EVALUATION OF ALTERNATE DESIGNS FOR A FUTURE NMCUES

Brenda G. Cox and Ralph E. Folsom, Research Triangle Institute

The rapidly rising cost of medical services in the United States in recent years, together with a continuous effort to improve the quality, effectiveness, and availability of health care, has led to a continuing need for comprehensive data for individuals and families on health status, patterns of health care utilization, charges for services received, and payers of amount paid. Sponsored by the Health Care Financing Administration (HCFA) and the National Center for Health Statistics (NCHS), the National Medical Care Utilization and Expenditure Survey (NMCUES) was the second of a series of national medical care surveys planned to provide data on a regular basis. These surveys will permit in-depth statistical descriptions of the utilization of health care services and the associated costs for various population segments, including the nation as a whole. They will also provide valuable data for the evaluation of current public programs such as Medicare and Medicaid, for the assessment of inequity in access to the health care delivery system and other unmet needs, and for the comparison of alternative solutions to health policy issues. The findings from these studies will ultimately have an impact on public policy concerning health care for the entire nation.

Current planning for population based surveys conducted by the National Center for Health Statistics (NCHS) assumes that the data systems can be integrated to save on data collection costs, to reduce respondent burden, and to increase the utility of the resultant data. As a part of a larger NCHS effort to evaluate the advantages of an integrated data system, alternative designs were examined for integrating the National Medical Care Utilization and Expenditure Survey (NMCUES) with the larger National Health Interview Survey (NHIS). The NHIS is a continuing, cross-sectional survey of the health status and experiences of the civilian, non-institutionalized population of the United States. This paper summarizes the results of the demonstration of the possibilities when the two surveys are linked (Cox et al, 1983).

Organizationally, this paper is divided into several sections. The first section provides an overview of the characteristics of the various designs that were developed in the study. Next, the unlinked design for the NMCUES is described which selects the NMCUES sample independently of the NHIS using an area frame. After that, optimally-allocated linked designs for the NMCUES are described which use the NHIS sample as the frame for the NMCUES. The paper concludes by comparing the linked and unlinked designs and making recommendations for future research.

1. OVERVIEW

Except when oversampling of certain population subgroups is specified, the allocation of the sample to strata is typically proportional to stratum size. However, when data collection costs and variances differ among strata, optimal allocation of the sample should be considered since it may result in substantial cost savings. Use of the NHIS sample as a frame for the NMCUES implies that information will be available for stratifixation and optimal allocation of the households selected for the NMCUES.

For a multi-purpose survey like NMCUES with many outcome measures of interest and numerous reporting domains, the preferred optimization strategy minimizes total survey cost subject to multiple variance constraints. Separate variance constraints are set to control the precision of key survey statistics for the total population and for important reporting domains. The NMCUES optimization was obtained using a survey design optimization approach developed by Chromy and described in Folsom, Williams, and Chromy (1980). Chromy's optimization algorithm is an iterative approach that provides an optimal solution when the convergence criteria are met.

As a baseline for comparison, specifications were developed for an unlinked NMCUES design. Selected independently of the NHIS, this unlinked design would result in a stratified, clustered area sample similar to that of the NMCUES conducted in 1980, hereafter referred to as the "1980 NMCUES." For the sake of flexibility for NCHS planning, two sample sizes were used: 6,000 versus 10,000 responding households. The 6,000 household design is roughly equivalent in size to the 1980 NMCUES. The 10,000 household design was added so that NCHS could evaluate the improved precision for smaller domains versus the increased survey cost when the larger sample was used.

Linkage of the NMCUES to the NHIS implies that the NHIS sample selections can be used to form the frame for the NMCUES. Unlike the area frame used for the last NMCUES, the NHIS-based frame will contain names, addresses, and individual and family level characteristics. This information could be used in sample selection and data collection to reduce costs and improve data quality. The elements of the NHIS-based frame could be either the NHIS sample dwelling units or the NHIS sample households. For simplicity of presentation, this paper will be restricted to the linked household design in which the frame is composed of NHIS sample households.

The linked household design would select a subsample of the NHIS sample households for inclusion in the NMCUES. The individuals within the subsampled households would be interviewed in Round 1 regardless of whether or not they lived in the clustered NHIS sample dwelling units. Rounds 2 through 5 data collection would follow the same rules as the unlinked design. Two sample sizes were again used in developing sample designs; these sample sizes were determined as the sizes required to yield the same precision as the unlinked design with 6,000 and 10,000 responding households. Optimal allocation was used to determine the number of PSUs and segments from the NHIS to include in the sample.

These two designs are self-weighting; that is, all sample individuals are selected with the same probability. In many ways, this eliminates the chief advantage of linkage with the NHIS. With knowledge of individual characteristics available for NHIS sample respondents, added precision can be obtained for small domains without proportionally increasing the size of the total NMCUES sample. To evaluate this feature of NHIS linkage, a third and final design type was investigated. This design is an optimally-allocated, linked-household design in which precision constraints were set for the total population and the Medicaid population based upon those achieved by the unlinked design.

An important result of this investigation is that there appears to be little relative gain from linkage when the final design has to be self-weighting. This result is reasonable when one considers that the principal gain from the linked self-weighting design is in the elimination of costs associated with counting and listing. Since the NMCUES mode of interview across rounds was not altered in this investigation (the first two rounds and the last round use personal interviews and the other two rounds use telephone interviews), there was little gain derived from the knowledge of names, addresses and telephone numbers for NHIS sample individuals. The pay-off came with the optimally-allocated design that used characteristics of the NHIS respondents to oversample heavy users of health care services and to increase the precision for small domains without proportionally increasing the size of the total sample.

2. THE UNLINKED NMCUES DESIGN

The unlinked NMCUES design that was studied in this investigation was patterned after the design used for the 1980 NMCUES. Specifically, an area sampling approach was assumed which would result in a self-weighting design in the sense that each sample individual would be selected with equal probability. The sample sizes required to yield 6,000 and 10,000 responding households had to be determined as well as the survey costs associated with these designs. The variances achieved by the unweighted, unlinked NMCUES design were then modeled for use in sample size determination for the remaining designs.

2.1 Definition of the Unlinked Design

The unlinked sample design would be a stratified, multi-stage area probability design in which each sample dwelling unit is selected with equal probability. The first stage sample would consist of primary sampling units (PSUs) which are counties, parts of counties, or groups of contiguous counties. The second stage sample would consist of secondary sampling units (SSUs) which are Census enumeration districts (EDs) or block groups (BGs). Smaller area segments would constitute the third stage. All of the dwelling units within the sampled segments would be listed. During the fourth stage of sampling, dwelling units within each sample segment would be designated for inclusion in the NMCUES sample. All civilian, noninstitutionalized individuals residing in the sampled dwelling units in the initial round of data collection would be included in the survey. To facilitate family-level analyses, single college students in the 17 to 22 age range would be linked to their parents' residence and included in the survey only if their parents' residence was selected.

2.2 Sample Size Determination

Two sets of sample sizes were required for the unlinked NMCUES design: a sample size sufficient to yield 6,000 responding households and a sample size sufficient to yield 10,000 responding households. In order to obtain these sizes, a precise definition was needed for what constituted a "responding household" since households change over the data collection year. The decision was made to use responding Round 1 households and to describe the sample sizes needed to yield 6,000 responding Round 1 households and 10,000 responding households. These Round 1 households are the Round 1 reporting units (RUs) after college student RUs are linked back to their parent RUs. The unit for which a questionnaire booklet is completed, a reporting unit is composed of individuals related by blood, marriage, or adoption who live in the same dwelling unit. Since data collection costs are for reporting units (RUs) and rounds, sample sizes had to be developed in terms of these units.

The first step in this process was to model the 1980 NMCUES experience. The modeling began with the set of control system records generated by responding Round 1 households. (In the 1980 NMCUES, a Round 1 household was defined to be responding if it was linked to an RU that completed an interview in any of the five data collection rounds.) The 1980 NMCUES contained 6,269 responding Round 1 households. These responding Round 1 households generated 6,603 completed RU interviews in Round 1; 6,519 completed RU interviews in Round 2; 6,528 completed RU interviews in Round 3; 4,559 completed RU interviews in Round 4; and 6,561 completed RU interviews in Round 5. There tended to be more RU interviews than there were responding Round 1 households since households containing college students required more than one RU assignment to handle the different addresses for data collection. These interviews occurred over 108 unique PSUs and 809 segments.

2.3 Variance Modeling for the Unlinked Design

As a baseline for comparison of the unlinked with the linked designs, the decision was made to fix the precision of the linked designs for selected key statistics and key domains to that of the unlinked design, and then to compare the designs with respect to sample sizes and costs. The domains of interest for this demonstration were the total population and Medicaid recipients. The statistics of interest were as follows:

- average number of hospital visits.

- average number of facility visits,
- average number of office visits,
- average annual expenditure for hospital visits,
- average annual expenditure for facility visits,
- average annual expenditure for office visits,
- average annual out-of-pocket (OOP) expense for hospital visits,
- average annual out-of-pocket expense for facility visits,
- average annual out-of-pocket expense for office visits, and
- the proportion with large out-of-pocket expenditures.

In order to determine the sample sizes required for the linked designs, the variance had to be modeled for the unlinked, self-weighting design.

The NMCUES estimation approach is to construct means in terms of total 1980 person-years rather than in terms of all persons ever existing in 1980. For domain k, the mean utilization or expenditure per person-year was estimated as:

$$\bar{Y}_k(\text{NMCUES}) = \sum_{i\varepsilon S} W(i)\ \delta_k(i)\ Y(i) \Big/ \sum_{i\varepsilon S} W(i)\ T(i)\ \delta_k(i) \qquad (1)$$

where

$\delta_k(i)$ is 1 if the i-th person belongs to the k-th domain and 0 otherwise,

$Y(i)$ is the response of the i-th person,

$W(i)$ is the analysis weight for the i-th person, and

$T(i)$ is the time-adjustment factor for the i-th person.

The numerator estimates total expenditures or utilization and the denominator the average annual number of persons in the population (i.e., the total person-years). The time adjustment factor, T(i), is the total days in the year that person i is eligible for NMCUES divided by the total days in the year.

Large out-of-pocket expenditures were defined as "annualized" out-of-pocket expenditures of $200.00 or more. The annualized out of pocket expenditure is the annual out-of-pocket expenditure divided by the fraction of 1980 that the person was eligible. For domain k, the proportion with large out of pocket expenditures was estimated as:

$$\bar{Y}_k(\text{NMCUES}) = \sum_{i\varepsilon S} W(i)\ T(i)\ \delta_k(i)\ Y(i) \Big/ \sum_{i\varepsilon S} W(i)\ T(i)\ \delta_k(i) \qquad (2)$$

where Y(i) is 1 if the person had large out-of-pocket expenditures and 0 otherwise.

The variance of $\bar{Y}_k$(NMCUES) was derived assuming a three-stage household survey design patterned after the 1980 NMCUES sample design, with SMSA/County-sized PSUs and area segments (SEGs) selected as noncompact clusters of dwelling units. Using this approach, the variance of $\bar{Y}_k$(NMCUES) may be modeled as:

$$\text{Var}[\bar{Y}_k(\text{NMCUES})] = \sigma_k^2(\text{PSU})/r + \sigma_k^2(\text{SEG})/r\bar{s} + \sigma_k^2(\text{HH})/r\bar{s}\bar{t} \qquad (3)$$

where

r is the number of PSUs,

$\bar{s}$ is the average number of segments per PSU,

$\bar{t}$ is the average number of responding households per segment,

$\sigma_k^2(\text{PSU})$ is the between-PSU, within-stratum variance component for domain k,

$\sigma_k^2(\text{SEG})$ is the between-segment (SEG), within-PSU variance component for domain k, and

$\sigma_k^2(\text{HH})$ is the between-household, within-segment variance component for domain k.

A variance components estimation program (Shah, 1979) was applied to the 1980 NMCUES data to produce the generalized composite components for PSUs, segments, and households. Table 1 presents the proportion of NMCUES expenditures and utilization variation explained by each component for the various types of service statistics for the total population and Medicaid recipients.

These three-stage variance component estimates were used to estimate the variances that would be achieved by self-weighting NMCUES designs with 6,000 and 10,000 responding households. The terms remaining to be specified in the variance expression presented in equation (3) are the number of PSUs (r), the average number of segments sampled per PSU ($\bar{s}$), and the average number of households sampled per segment ($\bar{t}$). For modeling purposes, the RTI general purpose sample was assumed for the next NMCUES, which contains 102 PSUs (r = 102). A future NMCUES should experience no worse than the nonresponse and attrition encountered by the 1980 NMCUES. Therefore, the 1980 NMCUES experience was ratio adjusted to produce the sample sizes required for the 6,000 and 10,000 household designs. Since the 1980 NMCUES had been designed to be optimal with respect to the number of selections per segment, the number of responding households per segment was set to the value that the 1980 NMCUES achieved or $\bar{t} = 8$. The total number of segments in the 6,000 responding OBRU design would then be 750 ($rs = 750$) and 1,250 for the 10,000 responding household design ($rs = 1,250$).

These estimated variances were used as precision criteria for the other designs that were investigated in this study. Table 2 presents the results of this variance modeling activity for the two domains of interest and the ten outcome measures. For convenience, percent relative standard errors are presented rather than variances. The percent relative standard error is the standard error (the square root of the variance) divided by the parameter being estimated, expressed as a percentage. The percent relative standard errors achieved by the 6,000 household design are sufficient for the estimates based upon the total domain but the increased precision that the 10,000 household design achieves for Medicaid estimates would be desirable.

2.4 Cost Modeling

Since cost comparisons would be needed between the unlinked designs and the linked designs, a systematic method was needed to generate the costs for all designs. The approach that was used in this study was to develop unit costs by task for each design. The NMCUES tasks that were included in the modeling were the basic sampling, weighting, data collection, and data processing tasks. The unit costs that were developed for each task were fixed costs, PSU-level costs, segment-level costs, and RU-level costs.

The first step in the process was to document what the actual cost experience had been for the 1980 NMCUES. An early decision was made to include only direct costs in the modeling. Indirect costs are the mechanisms whereby contractors recover costs that cannot be directly charged to a project, such as the costs for administration and building maintenance. Since these indirect costs vary across contractors as well as the accounting procedures used to recover these costs, direct costs only were modeled. For the 6,000 household design, direct costs would be $4,963,013. For the 10,000 household design, the direct costs would be $7,209,409.

3. OPTIMALLY-ALLOCATED LINKED DESIGNS

With knowledge of the characteristics of NHIS respondents, there are possibilities for gains due to stratification and to optimally allocating the sample. To investigate these possibilities, five optimally-allocated linked household designs were investigated. The first two designs were optimally-allocated self-weighting designs, one with the precision of the 6,000 household unlinked design and the other with the precision of the 10,000 household design. Next, the self-weighting requirement was dropped and two optimally-allocated designs were developed, one using the 6,000 household constraints and the second using the 10,000 household constraints. Since the main reason to increase the sample size to 10,000 households was to obtain improved precision for the smaller domains such as Medicaid recipients, the decision was made to investigate one last design in which the precision constraints for the total population were set to those achieved by the 6,000 household design, and the precision constraints for the Medicaid subpopulation were set to those achieved by the 10,000 household design.

3.1 The Optimization Problem

When NHIS households are used as the sampling units for the NMCUES, there is much useful information about the households from the NHIS interview. Most of this information is individual-level such as age, race, sex, relationship to head, limitation of activity, bed disability days, perceived health status, medical conditions, education level, marital status, and employment. Since NMCUES samples entire households to facilitate family-level analysis, these data would have to be aggregated to the household level in order to be used for stratification purposes.

Stratification of the NHIS sample prior to selecting the NMCUES sample could serve the dual purpose of providing control over the distribution of the sample while increasing the precision of survey estimates. The variance of survey estimates is reduced and hence the precision increased through stratified sampling when the strata are formed to maximize the between-stratum variation and minimize the within-stratum variation. Variables to use for stratification are those that result in homogeneity of the units within strata and associated heterogeneity between the stratum means.

Time constraints prevented the examination of 1980 NMCUES data to determine what variables might best be used for the stratification of the NHIS sample prior to NMCUES sample selection. Rather, variables that were generally known to be good predictors of health care utilization and expenditures were used for stratification. Specifically, black/nonblack, aged/ nonaged, poor/nonpoor, and self-perceived health status (healthy/non-healthy) variables were used for stratification in this demonstration. Sample size limitations of the 1980 NMCUES data base used to estimate variance components required collapsing the black strata over the poor/ nonpoor variable resulting in eight nonblack strata and four black strata.

To illustrate the advantages of an optimal-allocation approach, five optimal designs were developed. The domains that were included in the optimizations were the total population and Medicaid recipients. For use in stratification, household-level variables were defined that denoted the race (black versus nonblack), poverty status (above or below 150 percent of the value defined as poverty), aged status (containing no person 65 or over versus at least one), and health status (containing no person with poor or fair health versus at least one). The optimization was performed for nine utilization and expenditure means and the subpopulation proportion burdened with large out-of-pocket expenses.

3.2 Variance Modeling for the Stratified Design

The first step in developing optimally-allocated designs was modeling the variance for a stratified household sample drawn from the first phase NHIS sample households. To describe the variances for the linked household design, the characteristics of the NHIS sample need to be described, since the NHIS would be used as the NMCUES frame. The redesigned NHIS will have the same target population as the NMCUES. To represent this target population, the NHIS will include 200 sample PSUs and 8,750 segments from these PSUs. The segments will contain 40 addresses on the average with 6 addresses selected for inclusion in the NHIS. The sample segments will be partitioned into 52 sets which will be allocated to weeks so that each weekly sample is a valid national sample. An added feature of the NHIS is that blacks will be oversampled at a rate of 1.4 times that of nonblacks. Finer details of the structure of the redesigned NHIS had not been developed at the time this study was conducted.

Using a stratified sampling approach to subsampling NHIS sample households, NMCUES would estimate the mean for domain k as:

$$\bar{Y}_k(\text{NMCUES}) = \sum_{h=1}^{H} \hat{\pi}_k(h)\ \bar{Y}_k(h) \qquad (4)$$

where

$\bar{Y}_k(h)$ is the NMCUES estimated mean for stratum h,

$\hat{\pi}_k(h)$ is the NHIS-estimated fraction of the k-th subpopulation total person-years associated with the h-th stratum, and

H is the number of sample strata.

For the nine utilization and expenditure measures, the stratum mean is estimated as

$$\bar{Y}_k(h) = \sum_{i \epsilon h} W(i)\ \delta_k(i)\ Y(i) \Big/ \sum_{i \epsilon h} W(i)\ \delta_k(i)\ T(i) \qquad (5)$$

where

$Y(i)$ is the response of the i-th person,

$W(i)$ is the analysis weight of the i-th person,

$\delta_k(i)$ is one if the i-th person belongs to the k-th domain and zero otherwise, and

$T(i)$ is the fraction of the year that the i-th person was eligible for NMCUES.

For the proportion burdened with large out-of-pocket expenses, the stratum mean is estimated as

$$\bar{Y}_k(h) = \sum_{i \epsilon h} W(i)\ \delta_k(i)\ T(i)\ Y(i) \Big/ \sum_{i \epsilon h} W(i)\ \delta_k(i)\ T(i) \qquad (6)$$

where Y(i) is one if the annualized out-of-pocket expenses are large (> $200) and zero otherwise.

To simplify modeling the variance, it was assumed that NHIS oversampling of blacks would be at the last stage and that black/ nonblack would be a stratification variable. Under this assumption, the variance of the stratified estimate can be modeled as:

$$\begin{aligned}
&\text{Var}[\bar{Y}_k(\text{NMCUES})] \\
&\quad = \text{Var}_{\text{NHIS}}\ [E\ (\bar{Y}_k|\text{NHIS})] \\
&\quad + E_{\text{NHIS}}\ [\text{Var}\ (\bar{Y}_k|\text{NHIS})] \\
&\quad = \text{Var}_{\text{NHIS}}\ [\bar{Y}_k(\text{NHIS})] \\
&\quad + E_{\text{NHIS}}\ \{\sum_{h=1}^{H} \pi_k^2(h)\ S_k^2(h)[1-f(h)]/m(h)\} \\
&\quad \triangleq D_w(k)\ [\sigma_k^2(\text{PSU})/r + \sigma_k^2(\text{SEG})/r\bar{s} \\
&\quad + \sigma_k^2(\text{HH})/r\bar{s}\bar{t}] \\
&+ \sum_{h=1}^{H} \pi_k^2(h)\ S_k^2(h)\ [1-f(h)]\ /\ E[m(h)] \qquad (7)
\end{aligned}$$

where

$f(h)$ is the NMCUES subsampling rate for stratum h or m(h)/n(h),

$m(h)$ is the NMCUES stratum h household sample size,

$n(h)$ is the NHIS stratum h household sample size,

$D_w(k)$ is the design effect for NHIS unequal weighting for the k-th domain,

$\sigma_k^2(\text{PSU})$ is the between NHIS PSU variance component for domain k,

$\sigma_k^2(\text{SEG})$ is the between NHIS segment, within NHIS PSU, variance component for domain k,

$\sigma_k^2(\text{HH})$ is the between NHIS household, within NHIS segment, variance component for domain k, and

$S_k^2(h)$ is the stratum h variance for domain k.

Again, the variance components computed from the 1980 NMCUES were used to estimate the NHIS components. A Taylor Series approximation for the simple random sampling variance of a combined ratio estimator was used to estimate $S_k^2(h)$.

The expected NMCUES sample size from the h-th stratum can be expressed as

$$E[m(h)] = r\bar{s}\bar{t}\ f(h)\ \pi'(h) \tag{8}$$

where $\pi'(h)$ is the expected fraction of the NHIS sample from the h-th stratum or

$$\pi'(h) = M(h)\ o(h) \ / \sum_{h=1}^{H} M(h)\ o(h) \tag{9}$$

and $M(h)$ is the population count of households in stratum h. The $o(h)$ term represents the NHIS oversampling of black strata; that is $o(h) = 1.0$ for nonblack strata and $o(h) = 1.4$ for black strata.

Presuming that black/nonblack is used as a stratification variable with equal probability sampling within strata, the design effect for unequal weighting in domain k estimation may be modeled as

$$D_w(k) = \pi_B^2 \ / \ \theta_B + \pi_{NB}^2 \ / \ \theta_{NB} \tag{10}$$

where π_B and π_{NB} are the proportion of blacks and the proportion of nonblacks in the population and θ_B and θ_{NB} are the proportion of blacks and the proportion of nonblacks in the NHIS sample. Note that

$$\theta_B = 1.4\ \pi_B \ / \ (1.4\ \pi_B + \pi_{NB}) \tag{11}$$

and

$$\theta_{NB} = \pi_{NB} \ / \ (1.4\ \pi_B + \pi_{NB}) \tag{12}$$

since the NHIS will oversample blacks by 1.4 times the rate at which they occur in the population. Hence, $D_w(k)$ may also be expressed as

$$D_w(k) = 1 + (0.16\ \pi_B\ \pi_{NB} \ / \ 1.4). \tag{13}$$

For convenience sake, relative variance components were used in the optimization. To model the relative variances, note that

$$RV_k(NMCUES) = Var[\bar{Y}_k(NMCUES)] \ / \ \bar{Y}_k^2(NMCUES). \tag{14}$$

For domain k, the relative variance of a mean estimated using the linked household design can be expressed as:

$$RV_k(NMCUES) = \sum_{\ell=1}^{H} RV_k(\ell) \ / \ m(\ell) + \sum_{\ell=H+1}^{H+2} RV_k(\ell) \ / \ m(\ell)$$

where $\ell = 1,2,\ldots,H$ are the second phase strata used in selecting the NMCUES subsample and H+1 and H+2 are the first phase segment and PSU sampling stages.

3.3 Cost Modeling

The next stage in developing optimally-allocated designs was modeling the cost components associated with each second-phase NMCUES stratum and each stage of the first-phase NHIS design. Let $C(\ell)$ represent the variable unit cost for a selection from level ℓ. Then the optimization problem may be stated as follows:

$$\text{Minimize } CV(NMCUES) = \sum_{\ell=1}^{H+2} m(\ell)\ C(\ell) \tag{16}$$

subject to

1. $\sum_{\ell=1}^{H+2} RV_k(\ell) \ / \ m(\ell) \le RV_k^*$ for $k=1,2,\ldots,K$
2. $m(\ell) > 0$ for $\ell=1,2,\ldots,H+2$
3. $200 \le m(H+2) \le m(H+1)$
4. $m(\ell) \le m(H+1)$ for $\ell=1,2,\ldots,H$.

CV(NMCUES) is the total variable cost for NMCUES and RV_k^* is the relative variance constraint established for the k-th domain.

The variable costs for the PSU stage of sampling [C(H+2)] and the segment level of sampling [C(H+1)] were obtained by modifying the task-level unit costs produced as a part of the cost modeling of the unlinked household design. The unit costs for the subsampled households within NHIS-defined strata vary depending upon the response rate and movement rates within the strata. The 1980 NMCUES experience was used to estimate the roundwise rates at which ineligibles, nonrespondents, and movers would be encountered and to develop the household-level cost component for each of the 12 strata. Unit costs were developed for movers, tracing, interviewing ineligibles, and interviewing outside and inside the clusters and used in forming the overall unit costs for each stratum.

3.4 Self-Weighting Optimally-Allocated Designs

The first type of design that was investigated was a stratified, self-weighting linked household design. Using this design, the variance would be expressed as in equation (7) where $f(h) = f/o(h)$. The factor f is the overall subsampling rate desired for the NMCUES subsample of the NHIS after NHIS oversampling is removed. The Chromy optimization procedure was used to obtain optimum values for the number of PSUs, the average number of segments to sample per PSU, and the NMCUES subsampling rate to be used within the sample segments (r, s, and f).

The optimization was performed twice. When the variance constraints associated with the 6,000 household unlinked design were used, the optimal solution was 102 PSUs; 1,258 segments; and 5,980 responding households. With a subsampling rate f of 83 percent, black strata would be subsampled at a 59 percent rate (f/1.4) and nonblack strata at the 83 percent rate. When this design is used, the total cost for the design is $4,844,013 as compared to $4,963,013 for the unlinked design with the same precision.

When the variance constraints associated with the 10,000 household unlinked design were used, the optimal stratified linked household design had 103 PSUs; 2,117 segments; 9,960 responding households; and a subsampling rate f of 82 percent. Allowing for the NHIS oversampling implies that black strata would be subsampled at a 58 percent rate and nonblack strata at the 82 percent rate. When this design is used, the total cost is $6,931,233 as compared to $7,209,409 for the unlinked design with equivalent precision.

3.5 Nonself-Weighting Optimally-Allocated Designs

The next set of designs that was investigated was the stratified linked household designs without the self-weighting requirement. The advantage of this type of design is that heavy utilizers of health care services can be identified and oversampled. Since optimization occurs over PSUs (r), segments ($r\bar{s}$), and NMCUES strata (h=1,2,...,H), the stratified linked sample has H+2 design levels.

The first design that was investigated was an optimally-allocated design with the precision constraints of the unlinked 6,000 household design for the total and Medicaid domains. The optimal solution used 98 PSUs; 1,152 segments; and 5,880 responding households with subsampling rates ranging from 57 to 100 percent. In general, the unhealthy and nonblacks were sampled at a higher rate. The fact that greater percentages of NHIS nonblacks were selected than blacks is the result of the fact that blacks occur at a rate 1.4 greater than nonblacks in the NHIS sample. The total cost for this design is $4,770,353 as compared to $4,963,013 for the unlinked 6,000 household design and $4,844,013 for the equivalent self-weighting optimally-allocated design.

The next design that was investigated was an optimally-allocated design with the precision of the 10,000 household unlinked design for the total and Medicaid domains. The optimal solution used 106 PSUs; 1,811 segments; and 9,717 responding households with subsampling rates ranging from 59 to 100 percent. The total cost for the design was $6,758,063 as compared to $7,209,409 for the 10,000 household unlinked design and $6,931,233 for the self-weighting equivalent optimally-allocated design.

For household samples like the 1980 NMCUES that are drawn from area frames, there is little information available for use in sample stratification, and what information is available is for geographical areas rather than households or dwelling units. To obtain the required sample sizes for small domains, a larger than otherwise needed sample size is frequently used. With household-level stratification information, these small domains can be oversampled without having to correspondingly increase the size of the total sample.

To illustrate this advantage, an optimally-allocated design was created where the precision of the 10,000 household design was specified for the Medicaid domain but the precision of the 6,000 household design was deemed satisfactory for total population estimates. These constraints result in an optimal design with 95 PSUs; 2,092 segments; and 7,228 responding households with NMCUES subsampling rates ranging from 32 to 100 percent. The total cost for the (6,000/ 10,000) design was $5,601,533 which compares quite favorably with the $6,758,063 costs for the comparable nonself-weighting design with 10,000 household constraints for both the total and Medicaid domain statistics.

4. COMPARISON OF THE LINKED AND UNLINKED DESIGNS

Three types of sample designs have been described in this paper, including two unlinked designs, two optimally-allocated self-weighting linked household designs, and three optimally-allocated nonself-weighting linked household designs. Table 3 summarizes the sample sizes and costs for the designs investigated for potential use in future NMCUES. The optimally-allocated designs contrast quite favorably with the unlinked designs in terms of cost.

Table 3 also gives the months that the NHIS sample would have to be aggregated to obtain the required number of sample segments from the specified number of PSUs. These estimates of aggregation time were based upon the assumptions that NHIS would include 8,750 segments and 200 PSUs for an average of 43.75 segments per PSU in a year's time, and that the NMCUES would be selected from the 90 percent that were conducted by personal interview. The aggregation times range from 1.5 to 6.7 months with the longer periods of aggregation found for the optimally-allocated designs. The modeling of movement was only approximate so that the costs associated with movement may be understated, particularly for designs that aggregate over a longer period of time. More attention could be given to cost modeling of movement as the time between the NHIS and NMCUES increases.

Linkage of the NMCUES to the NHIS has the unique advantage of knowing the names, addresses, and personal characteristics of sample households in advance of data collection. The design with the most potential for exploiting this knowledge is the stratified nonself-weighting optimally-allocated design. Research could be conducted to produce such a design for the next NMCUES. This research would determine the domains and statistics of interest to the survey and the most appropriate set to include in the optimization. The gain from the use of an optimally-allocated design should far exceed the costs of developing such a design.

5. CONCLUDING REMARKS

The NMCUES has many small analysis domains of interest including the Medicaid population, the Medicare population, the aged, the poor, and blacks. The need for these domain analyses has led NMCUES in the past to use large self-weighting samples to obtain adequate precision for these small domains. This approach resulted in greater precision than was needed for large domains such as the nonaged or white domains. Without linkage, however, this is the best approach possible since household characteristics are not available for use in sampling.

With linkage to the NHIS, there is a plethora of information about the households that can be used to create an optimally-allocated design with increased precision for selected domains. This design strategy could be pursued much further than was possible in this demonstration. Precision constraints could be set for a larger group of policy relevant domains. The stability of the variance components would also need to be considered and the accuracy of the cost components. Finally, the effect of the length of aggregation of the NHIS sample could be built into the cost modeling.

By relaxing the self-weighting condition, an optimally-allocated design can be created that obtains the desired precision for a small domain by oversampling from strata where domain members are concentrated. To repeat the example cited earlier, if the required variance constraints for the Medicaid domain are those achieved by the 10,000 household unlinked design, then the self-weighting stratified linked design that would be used is the one that achieves the variance constraints of the 10,000 household unlinked design for all domains. If the variance constraints achieved by the 6,000 household unlinked design were acceptable for the total population, the nonself-weighting stratified optimally-allocated linked design can achieve both sets of variance constraints by oversampling strata with a high concentration of Medicaid recipients. The survey costs with the nonself-weighting approach would be $5,601,533 as compared to $6,931,233 with the self-weighting design.

In constructing an optimal design, careful attention needs to be given to the reporting domains to be included in the optimization. The survey planner is assured of acceptable levels of precision for those statistics and domains that are included in the optimization. The precision for other statistics and other domains will depend upon the extent to which they are related to the statistics and domains included in the optimization.

The disadvantage of the optimally-allocated nonself-weighting approach is associated with estimation for domains and/or statistics not included in the optimization. The nonself-weighting 6,000/ 10,000 design produces estimates of the desired precision for the total utilization and total expenditures statistics by oversampling from the unhealthy strata. If total income is being estimated instead, this design may or may not yield estimates of the desired precision since the design did not control for the precision of income estimates. Alternatively, if total utilization or total expenditures are being estimated for a domain not included in the optimization, such as the college educated domain for instance, then again the design may or may not yield estimates of the desired precision. The precision of estimates for domains and/or statistics not included in the optimization will depend upon the extent to which the statistics and/or domains are related to the statistics and domains included in the optimization.

In practice, most surveys have multiple domains that are of interest and a diversity of statistics to report. This does not imply that a nonself-weighting optimally-allocated design is unacceptable. In this situation, an appealing strategy is to consider several estimates simultaneously where the estimates are chosen by classifying their variance properties and selecting a typical variance model from each class. Similarly, the domains to include in the optimization can be chosen by listing the important domains of interest and selecting domains that represent diverse groups of the population. Care should be taken that the extremes of the domains of interest are represented in the set of domains subject to optimization, since the extreme groups are usually the rarest and hence an adequate sample size will not be obtained unless special steps are taken. Thus, a survey particularly interested in contrasting health expenditures for different income groups would want to include the poor and the wealthy as domains in the optimization. With a large proportion of the population middle income, there may be no need to explicitly include them as a domain, particularly if the total population is included as a domain in the optimization.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the assistance of Thomas Virag and Barbara Moser in developing the cost models used in this study and of Wendell Refior in developing variance models and the optimal allocations. The research in this paper was done for the National Center for Health Statistics (NCHS) and the Health Care Financing Administration (HCFA) by the Research Triangle Institute under Contract No. HRA-233-79-2032. The views expressed in this paper are those of the authors and no official endorsement by NCHS and HCFA is intended or should be inferred.

7. REFERENCES

Cox, Brenda G., Ralph E. Folsom, Thomas G. Virag, and Wendell F. Refior (1983). Design Alternatives for Integrating the NMCUES with the NHIS. Series 2 Report, National Center for Health Statistics, Hyattsville, Maryland. (In press).

Folsom, Ralph E. Jr., Rick L. Williams, and James R. Chromy (1980). Optimum Design of a Medical Care Expenditure and Utilization Survey Involving a Provider Record Check. Report No. RTI/1725/01-06S, Contract No. 233-78-2102, National Center for Health Statistics, Hyattsville, Maryland.

Shah, B. V. (1979). VMCPNLS: Program to Compute Variance Components. In-House Report, Research Triangle Institute, Research Triangle Park, North Carolina.

Table 1. Percent of NMCUES Expenditure and Utilization Variation by Domain and Type of Service

Domain	Statistic	Percent of Variation		
		PSU	Segment	Household
Total	Hospital Visits	0.61	0.07	99.32
	Facility Visits	1.34	5.17	93.49
	Office Visits	0.66	2.02	97.32
	Hospital Charges	0.02	0.28	99.70
	Facility Charges	0.59	3.38	96.03
	Office Charges	0.03	3.28	96.69
	Hospital OOP Expenses	0.02	0.65	99.33
	Facility OOP Expenses	0.48	0.92	98.60
	Office OOP Expenses	0.02	6.31	93.67
	Proportion With Large OOP Expenses	0.02	5.93	94.05
Medicaid	Hospital Visits	0.07	0.73	99.20
	Facility Visits	0.41	3.60	95.99
	Office Visits	0.49	0.56	98.95
	Hospital Charges	0.02	0.83	99.15
	Facility Charges	0.03	1.53	98.44
	Office Charges	0.50	0.02	99.48
	Hospital OOP Expenses	0.19	0.03	99.78
	Facility OOP Expenses	0.03	0.03	99.94
	Office OOP Expenses	0.02	0.20	99.78
	Proportion With Large OOP Expenses	0.25	2.06	97.69

Table 2. Estimated Means and Relative Standard Errors for the Unlinked NMCUES Design with 6,000 and 10,000 Responding Households

Domain	Statistic	Y_k(NMCUES)	Relative Standard Errors	
			6,000 Households	10,000 Households
Total	Hospital Visits	0.18	3.11	2.61
	Facility Visits	0.86	4.92	4.25
	Office Visits	4.18	2.02	1.69
	Hospital Charges	362.04	6.22	4.84
	Facility Charges	50.56	4.95	4.11
	Office Charges	117.71	2.42	1.88
	Hospital OOP Expenses	33.10	12.08	9.39
	Facility OOP Expenses	9.77	4.82	3.99
	Office OOP Expenses	53.70	2.43	1.89
	Proportion With Large OOP Expenses	0.24	7.03	5.47
Medicaid	Hospital Visits	0.33	6.63	5.20
	Facility Visits	1.36	7.70	6.27
	Office Visits	5.21	5.59	4.63
	Hospital Charges	691.56	13.56	10.55
	Facility Charges	78.09	7.45	5.80
	Office Charges	139.60	7.27	6.04
	Hospital OOP Expenses	36.18	29.97	23.98
	Facility OOP Expenses	7.39	20.80	16.19
	Office OOP Expenses	23.10	9.57	7.44
	Proportion With Large OOP Expenses	0.11	22.79	18.32

Table 3. Sample Size Summary for the Alternate NMCUES Design

Design Type	Sample Sizes			Aggregation Time	Direct Costs
	PSUs	Segments	Households		
Unlinked Designs:					
6,000 Households	102	750	6,000	N/A	4,963,013
10,000 Households	102	1,250	10,000	N/A	7,209,409
Linked Optimally-Allocated Designs:					
S.W. 6,000/6,000	102	1,258	5,980	3.8	4,844,013
S.W. 10,000/10,000	103	2,117	9,960	6.3	6,931,233
N.S.W. 6,000/6,000	98	1,152	5,880	3.6	4,770,353
N.S.W. 10,000/10,000	106	1,811	9,717	5.2	6,758,063
N.S.W. 6,000/10,000	95	2,092	7,228	6.7	5,601,533