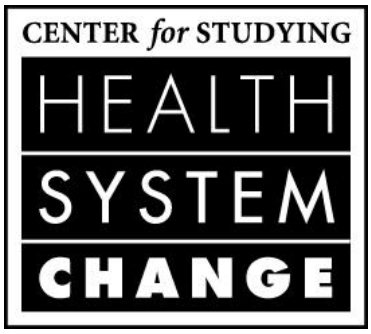

Site Definition and Sample Design for the Community Tracking Study

Charles E. Metcalf
Peter Kemper
Linda T. Kohn
Jeremy D. Pickreign

Technical Publication No.

1

October 1996

This is one of a series of technical documents that have been done as part of the Community Tracking Study being conducted by the Center for Studying Health System Change. The study will examine changes in the local health systems and the effects of those changes on the people living in the area.

The Center welcomes your comments on this document. Write to us at 600 Maryland Avenue, SW, Suite 550, Washington, DC 20024-2512 or visit our web site at www.hschange.com.

The Center for Studying Health System Change is supported by The Robert Wood Johnson Foundation and is affiliated with Mathematica Policy Research, Inc.

CONTENTS

TABLES

Table **Page**

I. OVERVIEW OF THE COMMUNITY TRACKING STUDY[1]

The Community Tracking Study is a national study of the rapidly changing health care market and the effects of these changes on people. The study will develop an information base designed to track and analyze change. This paper sets forth the sample design for the study. It reviews a wide range of issues concerning what communities are to be selected for the Community Tracking Study and how their selection relates to an overall set of sample designs for surveys of households and the family insurance units they comprise, employers, physicians, physician groups and organizations, hospitals, insurers, and managed care plans.

This chapter provides a brief overview of the Community Tracking Study. Chapter II considers issues related to site definition and strategies for their selection. Chapter III addresses sampling issues for the design of a household/family survey of health care consumers. It also integrates this design into the site sample design from Chapter II. Chapter IV extends the design principles established for the household survey to the physician survey and summarizes the design for this survey. Chapter V provides preliminary design considerations for the employer survey, to be designed by the RAND Corporation to serve a range of research objectives -- including those of the Community Tracking Study -- related to employer-provided health insurance.

This paper reflects the final designs adopted for site selection and for the household and physician surveys, except for sample size adjustments that might occur after data collection begins. The designs for the remaining surveys continue to evolve and are not yet final.

[1] The study is described in more detail in P. Kemper, D. Blumenthal, J.M. Corrigan, P.J. Cunningham, S.M. Felt, J.M. Grossman, L.T. Kohn, C.E. Metcalf, R.F. St. Peter, R.C. Strouse, and P.B. Ginsburg. "The Design of the Community Tracking Study: A Longitudinal Study of Health System Change and Its Effects on People." *Inquiry*, vol. 33, summer 1996, pp. 195-206.

The Community Tracking Study has three objectives:

1. ***Tracking Changes in Health Systems.*** The study's first objective is to document health system changes through intensive study of a selected number of communities. The major changes that have been reported in the health system include consolidation of the market at all levels (medical groups, hospitals, insurers, and health plans); vertical integration of providers (for example, hospitals and physicians) and of insurers and providers; increased risk sharing by providers; growth of large, national, for-profit health care enterprises; and adoption of new techniques for clinical care management (clinical information systems, quality improvement techniques, utilization management, and so forth).

2. ***Tracking Changes in Outcomes.*** The second objective of the study is to monitor the effects of health system change on people by tracking indicators of health system outcomes. This change could have important favorable or unfavorable consequences for individuals. It may alter their access to care, service use and delivery, and quality and cost of care -- referred to here as "outcomes."

3. ***Understanding the Effect of Health System Change on Outcomes.*** Given the absence of systematic, relevant local and national information, documenting changes in health systems and outcomes is of great interest in its own right. This documentation also lays the foundation for accomplishing the third objective of study: to understand how differences in health systems are related to differences in outcomes. This will be done by analyzing -- qualitatively and quantitatively -- the relationship between health systems and outcomes.

Central to the design of the study is its focus on communities. This focus is based on the fact that health care delivery is primarily local and differs because of history, culture, and state policy. Therefore, information at the local market level is needed to analyze and understand institutional changes in the delivery system and their effects on people.

Health care systems in selected communities will be followed over time through a variety of data collection activities. Site visits are planned to provide an understanding of health systems and the dynamics of change through interviews with key actors in the system. Surveys of health care organizations (insurers and health plans, hospitals, and physician organizations) will provide systematic information about the health care market in each community. Surveys of households, physicians, and employers in the same communities will provide information on

the outcomes defined earlier. By combining data on health systems and outcomes in the same communities, the study is designed to relate differences in outcomes to differences in health systems.

II. SITE SAMPLE DESIGN

What communities to study is one of the most significant decisions facing the Community Tracking Study team. The decision will affect the study for its duration and will determine how the results are received, as well as their ultimate credibility. Three issues are central to the site sample design: how sites are defined, how many are studied, and how they are selected.

A. DEFINITION OF SITES

When the health system or a market for a particular service is defined, it is important to distinguish analytically between services produced in a community and services consumed by residents of the community, including those provided outside the community's geographic boundaries. For example, to analyze practice patterns of physicians caring for low-income people, we could survey all physicians within a metropolitan statistical area (MSA), asking about the care they provide to low-income populations. Alternatively, we could identify low-income people through a community survey that asks which physicians they use. Subsequently, we could survey these physicians regarding their practice patterns. In the first example, we would be analyzing physicians *providing services in the geographic area* defining the site; in the second example, we would be analyzing physicians *serving the low-income people who live in the geographic area, even if the services were provided outside the area*.

Conceptually, we would like to define the health system according to who serves the population, that is, the actual market used by the population living in the area. What providers does the population use? What services are provided? We are interested in knowing the extent to which these markets change over time along with changes in the organization and structure of health care.

However, defining the market and collecting data based on what providers are located in the area is more operationally practical, even though residents may obtain services from a larger group of providers, and providers may serve a more dispersed population.[2] In the process of defining sites, therefore, it is desirable for the services *provided within* the geographic area and the services *provided to* the population residing in the geographic area to overlap as much as possible, with little import/export activity.

In previous studies, researchers have used multiple approaches to define market areas, including:

- Administrative units, such as MSA, county, or zip code
- Radius methods, in which a hospital (or other entity) location is the center, and a fixed radius is drawn around it. By creating clusters of overlapping hospital circles, researchers can build an overall market area.
- Clustering methods, in which patient origin data are used to define areas on the basis of the zip codes from which some predetermined percentage of patients are drawn (for example, 65 percent, 75 percent, 90 percent) by the providers of interest
- Markets serving the population in a community. As discussed, an alternative is to use a community survey to identify where people obtain care and define the market based on this information. Unlike radius or clustering methods that start from the provider perspective, this approach starts from the consumer perspective. (This approach still requires a geographic definition of "community" from which to draw the sample, implicitly defining the market.)

All these approaches have strengths and weaknesses; the choice must be based on a study's analytic goals. Methods that are based on current providers (radius measures, clustering methods, community survey approach) might not reflect new entrants into a market, an important consideration in tracking change over time. Although approaches based on administrative units do not face this constraint, they might bear little relationship to the market

[2] As delivery patterns change, so do patient travel patterns, and therefore, import/export patterns. There are two possible ways to identify the extent of import/export activities within an area. We can learn about patient travel patterns through the household survey. We can also use provider surveys to ask about the area from which patients are drawn.

for health care. Furthermore, the Community Tracking Study is not looking at a single sector of the health care system but at the *entire* system. Methods based solely on hospitals might not accurately reflect the market for health plans or physicians.

In this study, the method selected to define overall market areas should meet several criteria:

- It should be applicable to the whole health system -- hospitals, physicians, health plans, and so forth.
- It should be consistent over time; any changes to the definition over time will make valid comparisons difficult.
- It should be applicable to all communities across the nation.
- It should constitute the first-stage selection in sample design for the household, physician, employer, and organization surveys.

Because we want site definitions to remain constant over time, we have defined sites on the basis of counties or groups of counties. We have also adapted conventionally accepted definitions of statistical and economic areas. Our site definitions are based on MSAs as defined by the Office of Management and Budget and the nonmetropolitan portions of economic areas as defined by the Bureau of Economic Analysis (BEAEAs).[3] Both MSAs and BEAEAs are defined as counties or aggregates of counties, except in New England.[4] MSAs are readily understood and match the definition used by other analysts; BEAEAs, while not used universally, provide a rational and independent basis for assigning nonmetropolitan areas to defined economic markets. Use of counties also facilitates comparisons with secondary data.

---

[3] See *1990 Census of Population and Housing, Supplementary Reports, Metropolitan Areas as Defined by the Office of Management and Budget*, June 30, 1993 (1990 CPH-S-1), and Kenneth P. Johnson, "Redefinition of the BEA Economic Areas," *Survey of Current Business*, February 1995, pp.75-81. In the event of future modification of MSA and BEAEA boundaries, we will retain the site boundaries in force at the time of site definition.

[4] In New England, where MSAs do not conform with county lines, we have instead used New England County Metropolitan Areas, a conventionally accepted set of county approximations to MSAs used in this region.

The disadvantages of using county-based definitions involve variations in county size and composition across the country. Furthermore, county boundaries may be arbitrary relative to the underlying health care markets we are trying to characterize.

Each MSA was considered a self-contained site eligible for inclusion in the study. However, 18 large, complex metropolitan areas are designated as Consolidated Metropolitan Statistical Areas (CMSAs). Each CMSA has two or more component Primary Metropolitan Statistical Areas (PMSAs) -- for example, Denver, Boulder-Longmont, and Greeley, Colorado; Cleveland and Akron, Ohio; and the 15 component PMSAs that make up the New York-Northern New Jersey-Long Island CMSA. Since the health care markets in these large metropolitan areas are likely to mirror this complexity, our objective was to define submarkets within an overall market and sample one or more submarkets of the larger CMSA, rather than to view the most complex areas as single sites for potential selection. In the definition of submarkets, the conceptual principle was identifying subareas in which the services provided and the services used by residents overlapped as much as possible. After reviewing available evidence on health care markets in the 18 CMSAs, we adopted the operational procedure of dividing the CMSAs into their PMSA component parts as sites eligible for independent selection, after combining PMSAs of less than 350,000 population with an adjacent PMSA.[5]

To define nonmetropolitan areas, we adapted the BEAEAs -- which divide the nation into areas inclusive of both metropolitan and nonmetropolitan counties -- by defining the nonmetropolitan portion of each BEAEA as a nonmetropolitan "site" eligible for

---

[5] The New York PMSA was split into New York City and its suburban counties; all other sites consisted of one or more PMSAs. Of the 18 CMSAs, 6 remained as single sites, 6 were split into two sites, and 6 were split into 3 or more sites. The largest CMSAs, such as New York or Los Angeles, were likely to have multiple component sites selected in the larger sample of sites discussed later in this paper. Systematic sampling techniques used for geographic stratification prevented a nonrepresentative "overload" of sites selected from a single CMSA.

selection.[6] This approach produced both contiguous sites, some forming donuts around an MSA, and isolated sites clustered around economic centers too small to be designated as MSAs.[7] Both types of nonmetropolitan sites were sampled on a probabilistic basis.[8]

B. NUMBER OF SITES

Determining the number of sites to study reflects trade-offs among the following:

- Studying a limited number of communities extensively
- Collecting data to measure characteristics of a larger number of sites, in order to empirically examine causal relationships between system change and effects on care delivery and consumers
- Structuring site and sample selection to enhance the generalizability of the findings to the nation as a whole

In conjunction with these issues are corresponding trade-offs involving allocation of data collection and analysis costs.

Communities differ along multiple dimensions, such as state policy, extent of market consolidation, local economic conditions, region of the country, size, degree of urbanization, and sociodemographic composition. As a result, it is important to observe changes for communities representing the full range of these dimensions if we are to be able to characterize change in the nation. The evolution of health systems is likely to depend heavily on the history

[6] BEAEAs that crossed state lines were broken into single-state components; BEAEAs with nonmetropolitan populations of less than 50,000 were merged with adjacent BEAEAs.

[7] Nonmetropolitan areas surrounding metropolitan areas are influenced by and use health care providers in these areas. In addition, health plans are likely to expand their market into surrounding nonmetropolitan areas. As a result, we made a conceptual distinction between "contiguous" nonmetropolitan areas -- such as nonmetropolitan counties within, say, 75 miles of the center of a MSA -- and the remaining "isolated" nonmetropolitan counties that are not contiguous to metropolitan areas.

and culture of a particular community. The smaller the number of sites, the more likely our findings will pertain -- or will be perceived as pertaining -- only to the unique communities studied rather than to broader areas. For example, the Minneapolis health care market is generally viewed as very advanced. The reasons for this perception are unclear, however. It may be due to the local culture and populist politics, the presence of a strong group of Fortune 500 employers, a homogeneous population, a close network of health system leaders, or perhaps other factors.

Any case study design must confront the difficulty of distinguishing changes of general importance to people from idiosyncratic characteristics in a particular community. The smaller the number of cases, the greater this difficulty. The solution -- greatly expanding the number of case studies -- was not realistic for cost reasons. However, we adopted two changes in the original Robert Wood Johnson Foundation (RWJF) site design, to mitigate the problem of idiosyncracies with small numbers of case studies.

The original RWJF concept was to conduct 10 case studies, presumably 8 metropolitan and 2 nonmetropolitan. We concluded that this configuration was too limited to provide a representative picture of the nation -- in a judgmental or statistical sense. Thus, our first proposed change was to expand the number of case study communities to 12, and to concentrate them all in metropolitan areas. Although this change meant that we would have no case studies in isolated nonmetropolitan areas, these areas vary so widely that results from two cases would have been extremely difficult to interpret. These changes yield an increase of 50 percent in the case studies of metropolitan areas. While there was no formal scientific basis for deciding on the number of intensive case study sites, 12 reflected a balance between the

[8] We considered jointly selecting metropolitan sites with their adjacent nonmetropolitan counties but ultimately adopted an independent selection strategy. This strategy selects a "contiguous" site independently of its associated metropolitan site.

benefits of studying a range of different communities and the costs of case studies. Twelve sites would be sufficient, for example, to permit stratification along two dimensions with three and four categories -- for example, high-, intermediate-, and low-market consolidation and four regions -- or, as ultimately adopted, a finer classification into 12 regional strata. Modified systematic sampling techniques permitted a rough balance of variations in city size as well.

While 12 is a nontrivial number of observations for an intensive case study design, the problem of limited generalizability inherent in such designs remains. Therefore, we augmented the sample of intensive case study sites with a larger sample of metropolitan and nonmetropolitan areas -- a sufficient number to provide representativeness of communities nationwide and to provide a benchmark for interpreting the representativeness of the intensively studied sites, but studied much less exhaustively. The case study sites will be examined through site visits and family samples that are sufficiently large to allow us to make site-specific estimates of change. However, the larger sample of sites will be examined through telephone surveys of insurance plans, hospitals, and physician organizations; secondary data; and smaller samples from the family, employer, and physician surveys. This larger sample of sites will play two important roles in the study.

First, it will allow us to assess the extent to which the intensive case study sites represent health system changes taking place in the nation broadly. Data from the surveys will provide measures of market characteristics that will enable us to analyze the evolution of markets over time. Specifically, we will be able to construct measures such as concentration of the health plan market, extent of capitation of hospital and physician services, extent of integration of health plans and delivery systems, concentration of physician groups, and so forth. This information will be used to characterize the type and extent of market change that has occurred. This will in turn permit us to determine whether the case study findings reflect

more general experience or simply idiosyncracies of the study sites. It will also enable us to track the nature and pace of market evolution over time. This tracking will provide an important, nationally representative backdrop for the richer market data from the intensive case study sites, enabling us to validate our findings.

Second, the low-intensity sites and the information they provide on health systems for a larger number of sites will permit analysis of the effect of health system change on people. We will be able to relate differences in individuals' access, service use, outcomes, and satisfaction to differences in market type. Technically, multivariate statistical techniques will be used to test for effects of the type of market on access, use, outcomes, and satisfaction. Without the larger sample of sites, such an analysis would be unlikely to yield meaningful results.

C. SITE SELECTION

Selecting sites is one of the most important elements of the sample design. The selection methodology must protect against the risk that the results are -- or are perceived to be -- simply a collection of case studies of unique market experiences. Because the study is longitudinal, we will have to live with the selection for the life of the project (or change it at great analytic cost).

1. Site Selection in Previous Studies

Health services researchers with a variety of different analytic objectives have frequently had to confront site selection issues. Some of these issues involve:

- ***Evaluations of Demonstrations.*** Researchers conducting demonstrations and evaluations frequently encounter site selection issues. This type of research ranges from small one-site demonstrations to large experiments like the RAND health insurance experiments and the Channeling experiment. A common criticism of these evaluations is that the results are not generalizable because the number of sites was too small or that the sites were specially selected and are thus not representative. This criticism can occur, for example, because sites are selected at least partly based on their willingness and ability to conduct a demonstration. This type of selection makes the sites unrepresentative of all sites in the nation and limits the researchers'

ability to generalize results. Similar issues arise in before-and-after evaluations of policy changes instituted by states. By the very nature of such change, the most active states, such as California or Minnesota, are represented in the selection factor (Robinson and Phibbs 1989; and Langa and Sussman 1993). Because we do not need sites to decide to participate in the study, we will not be subject to the form of selection bias and criticism common in demonstrations. However, as indicated, the study will face questions about the generalizability of the results.

- ***Data Availability.*** Researchers frequently have selected sites for study because site data are available. For example, cost report data for hospitals in California were available and easily accessible (Melnick and Zwanziger 1988). Other researchers have taken advantage of the availability of a special, one-time, data set (Campbell and Fournier 1993) or a subset of data from a larger study (Shortell and Hughes 1988). Because much of the data being used in the Community Tracking Study are primary data, availability will not generally be a constraint.

- ***Study of the Effects of Change.*** Researchers have often sought to analyze the effect of a policy or market change by exploiting natural variation across communities. In these analyses, sites are selected to represent opposite and extreme ends of a continuum. For example, researchers have compared competition in California's health market to that of states with strong health regulation (Robinson and Luft 1988; and Anderson et al. 1993); the presence of state hospital rate setting (Cromwell 1987); or degree of HMO activity within the market (Luft et al. 1986). Such an approach could be adopted for the Community Tracking Study to analyze the effects of "market consolidation." Two types of sites would be selected -- the most consolidated and least consolidated -- with no intermediate sites included. The analysis would then focus on differences in outcomes between these two types of sites. The disadvantage of this approach is that it focuses on a single dimension of health system change and one that must be identified in advance, placing a premium on correct prediction of the essential difference in health systems of the future. Nonetheless, the strategy of placing greater emphasis on the extremes of health system change through oversampling is a potential strategy.

- ***Study of Health Care Institutions.*** Numerous published and unpublished studies of some aspect of health care have been conducted using a site visit or case study approach, often with the objective of documenting some aspect of health system change. Some studies selected sites of convenience (for example, the city in which a university is located), but most selected sites judgmentally to try to capture diversity in the sites, exemplary programs, and so forth. The Community Snapshots Project of health system change selected sites judgmentally to include those with a diverse set of characteristics. Two other approaches are also possible. One is to restrict greatly the population of interest in order to be able to study the entire population. For example, Merrill and McLaughlin (1986) selected the 25 largest MSAs for study. A second approach is to select sites randomly, stratifying to ensure that sites with features of interest are included. For example, Mathematica Policy Research, Inc., utilized a mixed random/nonrandom site selection methodology in a project for Health Resources and Services Administration in 1993 examining primary care staffing in HMOs (Felt-Lisk 1996). The analog to studying "exemplary programs" in the Community Tracking Study would be studying only the communities with, say, the most "progressive" or "consolidated" markets. Adopting such an approach exclusively would be inadvisable for this study for two reasons. First, it would not enable us to discuss health system change in the nation outside the "consolidated" markets.

Second, we would risk mistakenly identifying the type of market of greatest interest. For this study, the choice is between judgmental and random selection.

2. Approach

The approach that we recommended and ultimately adopted involves stratifying sites geographically and selecting them randomly with probability in proportion to population. There are separate strata for small MSAs (population of less than 200,000) and for nonmetropolitan areas. This approach provides maximum geographic diversity -- judged critical for the 12 high-intensity sites, in particular -- and acceptable natural variation in city size and degree of market consolidation.

Random selection has a number of important advantages:

1. It protects against researcher bias in site selection -- for example, an inadvertent preference to study sites that are known to the researcher, prominent in the press, experiencing particular types of market changes, or particularly desirable to visit.

2. It also increases the confidence that the communities reflect changes that are representative of the nation.

3. It increases the face validity of the sites studied.

4. It protects against incorrect judgments about which markets are most "interesting." Because of changing health systems and public policy, areas that are interesting today may not be so tomorrow. Similarly, sites that seem uninteresting today may be the most interesting sites tomorrow. Given that tomorrow's interesting sites are not known today, a random draw can help to ensure representation of different types of sites.

5. It ensures that the family sample, which is nested within the sample of sites, is representative of the nation.

6. It protects against the perception that sites have been chosen to support a particular approach to health reform.

We had two concerns about random selection. One was that a chance draw could result in a distribution of sites that was concentrated along some dimension out of proportion to its importance in the population. For example, a draw could by chance include a disproportionate

number of sites from the Northeast. Stratification and systematic sampling were used to reduce the likelihood of this.

A second concern was that even a stratified sample would not include a large enough number of sites of enhanced policy interest. For example, these sites might be experiencing rapid change, exhibiting innovations of interest, or representing the extremes of high and low market consolidation. If the primary objective of the study were to contrast these extremes, metropolitan areas in the extreme categories could have been oversampled.

We considered a number of predetermined stratification criteria and assessed whether to oversample according to any criterion. The following three criteria received the greatest design attention:

1. ***Region.*** Stratification by region and systematic sampling by state ensured the full diversity of health delivery systems across the nation, as well as diversity with respect to historical evolution and community "culture," as reflected by differences across regions.

2. ***City Size***. We desired diversity along the full range of MSA sizes. We created a separate stratum of the smallest MSAs (for the larger sample of sites only) and built implicit stratification by city size into the geographic stratification. Specifically, the New York and Los Angeles CMSAs and the large cities of the Northeast Corridor (excluding New York) were designated as three separate strata for selection of one intensive site each and a controlled number of selections into the larger sample of sites.

3. ***Market Stage.*** Stratification by the stage of market development would ensure that we studied the range of market development that exists -- and would not, by chance, have a disproportionate share of areas that have already undergone change or experienced little change to date. Measures of market stage for use in stratification are limited, with HMO penetration the only feasible proxy measure we have identified. After extensive testing, we dropped HMO penetration as a formal stratification criterion, since the strong geographic concentration of high penetration sites produced natural variation as a consequence of the geographic stratification procedure.

We also considered (and rejected) two dimensions of potential stratification involving data availability -- hospital discharge records data systems and community snapshots. Rather than use availability of discharge records and snapshot data as criteria in selecting

communities, we selected communities randomly, having discharge and snapshot data only for those communities in which they are available.[9]

We considered a number of issues involving potential oversampling of sites by stratification criteria before concluding that oversampling would introduce more risks than benefits. First, in analyses of the case study sites, meaningful adjustment cannot be made for oversampling of high-intensity sites. Furthermore, because case study sites are presented as separate observations, readers will interpret the findings on the basis of an implicit and roughly equal weighting of the sites; the analyst has relatively little influence over this interpretation. A sample that is overloaded in a particular dimension is at risk of being misinterpreted as typical of the nation. Thus, we were reluctant to sample intensive case study sites out of proportion to their population.

Second, there were other drawbacks to the potential criteria for oversampling that involved high- and low-intensity sites:

- ***Region and City Size.*** While we could think of no plausible basis for sampling regions at different rates, we did consider over-representing smaller communities at the expense of larger cities, to ensure that the issues facing such markets were adequately represented. However, we concluded that sampling with probability in proportion-to-size would fulfill this purpose -- approximately 18 percent of the nation's metropolitan population (residing in MSAs of 200,000 or larger) live in MSAs with populations of less than 675,000. Thus, of 48 metropolitan sites ultimately included in the design, proportion-to-size sampling would be expected to yield eight or nine moderate-sized MSAs.[10] Oversampling smaller MSAs would require that we under-represent larger MSAs where most people live; there appeared to be no compelling justification for this strategy, given our overriding focus on impacts on people.

[9] We also considered oversampling sites based on data availability but did not do so because data availability is likely to be associated with other site characteristics.

[10] The final sample included 8 MSAs with populations in the 200,000 to 675,000 range, of which 2 were selected among the 12 intensive sites.

- ***Market Stage.*** We considered two reasons for sampling disproportionately by a proxy for market stage: (1) to ensure a sufficient number of markets that had already undergone change to permit observation of the direction markets may take and to observe implementation of managed care concepts and other innovations; and (2) to ensure that the range of markets by degree of change to date are represented, so we could contrast the extremes (see "Study of the Effects of Change" in Section C.1). Given the inherent deficiencies in available measures of market stage and the strong geographic concentration of markets with high levels of managed care penetration, however, it was unnecessary to *stratify* by a measure of market stage. Furthermore, we were hesitant to *oversample* by such a measure -- for much the same generic reasons that we could not effectively oversample "interesting" sites if we could not predict what sites would be "interesting" in the future.[11]

[11] A deficient classification measure could have been used to some benefit without distorting the representativeness of the sample. However, oversampling by such measures can distort the sample without efficiently serving the purpose of the oversampling. It can also cause technical analytic problems if the criteria for oversampling are correlated with outcomes of policy interest.

III. SAMPLE DESIGN FOR A HOUSEHOLD SURVEY OF HEALTH CARE CONSUMERS

Sample size and design requirements for any survey are ultimately driven by the intended inferences the data will be used to support. Various objectives include describing and analyzing change at the site level, describing and analyzing subgroups of special interest, making cross-site comparisons of communities that represent different models of health care access, and analyzing health care policy at the national level.

In this chapter, we review the sample size considerations related to the planned survey of households, "family insurance units" (families) within households, adults in the sampled families, and a subsample of children. We also discuss which considerations will be extended as appropriate to other surveys conducted for the study.

To clarify the discussion that follows, it may be useful to distinguish among households, families, and individuals in terms of the role they will play as points of sample access and units of analysis. *Households* define the primary point of sample access (within sites), either by telephone or in-person interviews; however, households will not be units of intrinsic interest for analysis purposes. Within each household will be one or more "family insurance units" or *families*, each of which will be interviewed separately. Each interview will collect information about the family unit, *all adults* in the family, and *one randomly selected child* (in families with one or more children). Different analyses will focus on either *families* or *individuals* (adults and/or children), depending on the subject of the analysis. In the sampling discussion that follows, all sample sizes will refer to the number of families, unless specified otherwise. Later in the narrative we address sample sizes and other issues related to analyses of individuals.

It is important to focus on how a series of separate design issues dovetail. For example, separate decisions about (1) precision requirements for describing each site at a point in time, (2) precision requirements for cross-wave and cross-site descriptions and hypothesis tests, (3) subgroup analysis plans, and (4) the number of sites that produce an implied "national" design could produce an inefficient design for analyzing objectives related to national policy.

Gold et al. (1995) reviewed some issues related to creating a national Medicaid access survey from building blocks designed to meet statistical precision criteria for each state. Many of the issues they describe are relevant for a design intended both to provide information about specific sites and to inform policy analyses with nationwide implications. Two conclusions are worth adapting to our circumstances:

1. Descriptive measures of access at the site (or state) level require an "effective" sample of 400 to produce measurements with a 95 percent confidence interval of ± 5 percentage points. However, adjustments required to provide sufficient sample to account for precision losses resulting from "design effects," power requirements for measuring changes in access over time at acceptable precision, subgroup analyses at the site level, comparisons of individual sites with national estimates, and multiplying sample requirements by the number of sites produced aggregate sample requirements that in some configurations reached excessive dimensions. Reducing the sample totals to more manageable levels required making compromises among design objectives, such as relaxing precision standards for subgroup analyses at the individual site level.

2. A design created from individual site building blocks was not particularly efficient for analyzing issues at the national level. In the design adopted here, we combined samples that met fairly stringent precision requirements in the 12 high-intensity sites with smaller samples in a larger number of supplemental sites. This produced a design that can accommodate a broad range of analysis objectives. In addition, we included a modest-size independent national sample *not clustered into sites*, to enhance our ability to make national estimates.

We have divided our design discussion into three categories: (1) design requirements for describing attributes of a single site with a single interview wave, for measuring change over two interview waves, and for making cross-site comparisons; (2) properties of the proposed two-tier sample of sites, combined with the national supplement, for making national

estimates and comparisons; and (3) effective sample sizes available for analyses at the individual (rather than family) level.

A. PRECISION REQUIREMENTS FOR A SINGLE SITE

In the 12 high-intensity sites, we plan to interview approximately 1,225 families in each of two interviewing waves.[12] While a minimum sample of 400 families is sufficient for measuring attributes of a population at a point in time, we arrived at the required sample of 1,225 by considering a sequence of design considerations that increased sample requirements:

- Dealing with minimizing design effects resulting from clustering of multiple families into households and sampling methods for coverage of nontelephone households
- Increasing the sample to permit analyses of subgroups of interest
- Measuring (and testing hypotheses about) change over two interviewing waves
- Making cross-site comparisons

1. Site Descriptions at One Point in Time

A commonly accepted building block for descriptive statistics is a simple random sample (SRS) of 400, which permits attribute descriptions with a 95 percent confidence interval of no greater than $\pm$5 percentage points. If all or a portion of the sample is clustered, or if portions of the sample are over- or under-represented, design effects resulting from clustering and weighting would increase the *nominal* sample required to produce the precision of an *effective* SRS sample of 400:

- A sample of families accessed through an SRS sample of households would have design effects resulting from the presence of multiple "family insurance units" in some households. Based on the RWJF Family Survey on Health Insurance, we estimate that there will be an average of 1.48 family insurance units per household.

[12] In addition, the supplemental national sample (totaling 3,500) is expected to provide 6 to 59 additional sample points to each of the intensive sites, depending on site size (mean = 25 per site).

- The use of a telephone sampling frame would not produce material design effects, since there would be no need to cluster the sample into a small number of subareas of a site.[13] However, such a sample would be *representative of telephone households only*, not of all households. In order to prevent underrepresentation of the uninsured and of disadvantaged populations, the telephone sample will be augmented by an in-person sample of nontelephone households in the 12 high-intensity sites. For cost reasons, the sample of nontelephone households will be geographically clustered, and also under-represented relative to their incidence in the population. Thus, there will be further design effects resulting from clustering and weighting of nontelephone households.

We believe that the combined effects of these factors might produce design effects in the vicinity of 1.25, requiring a nominal sample in the vicinity of 500 to produce an effective sample of 400.[14]

2. Sample Sizes for Subgroup Analyses

If the precision standard discussed earlier were imposed on site subgroups, sample size requirements could escalate rapidly. For example, an effective sample of 2,000 would be required to meet this precision standard for every conceivable 20 percent subgroup; 4,000 would be required for analyzing 10 percent subgroups. One could limit this compounding of sample size by increasing the size of subgroups to be described at the site level, reducing the precision requirements for subgroup description, and/or by oversampling subgroups of special interest. For example:

- An effective sample of 800 (n_{nom} = 1,060 after allowing for design effects) would permit descriptions of 20 percent subsamples (n_{eff} = 160) within a 95 percent confidence interval of $\pm$7.9 percentage points.

- Aggregate sample requirements for subgroup analyses could be reduced if subgroups of special interest were oversampled. However, the range of subgroups of "special" interest was rather diverse for the various analytic uses of the sample. Furthermore, for the subgroup of greatest identified interest -- individuals with chronic medical conditions -- we could find no

[13] Methods of combining frames of listed and unlisted telephone numbers may produce some design effects.

[14] The Gold et al. (1995) estimate of design effects was higher, at 1.325, because of the higher incidence of nontelephone Medicaid households relative to the general population.

efficient way of identifying such individuals with a limited number of screening questions at the beginning of a survey. Thus, our plan calls for no oversampling calls for no oversampling of specific subgroups in the baseline survey.[15]

The following discussion of sample requirements for testing hypotheses of change over time describes how this process will lead to an effective baseline sample of approximately 975 per intensive site. Compared with an effective sample of 800, this number will also increase the sample available for subgroup analyses. With this increase, a 20 percent subgroup could then be described with a sample of n_{eff} = 195, permitting description of attributes with a 95 percent confidence interval of $\pm$7.2 percentage points.

3. Measuring Change Over Multiple Interview Waves

A primary focus of the Community Tracking Study will be tracking change over time and testing hypotheses related to causes of change. Options for measuring change in a site include comparing independent cross-sections, conducting multiple interviews of a longitudinal frame of households, and using a mixed strategy that combines these approaches.[16]

Holding statistical precision constant, larger samples are required for comparing populations than for measuring their attributes separately. For example, independent cross-sections of 800 each would be required to measure changes between two periods at the $\pm$5 percentage-point (95 percent) confidence interval. Samples of 400 would be required for describing attributes of the population separately at the two points in time at the same level of precision.

[15] We are investigating methods of oversampling, in subsequent interview waves, families containing individuals with chronic conditions, based on information obtained in the full baseline interview. In addition, the planned national supplement will permit enhanced subgroups analyses at the national (rather than individual site) level.

[16] Some measures of change may be obtained through including retrospective questions in the survey instrument. In this discussion, we focus on comparing responses to separate surveys not confounded by potential recall bias.

Longitudinal samples can detect change more efficiently, given the plausible presumption that individual responses are correlated over time. If we assume a 50 percent covariance in an attribute of interest over time, a longitudinal sample of 400 could be used to measure changes in attributes within $\pm$5 percentage points. However, reliance on longitudinal samples introduces other risks that compound over time in the event of repeated interviews. For example, a sample that is representative of a site at baseline will, by the time of a follow-up interview, suffer from sample attrition, age by two years compared with the population at large, and fail to capture compositional shifts in the site's population. These problems would intensify even further in the event of repeated interview waves.

We adopted a mixed strategy, under which approximately half the sample would be re-interviewed and the other half replaced by a new random sample draw at the time of the second interview wave. In the event of subsequent interview waves, each cohort would be replaced after being interviewed in two waves.[17] As one might expect, the mixed strategy has an intermediate sample requirement: a mixed sample of 625 per wave would be roughly equivalent in statistical power to independent cross-sections of 800 or to a longitudinal sample of 400.

Two further considerations caused us to recommend a further 56 percent increase in sample size to n_{nom} = 1,225, n_{eff} = 975 per site and per interview wave:

1. First, the effective sample of 625 per wave derived above is adequate to *describe* change within confidence intervals of $\pm$5 percentage points (for aggregate site measures regarding families), not to conduct hypothesis tests that require detecting *true* changes at higher power than that required to report descriptive measures of change.

   -For independent cross-sections, detecting a true change of 5 percentage points at 70 percent power with a 95 percent two-tail test (or 80 percent

[17] If we assume an 80 percent re-interview response rate, for each interview wave, 4/9 of the sample would be re-interviews and 5/9 would be newly drawn. The annual SMS survey of physicians commissioned by the American Medical Association utilizes a similar sample design.

power with a 95 percent one-tail test) would require effective samples of 1,250 per wave rather than the 800 per wave required for descriptive measurement.[18]

- For the mixed longitudinal/cross-section design, the same statistical power would be provided by an effective sample of about 975 per wave. With an expanded sample to compensate for design effects of approximately 25 percent, this calculation produces a required nominal sample per site, per wave of approximately 1,225.

2. The expanded sample enhances our ability to provide at least some subgroup description of changes. With an effective sample of 625 per site, change over time could be measured *only in the aggregate;* measurement of change at the subgroup level would require larger samples. As indicated, a 56 percent increase in sample size would provide an effective sample of 195 per wave for a 20 percent subsample. Subgroup measures of change over time could be described with a 95 percent confidence interval of $\pm$9.0 percentage points.

4. Cross-Site Comparisons

The sample requirements for describing differences in the attributes of two sites are identical to those already cited for comparing *independent* cross-sections for a single site. Effective samples of 800 per site are required for measuring differences and 1,250 per site are required for testing for true differences at 70 percent power. Alternatively, an effective site sample of about 940 could be compared with a larger effective sample of 1,875 (for example, two other sites combined or a national sample) at 70 percent power.

Sample requirements again compound if we wish to test cross-site hypotheses of change rather than point-of-time estimates. If we assume the mixed sampling strategy over time, the above sample sizes would have to be inflated further, to effective samples of either 1,250 or 1,950 per site (per wave) for descriptions of change or hypothesis tests at 70 percent power,

[18] In a sample large enough to measure change with a confidence interval of $\pm$5 percentage points, a *true change* of five percentage points would produce a sample estimate of greater than 5 percentage points about half the time (and less than 5 percentage points the other half). Thus, the true change would be detected about half the time, *or with 50 percent power*. A sample with 70 percent power would be large enough to detect a statistically significant *measured change* approximately 70 percent of the time in the event of a *true change* of 5 percentage points.

respectively, or to "unbalanced" comparisons of effective samples of about 940 and 1,875 (descriptions of change) or 1,460 and 2,925 (hypothesis tests at 70 percent power).

5. Site Sample Sizes

While some of the above calculations suggested even larger sample sizes, we concluded that an effective sample of 975 per site for each interviewing wave, *combined with a mixed longitudinal/cross-section design over time,* was an appropriate sample for each of the 12 high-intensity sites. With allowances for design effects, an effective sample of 975 would be produced by a combined telephone/in-person sample of about 1,225. A sample of this size in each site would permit the following types of analyses:

- Description of attributes at the site level, including description of targeted subgroups

- Tests of hypotheses of change at the site level, at 70 percent power for changes of 5 percentage points, or 80 percent power if the hypothesis being tested specifies a direction of change. For a subgroup representing 20 percent of the population, tests for changes of about 11 percentage points could be obtained at the same power levels.

- Comparisons of attributes between individual sites, detecting differences at 70 percent power between an individual site and either (1) a pair of other sites, or (2) a national sample with an effective size of 1,875 or greater

- Comparisons of *differences of change* between an individual site and either (1) a pair of other sites, or (2) a national sample with an effective size of 1,875, or tests at 70 percent power for detecting *true differences of change* between groups of two or more sites and other groups of sites or the nation

The following types of analyses could *not* be conducted unless even larger samples were utilized:

- Tests of hypotheses of 20 percent subgroup change of less than 11 percentage points at the individual site level

- Tests of change (at 70 percent power) between an individual site and either the nation or groups of other sites

B. NATIONAL ESTIMATES, THE SECOND-TIER SAMPLE OF SITES, AND THE SUPPLEMENTAL NATIONAL SAMPLE

Given the scale and significance of the Community Tracking Study, it would be extremely desirable to track changes in a way that permits statements about the nation, as well as about how individual sites compare with the nation. In the preceding discussion, there was an implicit requirement for an effective national sample in the vicinity of 1,875, or twice the size of the effective sample recommended for each of the 12 sites. Except for issues related to nonmetropolitan representation, this requirement appears to be redundant at first glance. Wouldn't 12 sites with effective samples of 975 each surely produce an effective national sample well in excess of 1,875? *Briefly stated, the answer is "no."*

1. Limitations of a 12-Site Sample for Making National Estimates

We selected our sites as a stratified random sample from a national (metropolitan) frame. As a result, we can assess the ability of a group of site samples to characterize the nation as a whole by viewing sites not in terms of being of direct interest, but as a first stage in a clustered random sample of the nation's households. From this national sampling perspective, a sample of 12 sites is *extremely* inefficient for producing national estimates.

The design effects that reduce the efficiency of national estimates depend on the proportion of total variance attributable to differences in site means rather than to differences among individuals within sites. This proportion (defined as) may be very small for variables such as age or sex, larger for economic-based variables such as employment status, and perhaps larger yet for variables affected by the local health care environment. Past experience suggests that about three percent of the total variance of many economic-based characteristics would be attributed to site differences. Although some health status variables may have $< .03$, we anticipate that many variables of interest relating to health care access would have $> .03$. For measures of HMO participation, for example, we have estimated .10. Even if $\leq .03$ for all

variables of analytic interest, the implications for national sample efficiency are not encouraging:

- With 3 percent of total variance attributable to site differences (= .03), an effective sample of 400 in each of 12 sites (n_{eff} = 4,800) produces an *effective national* sample of only 370.

- An effective sample of 975 in each of 12 sites (n_{eff} = 11,700) produces an *effective national* sample of only 387.

- A sample of 1 million in each of 12 sites (n = 12 million) produces an effective national sample of only 400.

What is the message? A paucity of sites cannot be compensated for by adding raw sample. Only adding sites -- or sample not concentrated in a small number of sites -- can achieve the required national effective sample. Remedying this problem by adding sites at 1,225 nominal observations would be extremely costly. For example, doubling the number of sites to 24 increases the nominal sample from 14,700 to 29,400 per wave, but it only increases the effective national sample from 387 to 774.

Thus, concentrating the sample in a small number of sites sharply constrains our ability to make representative statements about the nation.

2. Overview of the Case for a Three-Tier Sample Design

We considered two approaches to dealing with deficiencies in the 12-site sample for making national estimates:

1. Augmenting the effective national sample with an unclustered telephone sample of the entire nation

2. Expanding the number of sites to 60 -- 48 metropolitan sites inclusive of the 12 high-intensity studied sites plus 12 nonmetropolitan and small metropolitan (population < 200,000) sites -- *but with a smaller sample per site, except for the high-intensity sites*

The first approach could be implemented with a much smaller sample than site-based approaches, since there would be no site-cluster-based design effects -- it is the most efficient

method of expanding the effective size of the national sample. However, two important concerns led us to expand the sample of households clustered into sites:

1. The unclustered national sample would not measure characteristics of the community health care systems facing families, severely limiting the analyses and comparisons that could be conducted with the national sample.

2. As stated in Chapter I, 12 communities are insufficient for providing a fully representative range of communities nationwide, or for providing a benchmark for interpreting the representativeness of the high-intensity sites.

In our final design, we utilized both approaches, in effect creating a three-tier national sample. The second tier is an expanded sample of communities with smaller samples of families with telephones -- the design calls for effective telephone samples of 325 families per site, or nominal samples averaging 375.[19] This tier addresses our concerns about measuring site characteristics and increases the analytic power of the household and other planned data collection efforts by a substantial margin. It does so at considerably less cost than adding large-sample "first-tier" sites.

The third tier is an independent national telephone sample of 3,500 families, of which approximately 1,500 reside in the first- and second-tier sites -- augmenting available sample for site-based analyses. The remaining 2,000 sample members reside outside the selected sites. While the two-tier sample of sites would have been sufficient to provide the national precision requirements already identified, the addition of the third-tier sample substantially increases our ability to test hypotheses of change for regions and subgroups, with a relatively modest increase in total sample.

The design specifies that nontelephone households are not sampled at low-intensity sites or in the national supplement. We concluded that metropolitan nontelephone households

[19] The assumed design effect of 1.15 for the second-tier sites is smaller than the 1.25 assumed for high-intensity sites, but it relates only to the ability to characterize families *with telephones*, not all families.

could be adequately represented by the nontelephone sample in the 12 large-sample sites.[20] As a result, we rejected, on cost grounds, extending the nontelephone sample to represent nonmetropolitan and small metropolitan areas.

3. The Second-Tier Sample of Sites

The expanded sample of low-intensity sites is intended both to provide an enhanced national sample and to permit measurement of site characteristics with adequate precision for a range of analytic purposes. After summarizing the effects of adding the second tier on effective sample sizes for national estimates, we shall return to our rationale for the choice of sample size for the second-tier sites.

We estimate that including a second tier of low-intensity sites will have the following effects on effective national sample sizes:

- Adding 36 second-tier metropolitan sites with effective samples of 325 each (n_{nom} 375 per site) to the first-tier (nominal) sample of 1,225 family interviews in each of 12 high-intensity sites increases the total nominal metropolitan sample by about 92 percent, from 14,700 to 28,200. In contrast, the *effective national metropolitan sample* increases by 700 percent, from less than 400 (387) to about 3,100. This total exceeds, by a substantial safety margin, the requirement stated earlier for an effective national sample of 1,875.

- An additional 4,500 family interviews in 12 nonmetropolitan (and small metropolitan) areas rounds out a fully representative national sample, with a slightly smaller effective size of approximately 2,800.[21]

[20] The nontelephone sample will total approximately 576, or an average of 6 clusters of 8 interviews each, in each of the 12 high-intensity sites. With the 12 high-intensity sites being representative of all metropolitan areas with populations in excess of 200,000, we estimate the effective sample of metropolitan nontelephone households to be approximately 210. While this is too small a sample for separate analyses of nontelephone families, such families represent only 5 percent or less of all metropolitan families. Thus, the design effects on estimates pertaining to *all* metropolitan families -- resulting from clustering and underrepresentation of nontelephone households -- is well within acceptable bounds.

[21] The 12 sites include 9 nonmetropolitan and 3 nonmetropolitan sites. Because these segments of the population are somewhat underrepresented in the site sample -- 20 percent of the sites covering about 26 percent of the population -- the design effects resulting from weighting the sample to national proportions produce a somewhat smaller effective national sample than the effective metropolitan sample.

- Inclusive of all sites (but before adding the national supplement), a total of 32,700 family interviews (per wave) would be conducted in a nationally representative sample of 60 sites.[22]

Two separate factors account for the increase in effective sample sizes when the number of sites is increased:

1. The effect of increasing the number of sites from 12 to 60 (with the proposed sample size and site allocation configurations) increases the effective sample about four-and-a-half-fold, to approximately 1,725.

2. The 48 metropolitan sites represent about 15 percent of the universe of eligible MSAs and PMSAs. However, they represent more than 45 percent of the nation's metropolitan population because of the conventional procedure of selecting sites with probability proportional to size. This 45 percent coverage factor produces a *finite population correction*, which reduces design effects resulting from the clustering of the sample into sites, further increasing the effective sample to more than 2,800.

The Health System Change design group engaged in extensive debate about the required sample size for the second tier of low-intensity sites. We initially considered effective family samples of only 100 per site, which would have provided site-specific measurement of family attributes at a fairly large 95 percent confidence interval of $\pm$10 percentage points, normally considered too large for site-level descriptions. However, from the perspective of providing site descriptors for cross-site regression analyses, the relevant factor is how highly correlated sample-based site measures are with their true values. Thus, the appropriate question becomes, *"What proportion of the variance in a measured site characteristic is due to true variations rather than statistical error?"* For characteristics that vary substantially across the selected sites, we concluded that samples of only 100 families per site would have been sufficient to construct measures for which more than 90 percent of the variance is due to variations in the "true" values of site characteristics. We reached this conclusion after testing two such measures for some typical samples of 48 metropolitan sites:

[22] Increasing the number of sites to 60 with full samples of 1,225 per site would have increased effective national sample sizes only slightly, from 2,800 to 3,200, but would have involved an aggregate nominal sample of

1. We compared the variation in HMO penetration rates across MSAs to the sampling error associated from measuring HMO participation from consumer surveys of 100 per site[23]. We obtained a correlation between actual and sample-derived HMO participation in excess of 95 percent (r = .95, r^2 = .91).

2. We made a similar comparison of actual and sample-derived measures of the proportion of families who are black or Hispanic. It produced correlations (r) in excess of 97 percent.

In general, we have observed that about 10 percent of the variance in many measures of health care options -- HMO participation or health care options offered by employers -- is associated with the site rather than with the individual family or employer (for example, 1, for as previously defined in the design effect discussion). For such measures, the proportion of the variance (r^2) associated with the true value of the measure will exceed 90 percent for effective samples of 90 or more per site; for effective samples as small as 60 per site, r^2 will equal 85 percent.

Based on these examples, we concluded that measures of site characteristics derived from samples of 100 per site, or even smaller, would provide acceptable analytic variables for many quantitative purposes.

However, from the perspective of providing adequate *descriptive* measures for site groupings of interest, the design team was not comfortable with the degree to which sites would have to be aggregated with effective samples of only 100 families per site:

- Based on the statistical criteria already established, the following groupings would have been required for descriptions and hypothesis tests of various types:

  -Groups of 4 sites for descriptions of attributes

  -Groups of 6 sites for descriptions of *change over time*

  -Groups of 10 sites for *hypothesis tests* of change over time

---

73,500 family interviews, a further increase of 125 percent from the proposed sample allocation.

[23] We are focusing here on sample size issues, not on whether consumers can accurately report the characteristics of their health care plans

- In contrast, expansion of the effective family samples to 325 per low-intensity site provided the following more powerful capabilities:
  - Description of *individual* site attributes at slightly less than the established precision standard (confidence intervals of $\pm$5.5 percentage points rather than 5.0 percentage points)
  - Descriptions of change over time for *any pair* of sites
  - Hypothesis tests of change over time for *any three* sites

These groupings relate to descriptions and tests for the *specific sites* under consideration, *not* as representatives of a broader stratum of similar sites nationally. For example, groupings of 12 sites *chosen to reflect a meaningful stratum of sites* -- for example, the Northeast or the quintile of sites with the highest degree of managed care penetration -- would be sufficient to provide basic descriptions of the attributes of the stratum in question, with effective samples of approximately 550. While this capability is far superior to what could have been provided with the first-tier 12 site sample -- which would include only two to four sites in any site category of interest -- it provided only a limited capability to describe and measure change for national subgroups represented by various groups of sites. This consideration, plus the absence of a viable strategy for oversampling population subgroups, such as low-income people or those in poor health, provided the motivation for expanding the effective size of the national sample -- and of meaningful subgroups -- by adding an independent national sample as a third-tier supplement to the sample design.

4. The Third-Tier Supplemental National Sample

The third tier of the sample design is *an independent, geographically unclustered national telephone sample of 3,500 families*.[24] The addition of the national supplement has the following dramatic effects on estimated effective national sample sizes:

- A relatively modest (10.7 percent) increase in the total sample -- from 32,700 to 36,200 -- increases the effective sizes of (1) the national metropolitan sample by 89 percent, from 3,150 to 5,960; and (2) the national total sample by 144 percent, from 2,830 to 6,900.

- The incremental sample of 3,500 actually increases the effective national sample by 4,070, because its distribution compensates for several deficiencies in the site-based sample from the perspective of making national estimates:

  -The eight largest metropolitan sites eligible for sample selection include 18.2 percent of the U.S. population. They were selected with certainty into the sample of 60 sites (8/60 = 13.3 percent), but their site samples were not increased to their full national share of the population. Their full representation in the national supplement partially alleviates this situation.[25]

  -The national supplement partially compensates for underrepresentation of nonmetropolitan and small metropolitan sites in the 60-site sample.

  -The designation of strata for geographic stratification of the site sample led to modest deviations from strict PPS sampling of sites. Again, the national supplement partially alleviated this feature of the sample design.

- In the absence of the national supplement, the effective sample for a 20 percent subgroup of the population ranged from 550 (for a subgroup defined by 12 of the 60 sites) to 1,750 (for a subgroup spread evenly across all 60 sites).[26] Inclusive of the national supplement, the corresponding range of effective samples for a 20 percent subgroup is 1,380 to 2,500.

C. SAMPLE SIZES FOR ANALYSES AT THE INDIVIDUAL LEVEL

[24] Because of the clustering of family insurance units into households, we estimate the effective size of the supplemental national sample of families to be 3,390.

[25] In the absence of a national supplemental sample, we would have recommended increasing the samples in the larger sites to proportions closer to their national population shares.

[26] For subgroups spread across all sites, the effective size of a 20 percent sample is more than *60 percent* of the effective total sample. This seemingly counterintuitive result is due to a dramatic reduction in design effects with the reduction in the size of site "clusters" associated with subsampling.

The sample size requirements derived here focused on family insurance units. However, in the design for the Community Tracking Study, relevant units of observation include *households*, *families*, and *individuals*:

- Households are the initial unit for sampling purposes, with each household broken into family insurance units.
- Each family unit is interviewed separately, with basic demographic information collected for all individuals.
- Detailed information is collected for all adults and for one randomly selected child per family unit (for families with children).

1. Nominal Sample Sizes

Given the sample of families, samples of individuals for some analyses will be considerably larger. Based on the results of the RWJF Family Survey, we expect to observe an average of 1.4 adults per family insurance unit. We also expect that 38 percent of all family units will have one or more children. Furthermore, we assume that approximately 30 percent of all adults will be defined as heavy users of the health care system for purposes of separate analyses. These assumptions imply the following nominal sample sizes of adults and children resulting from the design for the household survey:

- 1,715 adults in each of the high-intensity sites, 455 adults in each of the remaining sites, and 4,900 adults in the national supplement, for a total of 47,320 adults
- Averages of 514.5 and 136.5 high-use adults per site (for the two site categories) and 1,470 in the national supplement, for a total high-use sample of 14,196
- Averages of 465.5 and 142.5 children per site and 1,330 in the national supplement, for a total sample of 13,756 children under 18 years of age

2. Effective Sample Sizes

In order to estimate effective sample sizes of individuals, we must augment the design effects assumed thus far with (1) design effects resulting from the clustering of more than one adult in some households, and (2) design effects resulting from weighting, given that the

selection of one child per family implies lower probabilities of selection of children from families with more than one child.

Consider, for example, a sample of 1,715 adults residing in the 1,225 family units in 1 of the 12 intensively studied sites. Design effects already identified for families would reduce the effective sample of adults to 1,365 (1,715 × 975/1,225). Because of similarity in insurance coverage by family unit, within-family design effects would be quite severe for evaluating health insurance issues. On the other hand, further design effects might be quite low for analyses of individual health characteristics and use, which have relatively low correlations within families. Depending on the variable being analyzed, we expect the range of effective sample sizes to vary widely -- from perhaps 1,075 for studying insurance issues to about 1,330 for measuring health characteristics. The corresponding range of effective samples for a low-intensity site would be approximately 360 to 445; for the national supplement, 3,750 to 4,625; and for the full sample, 7,360 to 8,500.

The effective sample of adults is similar to the number of families for analyses of health insurance, making the precision of estimates roughly comparable to what has been presented for families. Precision is somewhat higher for analyses of individual characteristics, with the effective samples about 35 percent larger than the number of families for site-based analyses, and about 23 percent larger for the national sample.

The samples of high-use adults would have only minor additional design effects, since the proportion of families with multiple high users would be far smaller than the proportion with more than one adult. We estimate that the nominal sample of 515 high users per high-intensity site would yield an effective sample of 410, sufficient for description of attributes at the site level. For testing hypotheses of change over time (70 percent power, two-tail test), the minimum detectable change would be 7.7 percentage points for one site, or 5.5 and 4.5

percentage points for groups of two and three sites, respectively. Inclusive of effective samples of 135 per low-intensity site and 1,425 in the national supplement, the effective sample of high users for making national estimates would be about 3,980, sufficient even for detecting national measures of change in the attributes of meaningful subgroups of high users.

Since the probability of selection for children is inversely proportional to the number of children per family, the sample would have to be weighted for estimates involving children. This process would produce significant design effects, which we estimate to be in the vicinity of 1.17, because of weighting. Inclusive of this factor, the effective sample for children would be about 317 per high-intensity site, sufficient for site descriptions with a confidence interval of $\pm$5.6 percentage points. Grouped analyses of three sites could detect changes of $\pm$5.1 percentage points. With effective samples of 105 children per low-intensity site and 1,100 in the national supplement, the effective national sample of children would be 3,400, again sufficient for meaningful analyses of subgroups of children.

D SUMMARY OF NOMINAL AND EFFECTIVE SAMPLE SIZES FOR THE HOUSHOLD SURVEY

Table III.1 summarizes our estimates of the nominal and effective sample sizes for the household survey of health care consumers:

TABLE III.1
SUMMARY OF NOMINAL AND EFFECTIVE SAMPLE SIZES, HOUSEHOLD SURVEY OF HEALTH CARE CONSUMERS

	High-Intensity Sites (12)	Low-Intensity Sites (48)	Independent National Sample	Combined National Sample
Families				
Total				
Nominal Sample	1,225	375	3,500	36,200
Effective Sample[a]	975 to 1,020	325 to 430	3,390	6,900
20 PercentSubgroup				
Nominal Sample	245	75	700	7,240
Effective Sample[b]	195 to 205	65 to 85	680	1,380 to 2,500
Individuals				
Total				
Nominal Sample	1,715	455	4,900	47,320
Effective Sample[c]	1,075 to 1,330	360 to 445	3,750 to 4,625	7,360 to 8,500
30 Percent Subgroup				
Nominal Sample	514.5	136.5	1,470	14,195
Effective Sample	410	135	1,425	4,040
Children				
Total				
Nominal Sample	465.5	142.5	1,330	13,755
Effective Sample	317	105	1,100	3,400

[a]Site-specific effective samples include portion of national sample falling within the sites; effective samples increase with site size.
[b]Site-specific effective samples increase with site size. For the national sample, lower estimate is for subgroup concentrated in 12 sites; upper estimate is for subgroups spread evenly across all sites.
[c]The lower estimate is for insurance-related variables; upper estimate is for typical individual attributes.

IV. SAMPLE DESIGN FOR THE PHYSICIAN SURVEY

In addition to the household survey, the planned data collection strategy for the Community Tracking Study includes surveys of physicians, employers, and a range of health care organizations. These organizations comprise physician groups, hospitals, insurers, and managed care plans. All the planned surveys share two design elements:

1. A common set of 60 randomly selected sites as the first stage of sample selection
2. Larger samples in the 12 high-intensity sites, and smaller samples in the remaining 48 sites

In this section, we extend the design principles established for the household survey to the physician survey, and then summarize the design for this survey.

A. OVERVIEW

The physician survey is very similar in design to the household survey, including the same three-tier and mixed longitudinal/cross-sectional design structures. However, the precision requirements -- and therefore recommended sample sizes -- are smaller.[27] The physician survey has target nominal samples of 450 physicians per high-intensity site, 125 per low-intensity site, and a supplemental national sample of 1,200, for a total sample of 12,600 per interviewing wave.

Two primary design issues must be addressed in specifying the structure of the physician sample:

1. We are particularly interested in having sufficiently large samples for conducting clinical vignettes with primary care physicians and for describing their attributes at the

[27] Given the Community Tracking Study's primary focus -- impacts on people -- the most stringent precision standards were established for the household survey.

site level, as well as for describing patient care physicians in the aggregate (inclusive of specialties). However, primary care physicians constitute only one-third of all physicians -- oversampling sufficiently to provide an adequate sample of primary care physicians will *reduce* the effective sample for describing attributes of physicians in general, because of the design effects resulting from weighting.

2. Particularly in small and medium-sized sites, a significant fraction of the physicians will be interviewed in high-intensity sites, producing finite population corrections that will substantially increase effective sample sizes. Thus, unlike the household survey, effective samples are *greater* than nominal samples for physicians.

After accounting for both oversampling and finite population corrections, the design will provide effective samples of 400 primary care physicians per high-intensity site -- sufficient for descriptions at the site level, and for detecting true changes of 5.5 percentage points (at 70 percent power) for two-site groupings. An effective sample of a total of 430 physicians will be more than sufficient for description at the site level and will permit detection of 5.3 percentage point changes in attributes for two-site groupings. Next, we address the oversampling issue and the effects of finite population corrections on the design structure and on effective sample sizes. We conclude with a discussion of the third-tier independent national sample, summarizing effective sample sizes for the physician survey.

B. OVERSAMPLING OF PRIMARY CARE PHYSICIANS AND FINITE SAMPLE CORRECTIONS

We would like to describe attributes of both primary care physicians and patient care physicians in the aggregate *at the site level*. These objectives are in potential conflict, since the oversampling of primary care physicians will reduce effective aggregate samples for any given total nominal sample. Table IV.1 illustrates the impact on effective total sample sizes of increasing the relative sampling rate of primary care physicians, within a total sample of 450 physicians in a single high-intensity site, assuming no effects of finite sample corrections.

TABLE IV.1
IMPACT OF OVERSAMPLING PRIMARY PHYSICIANS ON PHYSICIAN EFFECTIVE SAMPLE SIZES
(No Finite Sample Correction)

	Nominal Samples			Effective Sample[a]
Primary Care Sampling Rate	Primary Care Physicians	Non-Primary-Care Physicians	Total Physicians	Total Physicians
1.0	150	300	450	450
2.0	225	225	450	405
2.5	250	200	450	375
3.0	270	180	450	347
4.0	300	150	450	300

[a]The effective total sample $n_{eff} = 1/[P^2/n_{pc} + (1-P)^2/n_{npc}]$, where P equals the population proportion of primary care, and n_{pc} and n_{npc} are the sizes of the primary and non-primary-care samples.

1. Finite Population Corrections

Fortunately, accounting for finite population corrections increases the effective sample available for analyzing primary care physicians. It also alleviates the compromising effects of oversampling on effective total sample sizes. An important factor in producing this result is that finite population corrections are larger for primary care physicians because of their higher sampling rate, permitting the subsample target to be reached with a lower oversampling rate than would be required without finite population corrections.

Table IV.2 summarizes the impact of site size on physician effective samples for the size distribution of 12 high-intensity metropolitan sites selected for the Community Tracking Study.

After preliminary investigations of some alternative sample sizes and oversampling rates, the design for the 12 high-intensity sites was produced in the following manner:

TABLE IV.2
IMPACT OF SITE SIZE ON PHYSICIAN EFFECTIVE SAMPLES
Average Primary Care Sampling Rate = 2.66
(Distribution of 12 Selected Metropolitan Sites)

MSA Population	Number Of Sites	Physician Population			Physician Sample			Effective Sample[a]		
		Primary Care	Non-Primary-Care	Total	Primary Care	Non-Primary-Care	Total	Primary Care	Non-Primary-Care	Total
< 1 million	4	425	645	1,070	205	145	350	400	187	430
1 million to 2 million	2	1,285	2,020	3,305	300	175	475	400	192	4.30
2 million to 2.5 million	5	1,775	2,575	4,350	325	170	495	400	184	430
> 4 million	1	3,535	6,290	9,825	360	200	560	400	205	430
Total	12	16,500	25,785	42,465	3,011	2,389	5,400			
Average/Site		1,390	2,150	3,540	285	165	450	400	190	430

[a] The effective sample for site I, $n_{i,eff} = n_{i,nom} \div (1 - f_i)$, where f_i equals proportion of the population sampled; cell effective samples are means of separate calculations for each site in cell. Effective total samples also reflect design effects resulting from weighting. Columns may not sum to total because of rounding.

- For each site, the sample of primary physicians was set high enough to provide an effective sample of 400 for that site.

  -For example, in a site with a population of 640,000 and 425 primary care physicians, a sample of 206 (48.5 percent of the total) will provide an effective sample of 400. (Note that 206 ÷ (1-.485) is approximately equal to 400.)

  -In a site with a population of 2,250,000 and 1,775 primary care physicians, a sample of 326 (18.4 percent of the total) will provide an effective sample of 400.

  -Note that smaller sites have a *higher proportion* but a *smaller absolute number* of primary care physicians than larger sites. Thus, smaller sites "release" sample to permit larger nominal samples for the larger sites, which do not benefit as much from finite population corrections.

- The effective sample of non-primary-care physicians was set to equalize the effective *total* sample of physicians across sites, with an aggregate nominal sample of 5,400 (averaging 450 per site) for the high-intensity sites. This produced an average effective non-primary-care sample of 190 per site, and an effective total sample of 430 per site.

The resulting sample has an average of 285 primary care and 165 non-primary-care physicians per site, with primary care physicians being oversampled by a factor of 2.66, on average. However, as total sample size increases with site size to compensate for the weakening

of finite population corrections, the relative sampling rate increases as well -- from about 2.15 for the smaller sites to 3.2 for the largest site.

A similar sample construction exercise was conducted for the low-intensity sites, based on an average sample of 125 physicians per site. Because of the lower sampling rates associated with the smaller samples, finite population corrections were correspondingly smaller. The resulting sample achieved an effective sample per site of 100 for the primary care sample, an average of 50 for the non-primary-care sample, and 114 for the total sample. These effective samples are achieved with average nominal samples of 81 primary care and 44 non-primary-care physicians per site, with an average relative sampling ratio for primary care physicians of 2.82.

These sample sizes are lower than those available for constructing quantitative low-intensity site measures in the household survey discussed previously. However, based on the analysis presented in the discussion of the household design, we believe that the correlation of measured site physician characteristics with their "true" values will still be acceptable -- with r^2 = .91 for measures associated with primary care physicians (for variables having = .1), and r^2 = .92 for measures associated with all physicians.

Before adding a supplemental national sample of physicians, we estimated that the combined 60-site physician sample will provide effective national samples in the vicinity of 2,100 for primary care physicians, 1,535 for non-primary-care physicians, and 2,220 for all physicians.

C. INDEPENDENT NATIONAL SAMPLE AND SUMMARY SAMPLE SIZES

Like the design for the household survey, the third tier of the physician sample design is an *independent, geographically unclustered national sample of 1,200 physicians*.[28] In partial compensation for the underrepresentation of non-primary-care physicians in the site samples, primary and non-primary-care physicians will be proportionately represented in the independent sample. The addition of the national supplement has the following effects on estimated effective national sample sizes:

- A 6.5 percent increase in the nominal primary care sample -- from 7,308 to 7,783 -- increases the effective size of the national sample by 27 percent, from 3,100 to 2,660.

- For the non-primary-care sample, a 17.7 percent increase in the nominal sample -- from 4,092 to 4,817 -- increases the effective size of the national sample by 51 percent, from 1,535 to 2,320.

- The total nominal physician sample increases by 10.5 percent, from 11,400 to 12,600. This expansion increases the effective national sample of physicians by 63 percent, from 2,200 to 3,580.

Table IV.3 summarizes our estimates of the nominal and effective sample sizes for the physician survey.

[28] Because of the relatively high sampling rates of physicians in some of the 60 sites, some physicians in the independent sample will be duplicates of those already selected for the site samples. We estimate that approximately 30 physicians -- 2.5 percent of the independent sample or 0.26 percent of the site sample -- will be selected twice. In these rare cases, the physician will be interviewed once and double-weighted in the analysis when appropriate.

TABLE IV.3
SUMMARY OF NOMINAL AND EFFECTIVE SAMPLE SIZES, SURVEY OF PRIMARY CARE AND NON-PRIMARY-CARE PHYSICIANS

	High-Intensity Sites (12)	Low-Intensity Sites (48)	Independent National Sample	Combined National Sample
All Physicians				
Nominal sample	450	125	1,200	12,600
Effective sample[a]	430	114	1,200	3,580
Primary Care Physicians				
Nominal sample[b]	285	81	475	7,783
Effective sample	400	100	475	2,660
Non-Primary-Care Physicians				
Nominal sample[b]	165	44	725	4,817
Effective sample[b]	190	50	725	2,320

[a]Includes effects of both weighting and finite sample corrections.

[b]Site averages

V. THE EMPLOYER SURVEY

The employer survey has two primary objectives: (1) to track change in premiums and insurance offerings over time, and (2) to describe the role of employers in the market. From the perspective of the Community Tracking Study, the employer survey would be based on design principles similar to those established for the household and physician surveys. However, the employer health insurance survey will be designed by the RAND Corporation to serve a range of research objectives related to state policy initiatives, as well as those of the Community Tracking Study. Thus, the discussion in this chapter is intended to provide input into the RAND design.

Before turning to a discussion of sample sizes, we should specify two presumed design features of the employer survey:

1. Since our primary interest is in characteristics of health insurance plans *available to employees*, our interest in firms is proportional to their employment levels. We recommend sampling firms with probability in proportion to size, rather than sampling firms such as AT&T and the local delicatessen with equal probability.

   -This selection procedure would produce finite population corrections -- as it did in the selection of sites with probability proportional to size -- that increase effective sample sizes, because the sampled firms would represent a significant proportion of each high-intensity site's employment, especially in the small and medium-sized MSAs.

2. A mixed longitudinal/cross-sectional sample structure similar to the one specified for the household survey would be adopted for the employer survey as well, reducing sample sizes required for measuring change over time.

Our preliminary design called for a nominal employer sample of 450 per high-intensity site and 75 per low-intensity site, for a total sample of 9,000 per wave. For evaluation of sample precision, we consider the measurement of changes in insurance offerings and average premium levels. In Chapter III we stated that a mixed longitudinal/cross-section design would permit site-level descriptions of changes in attributes -- such as whether an employer offers an

HMO option -- to be measured with an effective sample of 625 employers per wave, for changes of ±5 percentage points at a 95 percent confidence interval. For a continuous variable such as monthly insurance premiums, sample size requirements depend on the coefficient of variation of the measure in question. Furthermore, the requirement would apply to the subsample of employers that offer health insurance:

- For employers offering health insurance, the 1993 RWJF Employer Health Insurance Survey reported within-state coefficients of variation for family premiums ranging from .34 to .45, averaging about .40 (Cantor et al. 1995). With a coefficient of variation of .40, change in average premiums could be measured at the site level at a 95 percent confidence interval of ±5 percent with an effective sample of 384 employers that offer insurance.

Based on reported data from the 1993 survey, we estimate that 57 percent of employers offer health insurance, or 77 percent on a size-weighted basis (Cantor et al. 1995, p. 203 and p. 210). Thus, for a sample of employees chosen with probability proportional to size, an effective sample of 500 employers per site would yield a sufficient number of employers offering insurance.

Based on these estimates in combination with cost considerations, we recommended an effective sample of 600 employers in each of the 12 high-intensity sites. This number is slightly below the target for measuring changes in employer attributes, but more than enough for measuring changes in average premiums among employers offering insurance.

Unlike the household survey, where required nominal samples exceeded effective sample sizes, we estimate that nominal samples averaging about 450 per site would be sufficient to achieve effective samples of 600 employers. Two factors are responsible for this conclusion:

1. Within-site design effects that reduced effective sample sizes in the household survey -- clustering of families within households and supplementation of a telephone sample with clustered samples of nontelephone households -- are not applicable to the employer survey.

2. Selecting employers with probability of selection in proportion to size results in finite sample corrections because of the significant fraction of employees covered by the sample, especially in the small and medium-sized sites:

   - In potentially selected MSAs with populations under 675,000 (average population, 306,000; average employment, 144,000), a sample of 327 employers, including all firms with 250 or more employees and almost 40 percent of employers with 50 to 249 employees, would be sufficient to provide an effective sample of 600.

   - Similarly, effective samples of 600 could be provided by nominal samples of 462 in medium-sized MSAs (population 675,000 to 2 million), and by samples of 500 to 560 in large MSAs and component PMSAs of large CMSAs.

The average samples of 450 per high-intensity site would be supplemented by samples of 75 employers in each of the 48 low-intensity sites, producing effective samples of approximately 80 per site. Samples of this size would permit construction of measures of average employer health plan characteristics that, with sampling error, would have an 89 percent correlation (r^2) with "true" measures (that is, without sampling error) of these characteristics. We estimate that the combined nominal sample of 9,000 employers would produce an effective *national* sample of approximately 2,000 employers, before the addition of a national supplement or design modifications to serve other research objectives.

REFERENCES

G. Anderson, R. Heyssel, and R. Dickler. "Competition Versus Regulation: Its Effect on Hospitals." *Health Affairs*, vol. 12, no. 1, 1993, pp. 70-80.

Campbell, E.S., and G.M. Fournier. "Certificate of Need Deregulation and Indigent Hospital Care." *Journal of Health Politics, Policy and Law*, vol. 18, no. 4, 1993, pp. 905-925.

Cantor, Joel C., et al. "Private Employment-Based Health Insurance in Ten States." *Health Affairs*, summer 1995, Exhibit 3, p. 203, and Appendix Exhibit 1, p. 210.

Cromwell, J. "Impact of State Hospital Rate Setting on Capital Formation." *Health Care Financing Review*, vol. 8, no. 3, 1987, pp. 69-82.

Felt-Lisk, Suzanne. "How HMOs Structure Primary Care Delivery." *Managed Care Quarterly*, vol. 4, no. 4, 1996, pp. 96-105.

Gold, Marsha, Jack Hadley, Donna Eisenhower, Charles Metcalf, Lyle Nelson, Karyen Chu, Richard Strouse, and David Colby. "Design and Feasibility of a National Medicaid Access Survey with State-Specific Estimates." *Medical Care Research Review*, September 1995, pp. 405-429.

Langa, K.M., and E.J. Sussman. "The Effect of Cost Containment Policies on Rates of Coronary Revascularization in California." *New England Journal of Medicine*, vol. 329, no. 24, 1993, pp. 1784-1789.

Luft, H.S., S.C. Maerki, and J.B. Trauner. "The Competitive Effects of Health Maintenance Organizations: Another Look at Evidence from Hawaii, Rochester and Minneapolis/St. Paul." *Journal of Health Politics, Policy and Law*, vol. 10, no. 4, 1986, pp. 625-658.

Melnick, G.A., and J. Zwanziger. "Hospital Behavior Under Competition and Cost Containment Policies, The California Experience, 1980-1985." *Journal of the American Medical Association*, vol. 260, no. 18, 1988, pp. 2669-2675.

Merrill, J., and C. McLaughlin. "Competition Versus Regulation: Some Empirical Evidence." *Journal of Health Politics, Policy and Law*, vol. 10, no. 4, 1986, pp. 613-623.

Robinson, J.C., and H.S. Luft. "Competition, Regulation and Hospital Costs, 1972-1982." *Journal of the American Medical Association*, vol. 260, no. 18, 1988, pp. 2676-2781.

Robinson, J.C., and C.S. Phibbs. "An Evaluation of Medicaid Selective Contracting in California." *Journal of Health Economics*, vol. 7, 1989, pp. 437-455.

Shortell, S.M., and E.F.X. Hughes. "The Effects of Regulation, Competition and Ownership on Mortality Rates Among Hospital Inpatients." *New England Journal of Medicine*, vol. 318, no. 17, 1988, pp. 1100-1107.