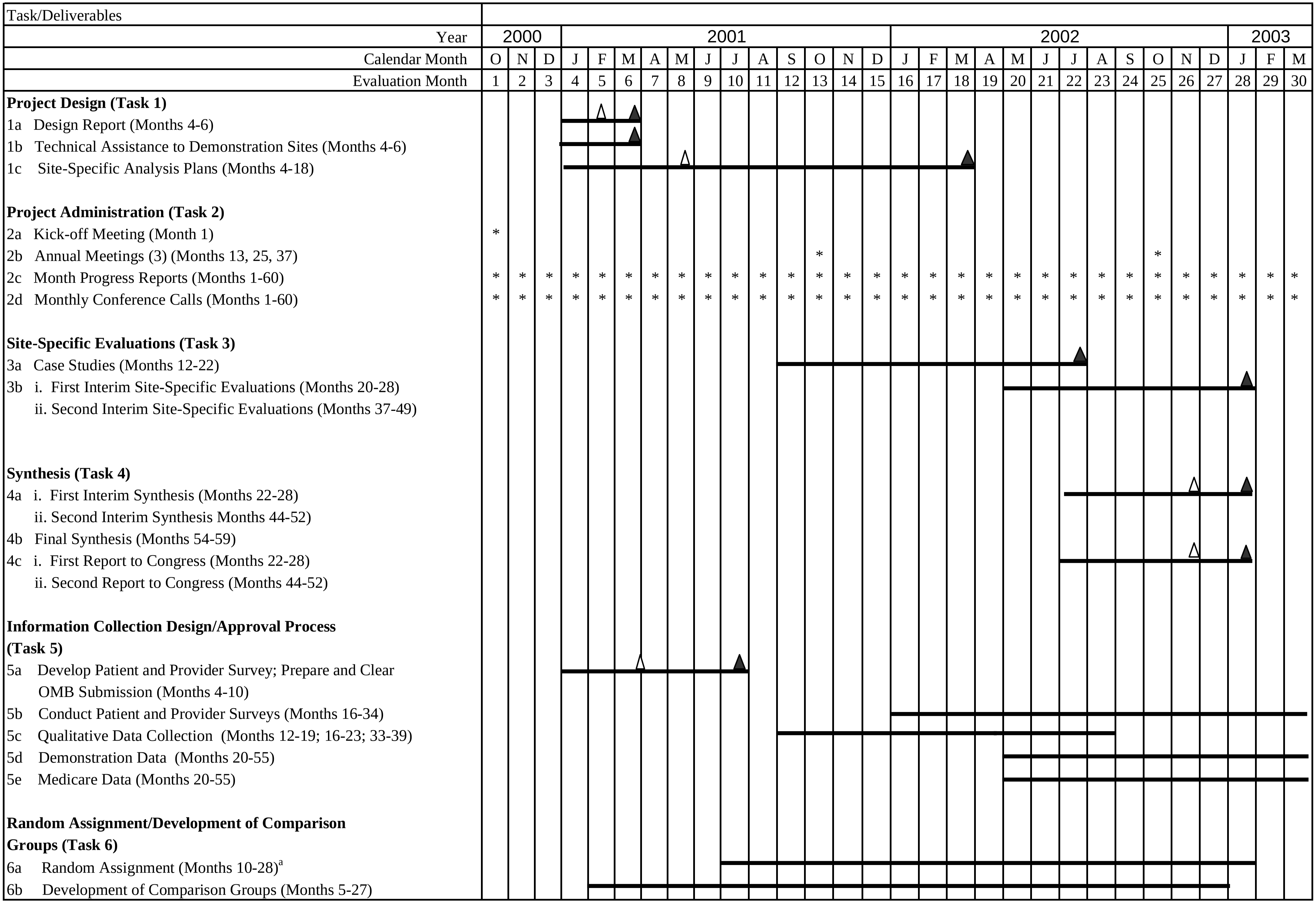
- We will survey patients 6 months after enrollment and follow them for a minimum of 12 months and a maximum of 24 months, using Medicare claims data.

- It appears that one site will require the development of a comparison group, but discussions with this site will determine whether it is possible to use a randomized design.

- We assume that MPR staff will have access to the HCFA data center to download enrollment and claims data.

FIGURE VI.1

PROJECT SCHEDULE



Project Design (Task 1)
1a Design Report (Months 4-6)
1b Technical Assistance to Demonstration Sites (Months 4-6)
1c Site-Specific Analysis Plans (Months 4-18)

Project Administration (Task 2)
2a Kick-off Meeting (Month 1)
2b Annual Meetings (3) (Months 13, 25, 37)
2c Month Progress Reports (Months 1-60)
2d Monthly Conference Calls (Months 1-60)

Site-Specific Evaluations (Task 3)
3a Case Studies (Months 12-22)
3b i. First Interim Site-Specific Evaluations (Months 20-28)
ii. Second Interim Site-Specific Evaluations (Months 37-49)

Synthesis (Task 4)
4a i. First Interim Synthesis (Months 22-28)
ii. Second Interim Synthesis Months 44-52)
4b Final Synthesis (Months 54-59)
4c i. First Report to Congress (Months 22-28)
ii. Second Report to Congress (Months 44-52)

Information Collection Design/Approval Process (Task 5)
5a Develop Patient and Provider Survey; Prepare and Clear OMB Submission (Months 4-10)
5b Conduct Patient and Provider Surveys (Months 16-34)
5c Qualitative Data Collection (Months 12-19; 16-23; 33-39)
5d Demonstration Data (Months 20-55)
5e Medicare Data (Months 20-55)

Random Assignment/Development of Comparison Groups (Task 6)
6a Random Assignment (Months 10-28)[a]
6b Development of Comparison Groups (Months 5-27)

* Meetings, conference calls Δ Draft deliverable ▲ Final deliverable

DRAFT

[a]Assuming last program starts by January 2002 and completes enrollment within 12 months.

FIGURE VI.1

PROJECT SCHEDULE

Task/Deliverables *(continued)*																														
Year	2003									2004												2005								
Calendar Month	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S
Evaluation Month	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Project Design (Task 1)																														
1a Design Report (Months 4-6)																														
1b Technical Assistance to Demonstration Sites (Months 4-6)																														
1c Site-Specific Analysis Plans (Months 4-18)																														
Project Administration (Task 2)																														
2a Kick-off Meeting (Month 1)																														
2b Annual Meetings (3) (Months 13, 25, 37)							*																							
2c Month Progress Reports (Months 1-60)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
2d Monthly Conference Calls (Months 1-60)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Site Specific Evaluations (Task 3)																														
3a Case Studies (Months 12-22)																														
3b i. First Interim Site-Specific Evaluations (Months 20-28)																														
ii. Second Interim Site-Specific Evaluations (Months 37-49)							━	━	━	━	━	━	━	━	━	━	━	━	▲											
Synthesis (Task 4)																														
4a i. First Interim Synthesis (Months 22-28)																														
ii. Second Interim Synthesis Months 44-52)														━	━	━	━	━	━	△	━	▲								
4b Final Synthesis (Months 54-59)																								━	━	━	△	━	▲	
4c i. First Report to Congress (Months 22-28)																														
ii. Second Report to Congress (Months 44-52)														━	━	━	━	━	━	△	━	▲								
Information Collection Design/Approval Process (Task 5)																														
5a Develop Patient and Provider Survey; Prepare and Clear OMB Submission (Months 4-12)																														
5b Conduct Patient and Provider Surveys (Months 16-34)	━	━	━	━																										
5c Qualitative Data Collection (Months 12-19; 16-23; 33-39)			━	━	━	━	━	━	━																					
5d Demonstration Data (Months 20-55)	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━					
5e Medicare Data (Months 20-55)	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━	━					
Random Assignment/Development of Comparison Groups (Task 6)																														
6a Random Assignment (Months 10-28)[a]																														
6b Development of Comparison Groups (Months 5-27)																														

* Meetings, conference calls △ Draft deliverable ▲ Final deliverable

[a] Assuming last program starts by January 2002 and completes enrollment within 12 months.

C. MANAGEMENT PLAN

The following task list illustrates the interrelatedness of project reports and data collection activities. The team member responsible for each task is listed at the end of the task description.

Project Design (Task 1)	
1a	**Design report.** Framework design based on evaluation proposal, work conducted for Coordinated Care design project, and input from HCFA at kick-off meeting. Draft due M5, final due M6 **[Brown]**
1b	**Technical assistance to demonstration sites.** Telephone discussions. Site-specific memorandum on adequacy of experimental design and site data collection based on early discussions with sites. Due M6 **[Brown]**
1c	**Site-specific analysis plans.** Memoranda indicating site-specific variations in evaluation design. Draft due M8; final due after early case study calls (Task 5c, M12-19) **[Brown]**
Site Specific Evaluations (Task 3)	
3a	**Case studies.** Based on site proposals, other materials prepared by site since award, protocol-guided telephone discussions with key site staff (Task 5c, M12-19). Reports due 6M after site enrollment begins **[Schore]**
3bi	**First interim site-specific evaluations.** Based on site visits (Task 5c, M16-23), patient intake forms for patients enrolling in the first 9M of operations, 6M patient survey data and Medicare data (2M followup) for patients enrolling in the first 4M of operations, and cost/other site data from HCFA. Reports due 12M after site enrollment begins **[Brown]**
3bii	**Second interim site-specific evaluations.** Based on protocol-guided telephone calls (Task 5c, M33-39), patient intake data for all patients, patient and provider surveys conducted through M34, Medicare data (12M followup) for patients enrolled in the first 12M of operations, and site cost/other site data from HCFA. Reports due 33M after site enrollment begins **[Brown]**
Synthesis (Task 4)	
4ai	**First interim synthesis.** Synthesis of first interim site evaluations (Task 3bi). Preliminary impact estimates, cross-site comparisons. PD/task leader for quality analysis primary authors. Draft due M26. Will discuss at second annual meeting (Task 2b, M25). Final due M28 **[Brown]**
4aii	**Second interim synthesis.** Synthesis of second interim site evaluations (Task 3bii). Preliminary impact estimates, cross-site comparisons. PD/task leader for quality analysis primary authors. Draft due M50. PD/task leader will meet with HCFA to discuss in M50. Final due M52 **[Brown]**
4b	**Final synthesis.** Update all site evaluations with remaining survey and Medicare data. Draft due M57. PD/task leader will meet with HCFA to discuss in M58. Final due M59. **[Brown]**
4ci	**First report to Congress.** Section II of first synthesis (Task 4ai). Have allowed for several rounds of HCFA's comments on this high-visibility report. Due M28 **[Brown]**
4cii	**Second report to Congress.** Section II of second synthesis (Task 4aii). Have allowed for several rounds of HCFA's comments on this high-visibility report. Due M52 **[Brown]**
Information Collection Design/Approval Process (Task 5)	
5a	**Develop patient and provider surveys.** Includes preparing and clearing OMB submission. Draft surveys to HCFA by M6; draft OMB submission to OMB for *Federal Register* notice M7, pretest and revise submission, final to OMB M8, clear OMB by M10 **[Chen, Ensor]**
5b	**Conduct patient and provider surveys.** Patient survey administered 6 months after enrollment. Provider survey administered 9 months after site begins enrollment (M16-34) **[Ensor]**
5c	**Qualitative data collection.** Case study telephone calls (M12-19), site visits (M16-23), telephone follow-up calls (M33-39). Protocols developed with design report, based on materials in evaluation proposal, modified with what is learned in each site contact. Assume protocols do not have to be cleared through OMB **[Schore]**
5d	**Demonstration data.** Assume HCFA supplies site cost and staff hours reports; beneficiary enrollment/ disenrollment records. Collect for first and second interim site evaluations and final synthesis **[Khan]**
5e	**Medicare data.** Assume 4-month lag between date of service and claim on Medicare files. Download prior to first and second interim site evaluations and final synthesis. Claims extracts and constructed variables developed prior to first download **[Khan]**
Random Assignment/Development of Comparison Groups (Task 6)	
6a	**Random assignment.** Conducted by MPR for 16 sites, after receipt of patient intake forms **[Brown]**
6b	**Development of comparison groups.** Based on Medicare claims data for 1 site **[Brown]**

TABLE VI.1

SCHEDULE OF DELIVERABLES FOR THE EVALUATION OF
THE MEDICARE COORDINATED CARE DEMONSTRATION

Item	Task	Deliverable Description	Period of Performance (Project Month)[a]	Due Date (Project Month)
1	1a	Draft design report	4-5	5
2	1a	Final design report	6	6
3	1b	Evaluation of Site Methodology	4-8	6
4	1c	Draft site-specific analysis plans	4-18	8
5	1c	Final site-specific analysis plans	4-18	18
6	2a	Kick-off meeting	1	1
7	2b	Annual meetings	Yearly	13, 25, 37
8	2c	Monthly progress reports	Monthly	Monthly
9	2d	Monthly conference calls	Monthly	Monthly
10	3a	Site case studies	12-22	6 months after site enrollment begins
11	3bi	First interim site-specific evaluation	20-28	12 months after site enrollment begins
12	3bii	Second interim site-specific evaluation	37-49	33 months after site enrollment begins
13	4ai	Draft first interim synthesis	22-26	26[b]
14	4ai	Final first interim synthesis	22-28	28[b]
15	4aii	Draft second interim synthesis	44-50	50[b]
16	4aii	Final second interim synthesis	44-52	52[b]
17	4b	Draft final synthesis	54-57	57
18	4b	Final synthesis	54-59	59
19	4ci	First report to Congress	22-28	28[b]
20	4cii	Second report to Congress	44-52	52[b]
21	5a	Draft survey instruments	4-6	6
22	5b	OMB approval package for planned information collections	4-10	7

[a] Project month refers to the number of months after the award of MPR's contract (September 29, 2000), where October 2000 is month 1.

[b] Assumes first BBA program starts enrolling in July 2001 (month 10).

REFERENCES

Aliotta, Sherry L. "Components of a Successful Case Management Program." *Managed Care Quarterly*, vol. 4, no. 2, 1996, pp. 38-45.

American Healthways. *Standards for Disease Management Programs.* Nashville, TN: American Healthways, 1999.

American Diabetes Association. "New Tools for Measuring and Improving Diabetes Care." [http://www.diabetes.org/councils/spring/spring99/newtool.html]. July 17, 2000.

Archibald, Nancy, and Randall Brown. "Medicare Coordinated Care Demonstration: Characteristics of Applicant Programs." Princeton, NJ: Mathematica Policy Research, Inc., December 2000.

Aubert, Ronald E., William H. Herman, Janice Waters, William Moore, David Sutton, Bercedis L. Peterson, Cathy M. Bailey, and Jeffrey P. Koplan. "Nurse Case Management to Improve Glycemic Control in Diabetic Patients in a Health Maintenance Organization. A Randomized, Controlled Trial." *Annals of Internal Medicine*, vol. 129, no. 8, October 15, 1998, pp. 605-612.

Aubry, Barbara. "Bolstering Disease Management Programs." *Healthplan*, July-August 2000, pp. 11-12.

Beck, Arne, John Scott, Patrick Williams, Barbara Robertson, Deborrah Jackson, Glenn Gade, and Pamela Cowan. "A Randomized Trial of Group Outpatient Visits for Chronically Ill Older HMO Members: The Cooperative Health Clinic." *Journal of the American Geriatrics Society*, vol. 45, no. 5, May 1997, pp. 543-549.

Block, Gladys, Anne M. Hartman, Connie M. Dresser, Margaret D. Carroll, Janet Gannon, and Lilly Gardner. "A Data-Based Approach to Diet Questionnaire Design and Testing." *American Journal of Epidemiology*, vol. 124, no. 3, 1986, pp. 453-469.

Bodenheimer, Thomas. "Disease Management—Promises and Pitfalls." *New England Journal of Medicine*, vol. 340, no. 15, April 15, 1999, pp. 1202-1205.

Boult Chad, Lisa Boult, Lynne Morishita, Stanley L. Smith, and Robert L. Kane. "Outpatient Geriatric Evaluation and Management." *Journal of the American Geriatrics Society*, vol. 46, no. 3, March 1998, pp. 296-302.

Brown, Randall S. "Demonstration Design for the Medicare Coordinated Care Project." Princeton, NJ: Mathematica Policy Research, Inc., January 2000.

Case Management Society of America. *Standards of Practice for Case Management*. Little Rock, AR: CMSA, 1995.

Center for Studying Health System Change. "Community Tracking Study: Physician Survey Instrument." Washington, DC: CSHSC, September 1997.

Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System: BRFSS Questionnaires." [http://www.cdc.gov/nccdphp/brfss/brfsques.htm]. February 12, 2001.

Chen, Arnold, Randall Brown, Nancy Archibald, Sherry Aliotta, and Peter Fox. "Best Practices in Coordinated Care." Princeton, NJ: Mathematica Policy Research, Inc., February 29, 2000.

Chin, Marshall H., and Lee Goldman. "Factors Contributing to the Hospitalization of Patients with Congestive Heart Failure." *American Journal of Public Health*, vol. 87, no. 4, April 1997, pp. 643-648.

Ciemnecki, Anne B., and Karen CyBulski. "Interviewing Populations with Disabilities by Telephone: Survey Design and Operations." Paper presented at the 55th American Association for Public Opinion Research Annual Conference, Portland, Oregon, May 18-21, 2000.

Clark, Noreen M., and Molly Gong. "Management of Chronic Disease by Practitioners and Patients: Are We Teaching the Wrong Things?" *British Medical Journal*, vol. 320, February 26, 2000, pp. 572-575.

Creditor, Morton L. "Hazards of Hospitalization in the Elderly." *Annals of Internal Medicine*, vol. 118, no. 3, February 1, 1993, pp. 219-223.

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.

Dehejia, Rajeev H., and Sadek Wahba. "Propensity Score Matching Methods for Non-Experimental Causal Studies." Working paper 6829, Cambridge, MA: National Bureau of Economic Research, December1998. Also at http://www.nber.org/papers/w6829. Accessed December 1999.

Dehejia, Rajeev H., and Sadek Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." Journal of the American Statistical Association, vol. 94, no. 448, 1999, pp. 1053-1062.

Dixon, Melissa K., Pamela B. Kirschner, Helen K. Edelberg, John Z. Ayanian, and Jeanne Y. Wei. "Physician Perceptions of HMO Care for Older Persons." *Journal of the American Geriatrics Society*, vol. 48, no. 6, June 2000, pp. 607-612.

Donabedian, Avedis. "The Role of Outcomes in Quality Assessment and Assurance." *Quality Review Bulletin*, November 1992, pp. 356-362.

Donabedian, Avedis. *Explorations in Quality Assessment and Monitoring, Volume 2, the Criteria and Standards of Quality*. Ann Arbor, MI: Health Administration Press, 1982.

Donabedian, Avedis. *Explorations in Quality Assessment and Monitoring. Volume I, the Definition of Quality and Approaches to Its Assessment.* Ann Arbor, MI: Health Administration Press, 1980.

Duan, Naihua, Willard Manning, Carl Morris, and Joseph Newhouse. "A Comparison of Alternative Models for the Demand for Health Care." R-2754-HHS. Santa Monica, CA: The RAND Corporation, January 1982.

Ferrans, C.E., and M.J. Powers. "Quality of Life Index: Development and Psychometric Properties." *Advances in Nursing Science,* vol. 8, no. 1, 1985, pp. 15-24.

Fleming, K.C., J.M. Evans, D.C. Weber, and D.S. Chutka. "Practical Functional Assessment of Elderly Persons: A Primary-Care Approach." Mayo Clinic Proceedings. vol. 70, 1995, pp. 890-910.

Fox, Peter. "Screening: The Key to Early Intervention for High-Risk Seniors." *Healthplan,* November-December 2000, pp. 56-61.

Greenfield, Sheldon, Sherry H. Kaplan, Rebecca A. Silliman, L. Sullivan, Willard Manning, Ralph D'Agostino, Daniel E. Singer, and David M. Nathan. "The Uses of Outcomes Research for Medical Effectiveness, Quality of Care, and Reimbursement in Type II Diabetes." *Diabetes Care,* vol. 17, 1994, pp. 32-39.

Guyatt, Gordon H., Leslie B. Berman, Marie Townsend, Stewart O. Pugsley, and Larry W. Chambers. "A Measure of Quality of Life for Clinical Trials in Chronic Lung Disease." Thorax, vol. 42, 1987, pp. 773-778.

Hagland, Mark. "Integrating Disease Management." *Healthplan,* January-February 2000, pp. 43-46.

Havranek, Edward P., Gerard W. Graham, Zhaoxing Pan, and Brian Lowes. "Process and Outcome of Outpatient Management of Heart Failure: A Comparison of Cardiologists and Primary Care Providers." *American Journal of Managed Care,* vol. 2, no. 7, 1996, pp. 783-789.

Heckman, J.J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." Annals of Economic and Social Measurement, vol. 5, 1976, pp. 475-492.

Holman, Holman, and Kate Lorig. "Patients as Partners in Managing Chronic Disease." *British Medical Journal,* vol. 320, February 26, 2000, pp. 526-527.

Krumholz, Harlan M., David W. Baker, Carol M. Ashton, Sandra B. Dunbar, Gottlieb C. Friesinger, Edward P. Havranek, Mark A. Hlatky, Marvin Konstam, Diana L. Ordin, Ileana Pina, Bertram Pitt, and John A. Spertus. "Evaluating Quality of Care for Patients with Heart Failure." *Circulation,* vol. 101, no. 12, March 28, 2000, pp. 122-140.

Landefeld C.S., R.M. Palmer, D.M. Kresevic, R.H. Fortinsky, and J. Kowal J. "A Randomized Trial of Care in a Hospital Medical Unit Especially Designed to Improve the Functional Outcomes of Acutely Ill Older Patients." *New England Journal of Medicine*, vol. 332, no. 20, May 18, 1995, pp. 1338-1344.

LeDuc, Nicole, Terry Nan Tannenbaum, Howard Bergman, et al. "Compliance of Frail Elderly with Health Services Prescribed at Discharge from an Acute-Care Geriatric Ward." *Medical Care*, vol. 36, no. 6, 1998, pp. 904-914.

Lorig, Kate, Halsted Holman, David Sobel, Diana Laurent, Virginia Gonzalez, and Marian Minor. *Living a Healthy Life with Chronic Conditions: Self-Management of Heart Disease, Arthritis, Diabetes, Asthma, Bronchitis, Emphysema, and Others.* 2nd edition. Palo Alto, CA: Bull Publishing, 2000.

Lorig, Kate, David Sobel, Anita Stewart, et al. "Evidence Suggesting that a Chronic Disease Self-Management Program Can Improve Health Status While Reducing Hospitalization." *Medical Care*, vol. 37, no. 1, 1999, pp. 5-14.

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.

Mahoney, Jane E., Mari Palta, Jill Johnson, Muhammed Jalaluddin, Shelly Gray, Soomin Park, and Mark Sager. "Termporal Association Between Hospitalization and Rate of Falls After Discharge." *Archives of Internal Medicine*, vol. 160, October 9, 2000, pp. 2788-2795.

Manian, Frank A. "Whither Continuity of Care?" *New England Journal of Medicine*, vol. 340, no. 17, April 29, 1999, pp. 1362-1363.

Manning, Willard G. "The logged dependent variable, heteroscedasticity, and the retransformation problem." *Journal of Health Economics*, vol. 17, 1998, pp. 283-295.

Monane, Mark, Dipika M. Matthias, Becky A. Nagle, and Miriam A. Kelly. "Improving Prescribing Patterns for the Elderly Through an Online Drug Utilization Review Intervention: A System Linking the Physician, Pharmacist, and Computer." *JAMA*, vol. 280, no. 14, October 14, 1998, pp. 1249-52.

Mullahy, John. "Much Ado About Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics." *Journal of Health Economics*, vol. 17, 1998, pp. 247-281.

Naylor, Mary, Dorothy Brooten, Robert Jones, Risa Lavizzo-Mourey, Mathy Mezey, and Mark Pauly. "Comprehensive Discharge Planning for the Hospitalized Elderly." *Annals of Internal Medicine*, vol. 120, no. 12, June 15, 1994, pp. 999-1006.

Oldridge, Neil, Gordon Guyatt, Norman Jones, Jean Crow, Joel Singer, David Feeny, Robert McKelvie, Joanne Runions, David Streiner, and George Torrance. "Effects on Quality of Life with Comprehensive Rehabilitation After Acute Myocardial Infarction." *American Journal of Cardiology*, vol. 67, 1991, pp. 1084-1089.

Patrick, Donald L., and Richard A. Deyo. "Generic and Disease-Specific Measures in Assessing Health Status and Quality of Life." *Medical Care.* vol. 27, no. 3, suppl., March 1989, pp. S217-S232.

Rector, Thomas, and Patricia Venus. "Judging the Value of Population-Based Disease Management." *Inquiry*, vol. 36, summer 1999, pp. 122-126.

Rector, Thomas S. and Jay N. Cohn. "Assessment of Patient Outcome with the Minnesota Living with Heart Failure Questionnaire: Reliability and Validity During a Randomized, Double-blind, Placebo-Controlled Trial of Pimobendan." *American Heart Journal*, vol. 124, October 1992, pp. 1017-1025.

Rich, Michael, Valerie Beckham, Carol Wittenberg, Charles Leven, Kenneth Freedland, and Robert Carney. "Multidisciplinary Intervention to Prevent the Readmission of Elderly Patients with Congestive Heart Failure." *New England Journal of Medicine*, vol. 333, no. 18, November 12, 1995, pp. 1190-1195.

Roter, Debra, Judith Hall, Rolande Merisca, et al. "Effectiveness of Interventions to Improve Patient Compliance." *Medical Care*, vol. 36, no. 8, 1998, pp. 1138-1161.

Safran, Dana Gelb, Mark Kosinski, Alvin R. Tarlov, William H. Rogers, Deborah A. Taira, Naomi Lieberman, and John E. Ware. "The Primary Care Assessment Survey: Tests of Data Quality and Measurement Performance." Medical Care, vol. 36, no. 5, pp. 728-739.

Sager, Mark A and M.A. Rudberg. "Functional Decline Associated with Hospitalization for Acute Illness." *Clinics in Geriatric Medicine*, vol. 14, 1993, pp. 669-679.

Schore, Jennifer, Randall Brown, Valerie Cheh, and Barbara Schneider. "Costs and Consequences of Case Management for Medicare Beneficiaries." Princeton, NJ: Mathematica Policy Research, Inc., April 30, 1997.

Spertus, John A., Jennifer A. Winder, Timothy A. Dewhurst, Richard A. Deyo, Janice Prodzinski, Mary McDonell, and Stephan D. Fihn. "Development and Evaluation of the Seattle Angina Questionnaire: A New Functional Status Measure for Coronary Artery Disease." *Journal of the American College of Cardiology*, vol. 25, no. 2, February 1995, pp. 333-341.

Spiegel, Jane S., Lisa V. Rubenstein, Bonnie Scott, and Robert H. Brook. "Who Is the Primary Physician?" *New England Journal of Medicine*, vol. 308, no. 20, May 1983, pp. 1208-1212.

Stewart, Simon, John E. Marley, and John D. Horowitz. "Effects of a Multidisciplinary, Home-Based Intervention on Unplanned Readmissions and Survival Among Patients with Chronic Congestive Heart Failure: A Randomised Controlled Study." *Lancet,* vol. 354, no. 9184, September 25, 1999, pp. 1077-1083.

Stewart, Simon, Sue Pearson, and John D. Horowitz. "Effect of a Home-Based Intervention Among Patients with Congestive Heart Failure on Readmission and Mortality." Archives of Internal Medicine, vol. 159, 1999, pp. 257-261.

The DCCT Research Group. "Reliability and Validity of a Diabetes Quality-of-Life Measure for the Diabetes Control and Complications Trial." *Diabetes Care*, vol. 11, no. 9, October 1988, pp. 725-732

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.

Tu, Shin-Ping, Mary B. McDonell, John A. Spertus, Bonnie G. Steele, and Stephan D. Fihn. "A New Self-Administered Questionnaire to Monitor Health-Related Quality of Life in Patients with COPD." *Chest*, vol. 112, no. 3, September 1997, pp. 614-622.

U.S. Department of Health and Human Services, Health Care Financing Administration, Office of Strategic Planning. "High-Cost Users of Medicare Services." *Health Care Financing Review*, statistical supplement, 1998, pp. 42-43.

U.S. Department of Health and Human Services, National Center for Health Statistics. *National Health Interview Survey, Supplement on Aging, 1984*. Hyattsville, MD: HHS, NCHS, 1984.

Valenti, L., L. Lim, R.F. Heller, et al. "An Improved Questionnaire for Assessing Quality of Life After Acute Myocardial Infarction." *Quality of Life Research*, vol. 5, 1996, pp. 151-161.

Vernarec, Emil. "Health Care Power Shifts to the People." *Business and Health: The State of Health Care in America 1999*, pp. 8-13.

Wagner, Edward H., Brian T. Austin, and Michael Von Korff. "Organizing Care for Patients with Chronic Illness." *Milbank Quarterly*, vol. 74, no. 4, 1996, pp. 511-544.

Ware, John E., Jr., Mark Kosinski, and Susan D. Keller. "A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity." *Medical Care*, vol. 34, no. 3, March 1996, pp. 22-223.

Wasson, John, Catherine Gaudette, Fredrick Whaley, Arthus Sauvigne, Pricilla Baribeau, and Gilbert Welsh. "Telephone Care as a Substitute for Routine Clinic Follow-up." *Journal of the American Medical Association*, vol. 267, no, 13, April 1, 1992, pp. 1788-1793.

Welch G.W., A.M. Jacobson, and W.H. Polonsky. "The Problem Areas in Diabetes Scale: An Evaluation of Its Clinical Utility." *Diabetes Care*, vol. 20, 1997, pp. 760-766.

Wells, Kenneth B., Cathy Sherbourne, Michael Schoenbaum, Naihua Duan, Lisa Meredith, Jürgen Unützer, Jeanne Miranda, Maureen F. Carney, and Lisa V. Rubenstein. "Impact of Disseminating Quality Improvement Programs for Depression in Managed Primary Care: A Randomized Controlled Trial." *JAMA*, vol. 283, no. 2, January 12, 2000, pp. 212-220.

Williams, Mark. "Chronic Care Clinics: Why Don't They Work?" *Journal of the American Geriatric Society*, vol. 47, no. 7, July 1999, pp. 908-909.