A METHODOLOGICAL STUDY OF FACTORS ASSOCIATED WITH WHETHER CHILDREN RECEIVE ADEQUATE MEDICAL CARE

by

Mary Grace Kovar

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1428

December 1982

A METHODOLOGICAL STUDY OF FACTORS ASSOCIATED WITH WHETHER CHILDREN RECEIVE ADEQUATE MEDICAL CARE

by

Mary Grace Kovar

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the School of Public Health

Chapel Hill, 1982

Approved by:

H. Bradley Wells
Advisor

Elizabeth J. Coulter
Reader

Gary G. Koch
Reader

Ivan Cornoni Huntley
Reader

Bernard G. Greenberg
Reader

MARY GRACE KOVAR. A Methodological Study of Factors Associated with Whether Children Receive Adequate Medical Care (UNDER THE DIRECTION OF H. BRADLEY WELLS)

The Federal government and private concerns sponser many national surveys based on samples with complex designs. The designs involve clustering to reduce data-collection costs and weighting to produce national estimates. Because of the complex samples designs, traditional methods of analysis that require the assmuption of simple random sampling are not appropriate for analyzing data from these surveys.

One such survey, the National Health Interview Survey, is designed to collect information on the health and use of medical care by the civilian noninstitutionalized population of the United States. This work is an investigation of the feasibility of applying one analytic approach to answer a question of substantive interest using data from that survey.

The question is; "What are the factors associated with children receiving medical care at an interval short enough to allow them to receive the recommended preventive care?" The analytic approach includes screening all of the possible variables for inclusion in a model, estimating the variance-covariance matrix using a Taylor-series approximation, and using that matrix and the Grizzle-Starmer-Koch linearized model to analyze the data.

The approach is useful and the results seem reasonable. Children whose mothers and fathers have received care within a year, children whose families have telephones, children in small families and in families where the head has more than the average education are more likely to have received care than other children. The results differ from those of some other studies because the use of a screening procedure to select the variables for inclusion in a model resulted in selecting some variables that have not been included in previous research and excluding others that have usually been included. Factors such as income and race, which are often viewed as being of great importance in univariate analyses, are not of great importance in a multivariate analysis. Their association with the probability of a child's receiving medical care during the mid-1970's was due to their correlation with other factors, such as education and family size.

The inclusion of the full matrix from the complex sample design resulted in smaller sampling errors and fewer differences being judged significant than would have been uder an assumption of simple random sampling.

Table of Contents

PREFACE v

CHAPTER ONE

The Problem 1

The Data 8

The Analytic Approach 13

References 21

CHAPTER TWO

History of the Survey 25

Sample Design 26

Estimation 29

Variances 30

References 35

CHAPTER THREE

Selection of Variables 36

Variable Definition 44

Selecting the Population 49

Supplemental Data 60

The Study Population 61

References 76

CHAPTER FOUR

Categorical Data Analysis 78

Variances and Covariances 83

Models 86

Analysis 94

Discussion 120

References 127

CHAPTER FIVE

Discussion 130

Comments on Survey Data 137

References 139

APPENDIX ONE

NHIS Sample Design 141

References 153

APPENDIX TWO

Creation of the Dataset 154

APPENDIX THREE

Variables used in the Study 164

National Health Interview Survey 164

Area Resource File 173

References 177

APPENDIX FOUR

Comparison with Other Data 178

APPENDIX FIVE

Variable Selection Tables 185

APPENDIX SIX

Analytic Set-Up Files	203
Example One	204
Example Two	208
Example Three	211
Example Four	218
Computer Time	220
References	221

PREFACE

This work was begun to study the factors associated with children's use of medical care by using data from a major survey, the National Health Interview Survey, combined with information about the areas in which the children live. It was intended to bring new information to light on a substantive matter of public health interest.

In the process it developed into a methodological study. What had been anticipated as a simple analysis of a well-documented survey turned out to be much more complicated. For example, no one had documented the survey after the sample was redesigned in 1973. Most of Chapter Two and Appendix One consist of either previously unpublished material or material which had been published to describe the Current Population Survey without any reference to the National Health Interview Survey. It is not necessary to have all the detailed information about the sample design that is presented in Appendix One before analyzing data from the survey. It is certainly necessary to have some of the information to understand why traditional analytic methods are not appropriate.

Nor had anyone attempted a multivariate analysis of data from this large survey which incorporated the full and complex survey design. It was necessary to have the information needed to compute the variances added to the data tape, to install new software on the computer, and then to modify software so that it could handle a problem of this size. The analytic methods used in Chapters Three and Four are not new but they had to be explored and modified for use on a survey of 73,817 people with a complex sample design.

The methodology used here was not chosen to present the best or optimal analysis of such data. All of the steps taken were not absolutely necessary. They were taken, however, because the approaches had never been applied to such data from a large complex survey and it was necessary to explore the possibilities. This investigation will give others the opportunity to address questions that might not readily be addressed by other means.

Thus, instead of a report on a finished project, this is a report of the first stage of a long term research project. As each stage unfolded it became apparent that much more needed to be done. For many aspects of the analysis of data from large-scale complex surveys such as the National Health Inteview Survey, we don't know the best way to do it; all we can do is state our assumptions, beware of limitations, and

proceed with as much honor and honesty as we can. There are no final answers from this study. It provides more questions than answers but it should serve as a guide to others who want to use the data from this survey. The items on the public-use data tapes and the analysis of data from the National Health Interview Survey have changed as a result. That may be accomplishment enough.

It could not have been done without help and one of the nicest discoveries of the whole endeavor was how very helpful and generous people are. I am grateful to Jean Williams who linked many files to create the data tape, to Sandy Rothwell and others at the computer center of the National Center for Health Statistics who put up with me while I taught myself to create a SAS data tape and to do my own programming, to Wilbur Hadden who put much of the new software on the system for me (and then modified it when it wouldn't handle the problem), to Ron Forthofer who gave me (and then modified) his version of the categorical data analysis program and who helped me use it, to Babu Shah who gave me (and then modified) his variance-covariance program, and to my husband and children who furnished love, patience, and support to "Mom's dissertation".

CHAPTER ONE

Children and Medical Care

The Problem

Almost all children and youths in the United States today receive medical care. They are born in hospitals and during childhood they are expected to receive well-baby care, immunizations, school examinations, camp physicals, and routine periodic physical examinations. They obtain that care from private physicians and a variety of publicly funded facilities.

Most of the medical care children receive is ambulatory care. In the mid-1970's, children and youths under the age of 18 had about 12 times as many ambulatory medical contacts of all kinds and 5.5 times as many visits to office-based physicians as they had days in the hospital. Nevertheless, children and youths received less ambulatory care than adults did. They had, on the average, 4.1 contacts (including telephone contacts) with physicians while people ages 18-64 years of age had 5.2 and people 65 years of age and older had 6.7 physician contacts per person per year (NCHS 1977a, NCHS

1977b, Kovar 1981a)

In recent years over half of the children and youths have had a contact with a physician within 6 months and about three-quarters have had a contact within a year. Almost 90 percent have had a contact within 2 years and 97 percent within 5 years (NCHS 1977a, 1977b, 1978a). In 1975-76, those with at least one contact had, on the average, 5.6 contacts during the year (Kovar 1981a). The national data show that the use of medical care has been increasing and that the differentials between "poor" and "nonpoor" children have decreased (Wilson and White 1977).

If children are receiving as much medical care as the data show, why then is it of interest to examine the determinants of care for children? One answer is that the national averages conceal wide disparities among subgroups of children in the amount of care, the interval at which care is received, and the place of care. In so far as any of these are related to the quality of care, the disparities may be even greater. If the issue is equity of care, that is, equal utilization for equal need, the disparities are certainly greater than the routinely published data reveal (Kovar 1982).

It is easy to demonstrate the lack of equity in children's receipt of medical care using data from the National Health

Interview Survey. For example, if one examines the data using the use-disability ratio suggested by Anderson (1978), the "poor" children had 33 contacts per 100 days of restricted activity in contrast with 43 for the children in families with incomes of $15,000 or more (Kovar 1982). The comparable figures per 100 days of bed-disability were 72 and 98 and per 100 days of school lost they were 46 and 77. If days of disability are a valid measure of need, then there is certainly no equity among children living at different levels of family income. Children in low-income families are less likely to have received care even though low income is the criterion for participitation in most public programs.

The disparities are even greater for characteristics of the child (or family) which are seldom examined. Even using the usually published measure of contacts per person, which does not take need into account, the differentials by education of the head or by family size are great. In 1975-76 children in families where the head had 8 or fewer years of education had 2.8 contacts per child per year; children in families where the head had 16 or more years of education had 5.4 contacts per child per year - approximately twice as many. Children in families with 6 or more persons had 2.7 contacts per child per year; children in families with 3 or fewer persons had 5.9 contacts (Kovar 1981a,1982).

Newberger et al (1976) have said that the country has two

systems of health care: " private and frequently inadequate care for those who can pay; and 'public' care, or no care, for those who cannot." Five years later, the second of the five major concerns listed by the Select Panel for the Promotion of Child Health (Office of the Assistant Secretary for Health 1981) was: "The health status of American children has improved dramatically over the past two decades, but not all groups have shared equally in the progress. Sharp disparities persist in both health status and the use of health services according to family income, parental ethnic background, parental education, and geographic location."

Thus, long after the implementation of Medicaid and other Great Society programs, differentials among population subgroups remain. Some children are not receiving adequate medical care and it appears likely that those children most likely to need care, those whose parents have trouble caring for them well because of lack of money, knowledge, or other resources, are those least likely to receive care.

Despite the focus in public policy on assuring that poor children (especially poor children in single-parent families) have access to medical care, it has long been recognized that having the money to pay for care is only one of the factors influencing whether children receive the care they need.

Newberger et al, when referring to the "two systems of health care", said that "When services are provided for the poor, they frequently neglect minimal needs for privacy, dignity, and convenience." They thus attributed this problem to the providers of services to the poor as if all other providers of care did insure privacy, dignity, and convenience. Kenneth Kenniston and the Carnegie Council on Children (1978) also saw the differentials as being due to the way services are provided even though a study by the Institute of Medicine (IOM 1974) had found that the quality of care from private physicians was not necessarily any better than that from public facilities. Haggerty (1975,1976) showed that, at least in Rochester, poor people tended to use their old means of receiving care, outpatient departments and public clinics, even though (at that time) Medicaid would have paid for care from a private physician. The recipients resisted change. Miller (1976) wrote of the need to reach out and involve children who cannot themselves enter the system and on whose behalf no present initiative is involved to bring them in. The report of The Harvard Child Health Project (1977) contained indications that access to care was not as simple as providing money or the means to pay physicians. The review of data from surveys published by the Health Services Administration (1978) contained the data to alert people to influences other than income on childrens' receipt of medical care but there was no comment on the tables to bring the information to people's attention. Wolfe (1980), like

Haggerty working in Rochester (which she says has generous Medicaid benefits and a university-sponsored pediatric project), found that children in lower income areas or from larger families were less likely to see a physician over a one-year period than others and that women who work full-time seem to take their children to providers less frequently. Thus, according to her analysis, even when the benefits to pay for and the facilities to provide the care are available, differentials in the use of care remain.

However, the availability of medical care has been shown to be important - at least for preventive care. Kleinman and Wilson (1977) found that children and youths under the age of 17 were less likely to have had a routine physical examination if they lived in a medically underserved area than if they lived in an adequately served area. Grossman et al (1981) found that the probability of receiving a checkup by a doctor was positively related to family income and to the number of pediatricians per capita in the county of residence. They also found 'little evidence that preventive care delivered to youths by physicians is effective in terms of their physical health' - an interesting attribution of causality from cross-sectional data. However, their study is one of the few where measures of health status, utilization of medical care, and a measure of medical care availability were examined simultaneously. In most of the other studies differential need is either assumed or is inferred from

references to other work; it is not measured at the same time as utilization.

Because most of the care children receive is ambulatory and because almost all preventive care is ambulatory, it is important in the context of promoting health and preventing disease, the current focus of Federal policy (Office of the Assistant Secretary for Health 1979,1980,1981a,1981b), to examine factors associated with differentials in children's receipt of ambulatory care. The interest, however, is not in how many visits children have - which may be a function of a transitory illness or injury - but in whether they are receiving care at an interval short enough to infer that they have had the opportunity to receive the recommended preventive care.

The American Academy of Pediatrics has established guidelines on the frequency at which children and adolescents should be seen to receive basic preventive care. (American Academy of Pediatrics, 1977). These guidelines can be used to study factors which differentiate among groups of children by ascertaining which children have not had any medical care in an interval recent enough to have received the recommended minimal preventive care.

According to the Academy, it is at present impossible to establish the optimal number of visits or preventive proce-

dures for all children or their parents. Therefore, their publication, Recommendations for Preventive Health Care of Children and Youth, represents a guide for the care of well children who receive competent parenting, who have not manifested any important health problems, and who are growing and developing satisfactorily.

For these children, the Academy recommends visits at 2-4 weeks of age, 2-3 months, 4-5 months, 6-7 months, 9-10 months, 12-15 months, 16-19 months, 23-25 months, 36-37 months, 5-6 years, 8-9 years, 11-12 years, 13-15 years, and 16-21 years. It should be noted that the recommendations for intervals depend on the age of the child.

The Data

Ideally, we would have national data to ascertain how many children are receiving preventive care on the recommended schedule and whether they are receiving the services suggested by the Academy at each visit. We do not have that information on any national data system.

There are two major requirements for data to be used to study which children have received care. First, all children and youths must be included, not just those who have received medical care. Therefore, the data must be from a population-based (in contrast to a record-based) survey. Second, all

medical care has to be included regardless of where it was received. Therefore, data from surveys which include information only on care received in private physicians' offices or in any other specific source are not appropriate.

The National Health Interview Survey (NHIS) came closest of the available data sources to meeting these requirements. It is population-based and is designed to include all children in the United States except those residing in institutions. This exclusion is not important for the present study as the recommendations of the Academy were not designed to pertain to institutionalized children who are for the most part chronically ill children (Census 1978). The National Health Interview Survey also had the advantage of the capability of linking it to other files to obtain information on other members of the family and on characteristics of the county in which the child resided. This latter capability was especially important as most of the public health endeavors have been directed towards providing care to children in low-income areas or in areas where there are few physicians. There has been little study of whether children in such areas actually receive less care than those in areas with higher incomes or higher physician to population ratios or whether living in such an area has any relationship to whether care is received.

The NHIS has information on the number of contacts with a

physician during the 12 months preceeding the interview, on the place, kind of physician, and reason for the contact if it was within 2 weeks of the interview, and on the interval since the last time the person saw or talked to a medical doctor.

Of these, the interval since the last visit is the most important for considering whether the child is failing to receive adequate preventive care. While the data do not permit us to know that the child did receive preventive services if he or she did see a physician, it is certain that the services were not received if there was no contact with a physician.

The intervals precoded on the NHIS questionnaire are: less than 2 weeks, 2 weeks-5 months, 6-11 months, 1 year, 2-4 years, 5 or more years, and never. These intervals are not ideal (the addition of 3 months and 3 years as dividing points would be helpful) but it appears reasonable to assume that if children under age 2 have not had any contact with a medical doctor within 6 months, children ages 2-5 have had no contact within a year, or children ages 6-17 have had no contact within 2 years for any reason, those children are unlikely to be receiving adequate preventive care. This age-specific interval since last contact is the response variable of interest for this study. It is a standard-based measure of the kind suggested by Elinson and Kovar to

determine wether there is an unmet need for medical care (Elinson and Kovar 1980).

Information about the child's health status, place of residence, socioeconomic characteristics of the family, and other demographic characteristics are coded onto the data tape which contains a record for each person in the survey. The complete questionnaire is printed in "Current Estimates" for the appropriate year (NCHS 1977a, 1977b).

Information about other members of the family or information from the special supplements included on the questionnaire during a year are not included on the record for the individual and had to be linked to the child's record.

Very little information about the area in which the individual lives is recorded on the tape. However, the county of residence is coded on the tape (only on the confidential tapes, not on the public use tapes) and that information could be used to link the child's record to the Area Resource File (ARF) maintained by the Bureau of Health Manpower. The ARF, which has one record for every county or county equivalent in the United States, contains information on medical manpower, medical facilities, and physical, social, and demographic characteristics of the counties. The information is obtained from a variety of sources which are identified on the documentation (Bureau of Health Manpower 1978).

All of these files were linked for the 73,817 children and youths under 18 years of age who were included in the NHIS in 1975 or 1976. These were the most recent years for which data were available at the time this work was begun and it was believed at that time that two years would be needed to have a large enough sample. A description of the NHIS comprises Chapter Two and the details of the linkage of the various files used is given in Appendix Two.

Without giving the description of the survey at this point, it is necessary to emphasize that the NHIS is a highly stratified cluster survey and that complex adjustment factors are used to produce estimates for the civilian noninstitutionalized population of the United States living at the time of the interview. The estimation is done in such a way that estimates from this survey are compatible with those of the Current Population Survey for certain age, sex, and color groups. Because the sample estimates are inflated to the official population count, there is, by definition, no sampling error for the number of persons aggregated by these specific age-sex-color categories. There is sampling error for all other aggregations of the population, for people in those age-sex-color classes if they are further subclassified, and for all dependent variables from the NHIS. It is a complex sample and the analytic methods developed for simple random sampling are not appropriate.

The Analytic Approach

The effect of the ratio estimating procedures used in the NHIS is to make the sample more closely representative of the civilian noninstitutionalized population by age, sex, color, and residence. These procedures and the adjustments for nonresponse reduce the sampling variance. However, the complex sample design and estimation procedures also make the variances from the survey difficult and expensive to compute.

The development of the methodology and computer programs to compute these complex variances has been slow even though the problem has long been recognized (Simmons and Bean 1969). Even today there are only a few software packages available and, in general, they take a great deal of computer time and are expensive to run (Cohen 1979, Cohen and Gridley 1981, NCHS 1982). Therefore, it is necessary as a practical matter to have a strategy for data analysis which conserves computer time and uses the methods which take the complex survey design into account as effectively and efficiently as possible.

Many users of the National Center for Health Statistics data tapes ignore the complex sample design and assume simple random sampling. Using the readily available software packages such as SAS and SPSS they compute means, regression

coefficients, and other statistics and then test differences as if the data were derived from a simple random sample. One problem with this approach is that the design effect is usually (but not always) greater than one. That is, the variance computed with the full sample design taken into account is larger than that computed assuming simple random sampling. Thus, fewer differences will be found to be significant when the full complex sample design is taken into account than would be under the assumption of simple random sampling (NCHS 1982).

One approach to using the analytic methods which assume simple random sampling is to compute a design effect, that is a multiplier, to take the complex design into account after the analysis has been made but before the test statistics are computed. The design effect, defined as the variance from the complex survey divided by the variance from a survey of the same size assuming simple random sampling (DEFF = Vc / Vsrs), is then used as a constant multiplier for variances computed from the software packages to adjust the test statistics. A design effect can also be computed for the standard error by taking the square root of the design effect for the variance.

A second approach is to compute variance curves which are applicable to a "set" of statistics. This is the approach that has been taken in publishing data from the National

Health Interview Survey (NCHS 1970, 1977a, 1977b). Using the empirical knowledge that if (a) the sampling rate is uniform and (b) the variable is of the "count" type, the relative variance decreases as the size of the estimate increases, variances have been calculated for a selected number of statistics which are grouped into classes of wide, medium, and narrow. A range of subpopulations from very small to very large was chosen for each. A curve was then fitted to each set of estimates using the formula $y=a+b/x$.

However, there are several problems with this approach also. First, the curves do not always fit very well. Bean (NCHS 1970) pointed out that the fitted estimate of the relative variance was too large for the smaller estimates and too small for the larger ones in the curves based on the Keyfitz procedure. The fit of the curves based on the repeated replications procedure is much better. Second, they are not useful for studying relationships or testing hypotheses as they only give the variances - and then only for classes of estimates. Third, as the curves are published in the Vital and Health Statistics Series 10 publications, they are so difficult to read that the user can easily make an error that would affect the analysis.

A further problem with using the variance curves as calculated, or with using a design effect as a constant multiplier, is that the impact of the complex survey design is not

constant for all measures of either dependent or independent variables. For any given dependent variable, the design effect varies among population subgroups, and the variation depends on the nature as well as the size of the subgroup. Many outside users and the NHIS variance curves attempt to take the impact into account for the dependent variables (NHIS variance curves are computed separately for wide, medium, and narrow range variables with the grouping done on the basis of the type of statistic and its expected variance). The variance curves take the size of the subgroup population into account, but they do not take the nature of the independent variables into consideration.

The variation in design effects among population subgroups may not be trivial and it is necessary to consider the kind as well as the size of the subgroup. Consider, for example, the variation among some of the population groups used in this study as shown in table 1.1. The design effects range from 1.12 through 3.41 even though only one dependent variable is considered. The largest effect is found when children are classified by family size, which suggests that most of the effect among children is due to the clustering in households.

Table 1.1. Estimates of standard errors and design effects for whether children have received care within an adequate interval for specified subpopulations of children and youths under 18 years of age, National Health Interview Survey, 1975-76 annual average.

	Number in Thousands	Proportion with Adequate Interval	Standard Error		Design Effect	
			SRS	Complex	S.E.	Variance
Total	73,234	.8566	.001295	.002152	1.66	2.76
Mother has had contact for medical care within year						
Yes	56,563	.8829	.001352	.001933	1.43	2.04
No	11,249	.7315	.004178	.006855	1.64	2.69
Phone in the household						
Yes	65,061	.8671	.001331	.001948	1.46	2.14
No	7,852	.7705	.004746	.008343	1.76	3.09
Family size in number of persons						
1-3	12,789	.9090	.002543	.002693	1.06	1.12
4	21,063	.8963	.002101	.002687	1.28	1.64
5	17,081	.8598	.002656	.003170	1.19	1.42
=>6	22,301	.7864	.002744	.005067	1.85	3.41
Education of the family head in completed years						
0-11	24,771	.7953	.002563	.004240	1.65	2.74
12	25,888	.8678	.002105	.002960	1.41	1.98
>12	21,874	.9161	.001875	.002383	1.27	1.62
Family income in thousands of dollars						
<5	8,418	.8186	.004200	.005437	1.29	1.68
5-9.9	15,171	.8231	.003098	.004678	1.51	2.28
10-14.9	17,558	.8543	.002663	.003760	1.41	1.99
=>15	26,697	.8932	.001891	.002822	1.49	2.23
Age of the child in years						
<6	21,130	.8789	.002244	.002796	1.25	1.55
6-11	24,298	.8693	.002157	.003257	1.51	2.27
12-17	27,806	.8287	.002265	.003284	1.45	2.11
Standard Metropolitan Statistical Area						
Yes	53,324	.8701	.001456	.002415	1.66	2.75
No	19,910	.8204	.002720	.004562	1.68	2.81
Outpatient department in the county						
Yes	49,782	.8718	.001498	.002514	1.68	2.82
No	23,452	.8242	.002486	.004182	1.68	2.83

A third approach is to calculate the variances for the estimates and assume that the covariance is zero. This is the approach that has been used by Freeman and coworkers for most of their investigations of the NHIS (Freeman et al 1975,1976). A problem with that approach, as they have demonstrated (Freeman and Brock 1978), is that the covariances are not always zero and are frequently negative. Thus, this approach, which had been thought to be conservative under the assumption that the covariances are positive, is not necessarily conservative. As yet there has not been sufficient work to demonstrate a pattern of when the covariances will be positive and when negative.

The variation among population subgroups in the size of the design effects is similar to that found by Hidiroglou and Rao (1981) for the Canada Health Survey. Their conclusion is that design effects should be published for each cell and marginal for a two-way table and also for each layer in a three-way table. (They have not yet addressed tables with four or more classifications.) The presentation of all of these effects is necessary so secondary users of the data can address any hypotheses they wish. Using their approach, the full variance-covariance matrix is not necessary, but the computation time is still great and the size of any published tables would still be doubled. It is an interesting approach that has not yet been applied to substantive analysis of any

survey.

Another approach, still under development, is that which Choi presented in his dissertation and has been carrying forward in a series of papers (Choi 1980,1981). His procedure to account for the clustering within households and to correct the chi-square statistic for that clustering is based on calculating an "effective" sample size and then using traditional methods of calculating variances and covariances that assume simple random sampling. One does not need to take the complex design into account. The later work consists of developing an average correction factor so that the correction does not have to be calculated for each individual estimate. This approach, which may be extremely useful in the future, is not yet developed to the point where it can be used in multivariate analysis. It does not yet take account of the clustering within segments or PSU's and, like the design effect or variance curve approaches, does not yet take the differential impact of the clustering on different population subgroups into account.

Finally, it is possible to calculate the full variance-covariance matrix for every step of an analysis. Given the extremely large sample and the large number of primary sampling units in the National Health Interview Survey, this approach, while theoretically possible, is not always feasible. It is necessary, however, to use the full matrix,

to account for both variances and covariances from the complex survey, before definitive statements are made. At the very least, if for example only the variances are used, that should be stated along with the possible implications for the analysis.

The approach adopted in this analysis is a combination of the ones listed above. At the early stages of selecting variables for inclusion in the analysis and the levels to be used for each variable, simple random sampling was assumed and all analyses were done as if the survey were a simple random sample and the estimates were those derived from the sample rather than the weighted national estimates. However, to guard against the possibility of excluding a variable from the later stages if the covariance was negative or the design effect was less than one, a cut-off point larger than 0.05 should be used. Where feasible or useful, the knowledge that the design effect was 2.76 overall was used to either change the significance level of the test statistic or, following Choi, to reduce the effective sample size to 36 percent of the actual size of 73,817. Another approach, which was not used, would have been to select only one child from each household thus reducing the clustering of children in families. Finally, the complete variance-covariance matrix was estimated and used for the final stages of fitting the model using the weighted least squares methodology proposed

by Grizzle, Starmer, and Koch (1969).

The weighted least squares approach is particularly appropriate for data from a survey with a complex sample design. Also, it is appropriate for the analysis of categorical data and much of the data from the National Health Interview Survey are categorical either by their nature (marital status, whether residence is in a Standard Metropolitan Statistical Area or not) or by the way they are precoded on the questionnaire (family income, interval since last contact with a doctor or dentist). The data presentation in the Series 10 publications of the National Center for Health Statistics is always categorical. Continuous variables such as age and level of education are grouped into categories. Both from the nature of the data and the usual way of presenting data from the survey, a categorical approach was appropriate.

There was a further reason for using a categorical approach. The study was intended to be useful for policy, and people who make policy seem to find a categorical presentation more comprehensible than presentations based on, say, regression analysis. This may be due to the lack of statistical expertise among policy makers or the lack of communication expertise among statisticians. It is probably due to both. Nevertheless it is real and should be accepted as reality.

References

American Academy of Pediatrics. Standards of Child Health Care. Committtee on Standards of Child Health Care, Evanston, IL., 1977.

Anderson,R. Health Status Indices and Access to Medical Care. Am.J.Pub.Health 68(5):458-463, May 1978.

Bureau of the Census. 1976 Survey of Institutionalized Persons: A Study of Persons Receiving Long-Term Care. Current Population Reports. Series P-23, Number 69. U.S.Government Printing Office, June 1978.

Bureau of Health Manpower. The Area Resource File; ARF A Manpower Planning and Research Tool. DHEW Publication Number(HRA) 80-4. Hyattsville,Md. October 1979.

Choi,J.W. Analysis of Categorical Data in Weighted Cluster Sample Surveys. Proc.Soc.Stat.Sec. of the American Statistical Association. 1980.

Choi,J.W. A Further Study on the Analysis of Categorical Data from Weighted Cluster Sample Survey. Presented at the American Statistical Association Annual Meet-ing, August 1981.

Cohen,S.B. An Assessment of Curve Smoothing Strategies Which Yield Variance Esimates From Complex Survey Data. Presented at the American Public Health Association Annual Meeting, November 1979.

Cohen,S.B. and G.Gridley. Present Limitations in the Availability of Statistical Packages for the Analysis of Complex Survey Data. Presented at the American Statistical Association Annual Meeting, August 1981.

Elinson,J. and M.G.Kovar. Strategies and Tactics of Gauging Unmet Health Care Needs in Local Populations. Presented at the American Public Health Association Annual Meeting, November 1980.

Freeman,D.H., J.L.Freeman, G.G.Koch, and D.B.Brock. An Application of Categorical Data Analysis to the National Health Interview Survey. Presented at the American Statistical Association Annual Meeting, August 1975.

Freeman,D.H., J.L.Freeman, G.G.Koch, and D.B.Brock. An Analysis of Physician Visit Data from a Complex Sample Survey. Am.J.Pub.Health 66(10): 979-983, October 1976.

Freeman,D.H. and D.B.Brock. The Role of Covariance Matrix Estimation in the Analysis of Complex Sample Survey Data. Academic Press, Inc. 1978.

Grizzle,J.E., C.F.Starmer, and G.G.Koch. Analysis of Categorical Data for Linear Models. Biometrics (25), 489-904,1969.

Grossman,M., D.Coate, L.N.Edwards, R.A.Shakotko, D.Chernichovsky. Determinants of Children's Health. NCHSR Research Summary Series. DHHS Publication Number(PHS) 81-3309. Hyattsville, Md. July 1981.

Haggerty,R.J. Who Will Monitor Access?. Pediatrics (57), 169-170, February 1976.

Haggerty,R.J., K.J.Roghmann, I.B.Pless. Child Health and the Community. New York: John Wiley and Sons,1975.

Harvard Child Health Project. Report of the Harvard Child Health Project Task Force. Volumes I-III. Cambridge, Ma.: Ballinger Publishing Co., 1977.

Health Services Administration. Health Status of Children: A Review of Surveys 1963-1972. DHEW Publication Number(HSA) 78-5744. Washington,D.C.: U.S. Government Printing Office, 1978.

Hidiroglou,M.A. and J.N.K. Rao. Chisquare Tests for the Analysis of Categorical Data from the Canada Health Survey. Presented at the International Statistical Institute, Buenos Aires, Argentina, December 1981.

Institute of Medicine. Assessment of Medical Care for Children. Contrasts in Health Status, Vol. 3. Washington,D.C.: National Academy of Sciences, 1974.

Kenniston,K. and The Carnegie Council on Children. All Our Children. New York: Harcourt Brace Jovanovich, 1977.

Kleinman,J.C. and R.W. Wilson. Are "Medically Underserved Areas" Medically Underserved? Health Services Research 12(2): 147-162, Summer 1977.

Kovar,M.G. Children and Youth: Health Status and Use of Health Services. Health United States, 1978. DHEW Publication Number(PHS) 78-1232. Public Health Service. Washington,D.C.: U.S.Government Printing Office, December 1978.

Kovar,M.G. A Statistical Profile: Volume III of Better Health for our Children: A National Strategy. The Report of the Select Panel for the Promotion of Child Health. DHHS Publication Number(PHS) 79-55071, Public Health Service. Washington, D. C.: U.S.Government Printing Office, February 1981.

Kovar.M.G. Children's Health Status and Use of Medical Care. Public Health Reports Jan-Feb,1982.

Miller,C.A. A Bicentennial Generation of Healthy Children. Presented at the Bicentennial Conference on Children, National Council of Organizations for Children and Youth, February 1976.

National Center for Health Statistics. Current Estimates from the Health Interview Survey: United States-1975 by T.F.Drury. Vital and Health Statistics. Series 10, Number 115. DHEW Publication Number(HRA) 77-1543, Public Health Service. Rockville,Md.: U.S.Government Printing Office, March 1977.

National Center for Health Statistics. Current Estimates from the Health Interview Survey: United States-1976 by E.R.Black. Vital and Health Statistics. Series 10, Number 119. DHEW Publication Number(PHS) 78-1547, Public Health Service. Hyattsville,Md.: U.S.Government Printing Office, November 1977.

National Center for Health Statistics. A General Statistical Methodology for the Analysis of Data From a Complex Survey: NHANES I by Landis,J.R., J.M.Lepkowski,

S.A.Eklund, and S.A.Stehouwer. Vital and Health Statistics. Series 2, Number 92. DHHS Publication Number(PHS) 82-1366, Public Health Service Hyattsville,Md.:U.S.Government Printing Office, September 1982.

Newberger,E.H., C.M.Newberger, and J.B.Richmond. Child Health in America: Toward a Rational Public Policy. Milbank Memorial Fund Quarterly 249-298, Summer 1976.

Office of the Assistant Secretary for Health and Surgeon General. Healthy People - The Surgeon General's Report on Health Promotion and Disease Prevention, 1979. DHEW Publication Number(PHS) 79-55071. Public Health Service. Washington, D.C., U.S. Government Printing Office, 1979.

Office of the assistant Secretary for Health and Surgeon General. Promoting Health/Preventing Disease - Objectives for the Nation. Washington D.C.: U.S. Government Printing Office, 1980.

Office of the Assistant Secretary for Health. Better Health for our Children: A National Strategy. The Report of the Select Panel for the Promotion of Child Health. DHHS Publication Number(PHS) 79-55071, Public Health Service. Washington, D. C.: U.S.Government Printing Office, February 1981.

Office of the Assistant Secretary for Health. The Surgeon General's Workshop on: Maternal and Infant Health. DHHS Publication Number(PHS) 81-50161.

Simmons,W.R. and J.A.Bean. Impact of Design and Estimation Components on Inference in New Developments in Survey Sampling eds. N.L.Johnson and H.Smith Jr. Wiley-Interscience,New York,1969.

Wilson,R.W. and E.L.White. Changes in Morbidity,Disability, and Utilization Differentials Between the Poor and the Nonpoor. Data from the Health Interview Survey, 1964 and 1973. Medical Care 15(8): 636-646, August 1977.

Wolfe,B.L. Children's Utilization of Medical Care. Medical Care 18(12): 1196-1218, December 1980.

CHAPTER TWO

The National Health Interview Survey

History of the Survey

The National Health Interview Survey (NHIS), one of a program of surveys designed to collect data on the health and use of health services of the population of the United States, was authorized by the National Health Survey Act of 1956. A pretest of 1,200 households was conducted in Charlotte, N.C. in February 1957. During May and June of that year the nationwide operation was given a dry run; interviewing and editing proceeded as if the survey were in operation (NCHS 1958). Official data collection began in July of 1957 and the survey has been continuously in the field ever since. The sample has been updated after every decennial census and the questionnaire has been modified at approximately ten-year intervals but there has been very little change in the concepts and procedures of the survey. Opinions about what can and should be collected by means of household interviews or who is an eligible respondent, and the presentation of the point estimates and variances remains virtually unchanged. Publications still show only descriptive statistics and variance curves. The paired primary sampling units (PSU's), which are needed to compute variances, are not even on the

data tapes. No one outside the NCHS can analyze the data with much confidence because the required information is not published or on the public-use data tapes.

It is a household interview survey in which one respondent answers for all children in the family and for all adult family members who are not at home when the interviewer calls. The response rate is remarkably high with about 97 percent of the occupied households selected in the sample participating in 1975. The response rate for households with children is probably higher as it is easier to find a respondent at home in families with several members than in one-person households and about half of the nonresponse was due to the failure to find an eligible respondent at home after repeated calls. In 1975 there were 37,978 and in 1976 there were 35,839 children and youths under age 18 in the interviewed households.

Sample Design

Details of the sample design are given in Appendix One. Only a brief overview is given here.

From the beginning, the Bureau of the Census has been responsible for the sample design and for the data collection of the NHIS. Historically, therefore, the NHIS is linked with the Current Population Survey (CPS) which is conducted

primarily to produce data on labor force characteristics of the U.S. population. The sampling frames, developed for the CPS, are used to select the sample for the NHIS although the latter is conducted independently. The CPS which had been based on a probability sample since 1943 (Census 1963) and the principles and procedures used for it were adopted when the NHIS was designed in 1956 (DHEW 1958b). Understanding the origin and the linkage makes it easier to understand the NHIS design and estimation procedures.

The NHIS sample design is based on the design for the national sample of the Current Population Survey (CPS). The two samples are selected simultaneously and most of the literature on the CPS also applies to the NHIS (Census 1963, Census 1978). Details about the sample design have been abstracted from the two Census publications and, along with information from NCHS staff, are given in Appendix One.

Without going into the detail, there are aspects of the design that are necessary for understanding how to use data from the NHIS. No one should try using data from this or any other survey with a complex design without some understanding of how the sample was drawn.

The sample is designed to produce national estimates of the civilian noninstitutionalized population residing in the United States. The approach to doing so is to divide the

United States into geographically defined areas called primary sampling units (PSU's). A PSU consists of a county, a small group of contiguous counties, or a Standard Metropolitan Statistical Area (SMSA). The PSU's collectively cover the 50 states and the District of Columbia.

The 1,924 PSU's are classified into 376 strata - combinations of PSU's with similar characteristics. One PSU is then selected from each of the 156 self-representing strata and one from each of the 220 non-self-representing strata (modifications of this procedure are discussed in Appendix One).

Within the selected PSU's, small clusters of 4 housing units are then selected. This stage of selection is accomplished by listing small areas of the PSU called segments and then selecting 4 adjacent housing units from the segment.

Since each family member in the selected housing units is in the sample, there is clustering within the household, within the segment, and within the PSU. It is this clustering that causes the procedures for estimation and analysis to differ from those in simple random sampling.

Another important aspect of the design is that it is a multistage probability design which permits a continuous sampling of the civilian, non-institutionalized population of

the United States and is designed in such a way that each week is an independent sample of the population; the weekly samples are additive over time. Thus, the design permits estimates for high-frequency measures or for large population groups to be produced from a short period of data collection and estimates for low-frequency measures or for smaller population subgroups to be obtained from a longer period of data collection. Because interviewing is done throughout the year, about 800 households are interviewed each week, there is no seasonal bias.

Estimation

The survey is designed to make inferences about the civilian noninstitutionalized population of the United States. To do so it is necessary to inflate the sample counts to national estimates.

Four basic operations are involved:

1. Probability of selection inflation. The estimates are multiplied by the reciprocal of the product of the probabilities of selection from each step in the design (PSU, segment, and household).
2. Nonresponse adjustment. The estimates are inflated by a multiplication factor which has as its numerator the number of sample households in a given segment where interviews should have been obtained and as its denominator the number of households in that segment where interviews actually were obtained.
3. First-stage ratio adjustment. Sampling theory indicates that the use of auxiliary information which is highly correlated with the variables being estimated improves the reliability of the estimates. To reduce the variability

among the PSU's within a region, the estimates are ratio adjusted to the 1970 populations within 12 color-residence classes within each region.

4. Poststratification by age-sex-color. The estimates are ratio adjusted within each of 60 age-sex-color cells to an independent estimate of the population of each cell for the survey period. Both the first-stage and the poststratified ratio adjustments take the form of multiplication factors applied to the weight of each elementary unit (household, person, condition, hospitalization). These independent estimates are prepared by the Bureau of the Census.

Variances

An estimate of the reliability of the results from a probability sample can be made from the sample itself. The form of the variance estimator depends on the sample design and the estimation procedures. It is desirable that the variance estimator be unbiased (and it should be consistent if not unbiased), and that it have a small sampling error. It is also desirable for computation to be simple.

At the time the National Health Interview Survey was initially designed, there had been very little work on methods of estimating variances for surveys with complex sample designs. Computers were slow, awkward, and had very little memory. Computer software was virtually nonexistent. Multivariate analysis couldn't even be contemplated. Calculating and presenting sampling errors for the national estimates from the NHIS in the reports was a great advance in the presentation and interpretation of data.

Initially variances for the NHIS were calculated by a technique called the random group method (NCHS 1958). The method was used for data collected in 1957 and until the variances were updated in 1972. The variance charts constructed in 1961 based on the random group method of variance estimation were used until the late 1960's when some of them were replaced by charts based on the Keyfitz estimation procedure (Keyfitz 1957). For data collected in 1972, the Keyfitz estimation procedure was used to calculate a complete set of new variance curves. The Keyfitz procedure was selected rather than the balanced half-sample replication variance estimator because of the large number of strata in the NHIS design. At that time, no one knew how to form the orthogonal patterns for more than 100 strata.

A report on the research on the Keyfitz procedure (NCHS 1970) also gave the argument for presenting curves for classes of estimates: " Instead of presenting variances for each statistic, the data can be grouped and 'average variances' given. ... Grouping the statistics is not easy, but two points to consider are that the survey characteristics such as prevalence of any diseases represented in a group should have similar design effects and that the groups should cover the possible range of variation in the data." (NCHS 1970). Intead of presenting the individual estimates, a set of relative variance curves was drawn using the empirical relationship between the size of the estimate and the

estimate's relative variance.

When the redesigned sample for the survey was implemented in 1973, new relative variance curves were needed. Balanced half-sample replication had been considered before, and early work on the method had been funded by the Center (NCHS 1966). In the meantime there had been further work (NCHS 1969), and a program had been written to apply the technique to the data from the National Health Examination Survey (NHES now NHANES). At that time no one knew how to form orthogonal patterns when there were many PSU's. That problem was solved when George Schnack formed the orthogonal patterns for more than 100 strata. The program was extended so that it could be applied to the NHIS (NCHS 1977). Using this program, curves of relative variances were calculated based on data from 1973 (NCHS 1977). Those calculations are the basis for the variance curves published in Series 10 reports for data from 1973 through the present.

These variance curves are published in all Series 10 reports from the NHIS. There are curves for wide-range, medium-range, and narrow-range response variables to allow for the differing distributions of the variables and their design effects. Curves are calculated for each of the recall periods used on the questionnaire, and they are either calculated or adjusted to allow for the sample size. The curve for this study is the one for two years of data

collection of a narrow-range variable with a one-year recall period. A copy is given following the last appendix.

These curves serve a useful purpose by allowing the reader to obtain an estimate of the variability of the statistic of interest quickly. However, they are not sufficient for relational analysis for a number of reasons. Most important is that, although they take the size of the denominator into account, they do not take the differential effects of the clustered design into account. For example, there are approximately 22 million children in families with six or more members in the household and there is approximately the same number in families where the head has 12 or more years of education. Nevertheless, as can be seen in table 1.1 of Chapter One, the design effect for the first category is 3.41 and for the second it is 1.62. The standard error of the proportion for the first group is twice that for the latter.

It is also important for the user to recognize that each time the curves have been recalculated, they have been used without modification in all succeeding years until the sample is redesigned as if there has been no change in the population or the dependent variables over an extended time period. The effect of the changes over time is not trivial. In evaluating the change from 1973 through 1979, Givens found that the weighted relative standard errors for each of the eight curves he evaluated showed consistent increases over

time because of the attrition in sample size (Givens 1982). Moreover, the values predicted from the curves differed from the computed values. In 1980, which was a special case because the sample was reduced by four weeks, one third of the variance esimates were in error by more than 20 percent.

It is also important to note that the NHIS was designed with one sample PSU selected from each stratum (there is more than one if the stratum is an non-self-representing stratum and the PSU is a rotation cluster containing more than one PSU, although it is still treated as one). The calculation of variances requires two PSU's in each stratum. Therefore, pseudo strata, each containing 2 PSU's had to be created. Strata, each containing one PSU, were paired to form pseudo strata, each containing two PSU's. This was done in the self-representing strata by ordering the segments and then putting the even-numbered segments into one pseudo-PSU and the odd-numbered segments into another to form a stratum with two "PSU's". The non-self-representing strata were paired with one another and a pseudo PSU number was assigned to one member of the pair and that number plus one to the other (Schnack 1974).

The number of persons in the total U.S. population or in an age-sex-color class used for poststratification is not subject to sampling error because these numbers are adjusted to figures supplied by the Bureau of the Census (NCHS 1977,

1978). None of these 60 classes were used in this study. Therefore, all population estimates given here are subject to sampling error.

References

Bureau of the Census. The Current Population Survey -- A Report on Methodology. Technical Paper No. 7. U.S. Government Printing Office, Washington,D.C. 1963.

Bureau of the Census. The Current Population Survey -- Design and Methodology. Technical Paper No. 40. U.S. Government Printing Office, Washington, D.C. January 1978.

Keyfitz,N. Estimates of Sampling Variance Where Two Units Are Selected from Each Stratum. JASA (52)280, 503-510, Dec., 1957.

National Center for Health Statistics. The Statistical Design of the Health Household-Interview Survey by staff of the U.S.National Health Survey and the Bureau of the Census. Health Statistics. PHS Publication Number 584-A2. Public Health Service. Washington,D.C. U.S.Government Printing Office, July 1958.

National Center for Health Statistics. Replication, an Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. PHS Publication Number 1000, Series 2 Number 14. Public Health Service, Washington,D.C. U.S. Government Printing Office, April 1966.

National Center for Health Statistics. Pseudoreplication; Further Evaluation and Application of the Balanced Half-Sample Technique by P.J.McCarthy.Vital and Health Statistics.DHEW Publication Number (HSM) 73-1270. Public Health Service, Washington,D.C., January 1969.

National Center for Health Statistics. Estimation and Sampling Variance in the Health Interview Survey by J.A.Bean. Vital and Health Statistics. PHS Publication Number 1000-Series 2-Number 38. Public Health Service, Washington,D.C. June 1970.

Schnack,G. Memorandum to Madigan on HIS variance-1973 dated April 3, 1974.

National Center for Health Statistics by G.Jones. HES Variance and Cross Tabulation Program,Version 2. March 10,1977. Unpublished.

Givens,J. Memorandum to Sirken on NHIS variance comparisons 1973,1979,1980. dated April 12,1982.

CHAPTER THREE

The Variables and the Study Population

Selection of Variables for Analysis

The selection of variables for analysis is a mixture of art, science, and luck. The analyst is always limited to the variables available on the dataset and seldom, especially in the secondary analysis of data which were collected for another purpose, will have the luck to find everything desired. The limitation to available data is a point too often ignored in reporting findings. Study data are often reported as if everything important was included and some sizable proportion of the variance is "explained".

The problem of selecting variables for inclusion in a model has received little attention in the literature compared to that given to selecting models or other aspects of statistical analysis. A search of the substantive literature reveals that papers generally begin with a literature review and then continue with a description of what the authors did. Sometime a hypothesis is stated; more often it is not. Economists appear to be more likely than demographers or

social scientists to have models and hypotheses. Rosenzweig and Schultz (1980) and Grossman (1980), for example, used data from NCHS with the variables selected as being reasonable for inclusion in their explicitly stated economic models.

It appears, then, that the selection of variables to be included in the analysis is often made on substantive, and sometimes on theoretical, grounds with the decision based on the results of previous research and a literature search. There is a risk, however, in using the same variables that have been used by others. The previous research may have been based on limited data which did not include all of the variables that were important even at the time that research was done. In addition, conditions may have changed so that the variables of importance when the original research was done have faded from importance while others, unimportant at the time, have become critical. This may be especially important in social research. For example, Hoffman (1980) has pointed out the dangers of interpreting research on the effect of working mothers or single-parent families on children which was done a decade or more ago. Then both working mothers and single-parent families were relatively rare but now both are common. The effect, if it existed at all, may be different when "deviant" behavior has become "normative".

It is desirable to have a method of examining all available variables, ranking them in importance according to some criterion, and selecting the most important. In the large dataset with many variables used for this analysis, a rational method of selecting the variables was crucial. Computational methods for simultaneously examining all the available information are simply not available. It would have been possible, of course, to have included the information which has been used in previous research on who receives health and medical services without any changes. If that had been done, it would still have been a contribution as much of the previous research refers only to adults or, occasionally, families. Repeating the same studies for children might have revealed previously undetected differences between the reasons children and adults receive medical care. It would also have been possible to have selected variables on the basis of the marginal distributions but, as Lilienfeld (1981) and others have pointed out, basing decisions on the effect of one factor alone may fail to reveal strong associations that exist in the presence of another factor.

Step-wise regression is sometimes used to screen and select variables. The major problem with using step-wise regression to screen the variables for an analysis such as the present one is the assumption of linearity. Usually there is no reason to assume that the relationships among the variables, if they exist, are linear. While fitting quadratic or higher

order equations is theoretically possible, it would not have been computationally feasible for these data with 73,817 observations and over 50 possible variables. Many of the variables are categorical which would have necessitated the creation of dummy variables. It is also difficult to allow for interactions in using regression as a screening technique. In addition, in using step-wise or other regression, one doesn't know the number of observations in each cell unless separate tabulations are made for that purpose. Everything appears to be equal and the analyst does not develop a "feel" for the data.

Higgins and Koch (1977) proposed a procedure for categorical data that is analogous to step-wise regression. The algorithm they proposed is designed to obtain a reasonable subset of variables. It may not be the "best" subset and it may not contain all of the variables of interest but it will have been formulated with two considerations in mind:

"The first is that once these variables are taken into account, none of the other variables have statistically significant effects. The second is that the total variation in the proportions associated with these variables is maximized relative to the associated degrees of freedom, which is analogous to maximizing the mean square due to regression in regression analysis."

The procedure uses Pearson chi-square statistics divided by their degrees of freedom to measure the relative importance of variables. The first variable selected is the one having

the largest chi-square per degree of freedom with regard to the first-order relationship with the response variable. This variable is combined with each of the others to produce higher order contingency tables and the additional variables are selected one at a time by similar rules using the chi-square per degree of freedom for the higher order contingency tables.

Two types of statistics are used for terminating the selection of statistically important variables: (1) the Pearson chi-square statistics of the relationship of a specific variable summed over the variables that have already been chosen, and (2) a statistic discussed by Mantel and Haenszel (1959), but without the continuity correction, which possesses a chi-square distribution and combines information with respect to the effect of a specific variable on the response variable over all possible combinations of previously selected variables.

All of these statistics are based on the assumption that the data are derived from a simple random sample. The National Health Interview Survey is not a simple random sample, however, but is a stratified sample with clusters of households in the segments selected in the sample, and with clusters of children in the households. Such clustering violates the assumptions underlying the statistics.

The violation is critical when the statistics are being used to terminate variable selection. In that case, the variables are being used for significance testing and neither the Pearson nor the modified Mantel-Haenszel statistic should be used as calculated for a termination statistic. There is no well-defined process that one can use as a rule.

The violation of the assumption of simple random sampling is less critical where the statistics are being used to determine the relative ranking of the variables rather than to determine statistical significance. If all are subjected to the same design effect, their importance relative to one another is not affected.

However, in the NHIS the design effect is not the same for all of the control variables (Table 1.1). The effects of clustering will be modified by the effects of the post-stratification procedures which take into account some, but not all, of the characteristics of the individual and the household. However, for the purpose of screening, the assumption of simple random sampling is reasonable enough to use the procedure.

The chi-square computed as the sum of the chi-square statistics from individual tables where each table is defined by one of the combinations of previously selected variables was used for screening rather than the screening statistic

proposed by Higgins and Koch. It provided an excellent opportunity for exploratory data analysis. It was possible to examine the individual tables and their chi-square statistics to see which population subgroups were contributing a large proportion of the total chi-square and which were contributing very little. If it became necessary to combine levels as the number of subpopulations into which the sample was divided increased, the information was available to combine those which were contributing least to the heterogeneity. The chi-square statistics for the higher-order tables were also calculated for the second and third level of variable selection for the study population so that the information would be available to determine whether different variables would have been selected if the Higgins-Koch procedure had been used.

Chi-square statistics for complex survey designs can be adjusted through the use of a design effect (Rao and Scott 1981, Hidiroglou and Rao 1981) or by calculating an effective sample size. One such procedure for obtaining an effective sample size is that proposed by Choi and McHugh (1980) and simplified by Choi (1981). As Choi pointed out, the use of the chi-square statistic to test hypotheses is generally inappropriate when the data arise from a cluster survey because of possible dependence between members of the same cluster. The chi-square is too large and leads to finding more differences significant than would be the case if the

clustered sample design were taken into account. (It is intuitively obvious that children living in the same household are more likely to be treated the same way than children with similar characteristcs living in different households.) To correct for the household clustering, he suggested dividing the chi-square by the factor "K" where:

$K = w(1+a\{m-1\})$, and

w = mean weight for national estimates
a = the proportional distribution of persons by the number of persons in the household
m = mean number of persons(children) per household

For this study, Choi's correction factors were calculated for households with children under age 18 because only children under age 18 are included in the study and the number of households containing children, the distribution of children by number in a household, and the mean number of children per household differ from the equivalent data for the total population. For this case:

	1975	1976	1975-76	unweighted
w	1744.71	1818.82	888.75	1
a	0.68	0.68	0.68	0.68
m	2.14	2.12	2.13	2.13

Thus, the correction factor for the chi-square statistics for the unweighted data used to screen the variables was: K = 1.77. If weighted data had been used, the factor would have been 888.75(1.77). This divisor could be applied to all

chi-square statistics before testing for significance. Because it is a constant divisor, the chi-square statistics are all affected equally. The uncorrected statistics are shown in the tables in Appendix Five, however, so that the reader will have the basic data.

The Higgins and Koch procedure as modified was used, therefore, to select the variables for analysis but before the procedure could be implemented effectively, there were several decisions that needed to be made. The decisions could be divided into several categories. The first category consisted of decisions about how to define appropriate levels for the variables and to eliminate redundant variables.

Variable Definition

Since the procedure is designed for use with categorical data, it was necessary to convert all the continuous variables into categorical ones and, in many cases, to reduce the number of categories for those variables which had too many categories.

The response variable, described in Chapter One, was already a categorical variable. If the child had received medical care within an interval short enough to have received the recommended preventive care, the response was coded Yes. Otherwise, it was coded No.

Redefining the control variables created a number of interesting problems.

Pragmatically, it was not feasible to have more than four categories for any variable or the data would be too thin for analysis. The limitation to four categories may reduce the power of the test (Cochran 1952). However, it was necessary for practical reasons and, because of policy interests, the classes were as small as feasible at the lower tail of the distribution. Combinations in that region were one of Cochran's concerns (Cochran 1954).

The general approach was to try to create a manageable number of levels for each variable which (1) led to heterogeneity of the response and (2) had enough children at each level to maintain adequate numbers in the cells for analysis. With a sample of 73,817 children, the second criterion might appear to be unnecessary. However, small numbers in the cells quickly became a problem despite the precaution.

The basic approach was to create categories on the basis of policy considerations, analytic needs, or on the basis of previous research.

Public policy, when it is used to encourage access to and utilization of medical care, is directed toward the poor

(Medicaid), the poor in families with no father (Medicaid), areas where a high proportion of the people are poor (Neighborhood Health Centers), or the so-called Medically Underserved Areas (National Health Service Corps). Therefore, instead of creating categories with equal numbers of children in each, finer divisions were made at the lower end of the economic or health resource scale and larger ones at the upper end in an endeavor to distinguish those children who, on the basis of economic considerations, are of policy interest from those who, on the same basis, are not. Sample distributions for some of these variables are given in Appendix Three.

A great deal of earlier work by the author, using data from all of the National Center for Health Statistics data collection systems that contain information on medical care utilization by children and youths, had demonstrated that the medical-care utilization patterns of young children, elementary-school-age children, and secondary-school-age children are different in many respects. Therefore, the 18 possible ages of the children were divided into three age categories (under 6, 6-11, and 12-17) on the basis of this knowledge.

A number of the variables were redundant - at least to some degree. For example, per-capita income and median family income in the county of residence are both measures of the

general economic status of people in the county. The prevailing charges for initial visits and for visits for established patients are both measures of medical care prices prevailing in the county. Simple 2 by n tables were calculated for all such variables and one of the redundant variables was eliminated by selecting the one with the higher value of chi-square divided by the number of degrees of freedom for further consideration.

For the variables about which little was known in advance, a different - more exploratory - approach was used. The approach was implemented differently for the categorical and the continuous control variables.

Two-way cross-classification tables were created for each of the categorical variables listed in Appendix Three by the response variable. For each table the Pearson chi-square (assuming simple random sampling) was computed and the chi-square was divided by the number of degrees of freedom. If adjacent levels had similar responses, the levels were combined and the chi-square divided by the degrees of freedom was recomputed. The number of levels for which this measure of heterogeneity was highest was retained. The exploration demonstrated, however, that additional combining was necessary to eliminate small numbers. In Table 4 of Appendix Five, the variables for which the number of levels was reduced for the study population are shown both ways.

Percentiles were calculated for the continous variables and used to divide them into four categories so that approximately one fourth of the children in the sample were in each category. An alternative procedure that would probably have been better is that which Cox suggested for normally distributed variables. He suggested that when continuous variables were being grouped into four categories, the percentages should be 16.4, 33.6, 33.6, and 16.4 because the function M is very flat in the neighborhood of its maximum. Thus the precise grouping in that area is not critical.

For those continuous variables where zero was relatively common (no outpatient department or no physicians in the county for example), a dichotomous presence-versus-absence variable was also created. The two-level or four-level variable was selected according to the same criterion as used to select the number of levels for the categorical variables.

This approach of combining levels with similar responses to increase the heterogeneity has implications for the latter analysis. Strictly speaking, hypotheses should be formulated prior to collecting the data and to any analysis. However, the data used in this study had already been collected and the purpose of this secondary analysis was to identify factors associated with childrens' use of medical care; the primary purpose was not the testing of previously formulated

hypotheses. Thus, the strict position was not only impossible, it would have been inappropriate for the purpose. It would have prevented the discovery of new information.

It should also be noted that some of the dangers of ex post facto interpretation were avoided as much as possible by applying the same approach to all variables and by screening all variables for possible inclusion. By using a standard approach the introduction of the analyst's personal biases was minimized as much as possible.

Selecting the Population for Study

The purpose of the analysis was to draw inferences about all children and youths in the United States who were not residents of institutions. The sample had been drawn from this population and the desire was to use the entire sample. However, all the information was not available for all the children.

The final data set contained 73,817 records of which 69,062 were matched with the records of one or both parents. The failure to match the child's record with a parent's record did not necessarily mean that there was no parent in the household. Because the relationship coded in the National Health Interview Survey is the relationship to the head of the household, the relationship could be established and the

record linked only when the parent was either the head or wife of the head of the household. Other evidence on the record indicated that between 1,161 and 2,369 of the 4,755 children whose records were not matched did have a parent in the household (Appendix Two).

The failure to match a record did mean, however, that no information was available about the parents. In particular, no information was available about the mother whose characteristics, since she is the assumed caretaker of the child, would theoretically have the greater influence on whether the child received medical care. For this and other reasons it was judged desirable to restrict the analysis to those children whose records had been matched to those of their mothers.

For practical reasons two further restrictions were also desirable. First, although lack of information about when a child last received medical care could in itself be an indicator of lack of interest in the child and lack of caretaking, it was of no practical use. The information was missing for only 468 of the 68,417 children with matched records for the mothers - a group too small for analysis. Second, it immediately became apparent that the interval since the mother's last medical care was more strongly associated with the interval since the child's last care than any other variable. It was therefore desirable to restrict

the analysis to those children for whom the interval since the mother's last medical care was known. This information was missing for 198 of the children. Of the 68,417 children with matched records for the mothers, information on either the interval since the child's or the mother's last medical care was missing for only 605 children.

Thus, there were three possible populations available for analysis:

1. All children under age 18 (n=73,817)
2. Children under age 18 with matched records for the mother (n=68,417)
3. Children under age 18 with matched records for the mother and with information on the interval since both the child's and the mother's last medical care (n=67,812)

Ninety-three percent of the children were matched with the mothers. The factors associated with the receipt of medical care for these children could be different from those for the total population of children as the children whose records were not matched were more likely to be living with distant relatives or non-relatives, in very large households, or in other circumstances where the individual needs of the child might be less likely to be recognized and cared for.

Of the children whose records were matched with those of the mother, 99 percent had the information on the interval since the last medical care available for <u>both</u> mother and child.

There was no theoretical reason to expect that the factors influencing receipt of medical care for these children would be different from the factors influencing receipt of medical care for all children with mothers.

Because the purpose of the study was to draw inferences about the factors influencing the medical care of all children - not just children for whom information was available - it seemed appropriate to investigate whether the same variables would be selected as the most important if the analysis were restricted to children with matched records or to the subset of children with matched records and with the two specified items of information available as would be if all children were included.

To do so statistics were computed for each of the three populations - first with no control variable, second with a control for the interval since the mother's last contact for medical care (MDMA), and third with controls for the mother's last contact (MDMA) and for family size (FAMSIZE).

A summary of the results is shown in Tables 3.1 - 3.3. More detail and a different arrangement of the data are shown in Tables 1-3 of Appendix Five. The questions used to obtain the information and the levels are given in Appendix Three. The use of acronyms in the tables is unfortunate for readability but was unavoidable because of space limitations.

When comparing data from the three populations, it is necessary to remember that the number of levels is different for some of the variables when people with missing information are excluded. Thus, the number of degrees of freedom used as a divisor will also be different. It is the relative value, not the absolute value, of the statistic in each population that is of interest at this point.

Table 3.1. Relative ranking of variables for all children under age 18.

Variable	No control	MDMA	MDMA by FAMSIZE
	n=73,817	n=73,234	
	X^2/D.F.	Sum 2/D.F.	
MDMA	400	--	--
FAMSIZE	238	99	--
EDMA	167	64	24
EDPA	159	67	25
MDPA	157	52	26
OUTCO	143	69	32
PHONE	140	59	26
SMSA	138	67	31
EDHEAD	118	45	18
HEALTHIN	233	39	17
INCPA	114	50	24
LIMITEDACT	104	47	22
CAPINC	104	49	23
FAMINC	91	32	19
RACE	76	34	13
AGEGROUP	76	29	17
REGION	67	36	22
MDCODE	61	31	15
AGEMA	61	14	11
ADJACENCY	53	26	13
AGEPA	51	16	11
MEDSCH	41	19	8
HEALTH	37	9	6
PCHNEW	37	19	8
HEALTHPA	36	15	9
HEALTHMA	35	10	6
INCMA	33	6	6
OUTAREA	32	19	10
ACTMA	31	2	6
AGEBIRTH	29	5	7
USUALACT	21	6	20
ACTPA	20	8	10
EMERCO	17	12	5
PARENTS	11	2	11
FEMFAM	6	5	3
SEXHEAD	4	5	11
AFDCCODE	2	2	2
SEX	1	4	4

NOTES: Data based on 73,817 children in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling. Definitions of acronyms are given in Appendix Three.

Table 3.2. Relative ranking of variables for children under age 18 with matched records for mothers.

Variable	No Control	MDMA	MDMA by FAMSIZE
	n=68,417	n=67,949	
	X^2/D.F.	Sum 2/D.F.	
MDMA	771	--	--
FAMSIZE	234	96	--
EDMA	158	63	23
EDPA	152	64	24
MDPA	144	50	25
PHONE	138	58	26
SMSA	137	66	30
OUTCO	133	67	31
INCPA	110	49	23
EDHEAD	107	43	17
HEALTHIN	106	38	16
CAPINC	97	47	22
LIMITEDACT	92	41	20
FAMINC	84	32	18
AGEGROUP	78	29	17
RACE	63	34	12
MDCODE	62	31	14
AGEMA	59	18	16
REGION	57	31	19
ADJACENCY	51	25	13
AGEPA	49	15	12
MEDSCH	38	18	8
PCHNEW	35	18	8
HEALTHPA	31	14	5
HEALTH	30	8	7
OUTAREA	29	16	9
EMERCO	21	12	5
AGEBIRTH	19	6	7
ACTMA	19	2	10
USUALACT	18	5	18
HEALTHMA	18	10	7
INCMA	17	5	6
ACTPA	16	6	7
FEMFAM	5	4	2
PARENTS	3	2	12
AFDCCODE	1	2	2
SEX	0	0	2

NOTES: Data based on 68,417 children in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling. Definitions of acronyms are given in Appendix Three.

Table 3.3. Relative ranking of variables for children under age 18 with matched records for mothers and with interval since last medical care known for both mother and child.

Variable	No Control	MDMA	MDMA by FAMSIZE
	n=67,812		
	X^2/D.F.	Sum 2/D.F.	
MDMA	982	--	--
FAMSIZE	460	127	--
EDHEAD	310	89	17
EDPA	296	85	16
EDMA	289	82	15
MDPA	276	66	18
SMSA	274	101	21
OUTCO	272	87	21
PHONE	265	78	18
INCPA	212	65	16
CAPINC	194	62	15
LIMITEDACT	176	54	13
HEALTHIN	173	50	11
AGEGROUP	143	38	11
FAMINC	131	42	13
MDCODE	122	41	9
RACE	116	45	9
OUTAREA	116	44	10
AGEMA	114	24	8
REGION	110	41	6
ADJACENCY	103	33	10
AGEPA	90	20	1
MEDSCH	76	24	13
PCHNEW	70	24	7
EMERCO	64	23	6
HEALTHPA	62	18	4
AGEBIRTH	36	8	6
HEALTH	33	11	3
USUALACT	28	6	2
ACTPA	23	8	4
HEALTHMA	16	13	3
INCMA	17	6	2
FEMFAM	11	5	3
ACTMA	6	2	2
PARENTS	4	3	3
AFDCCODE	1	2	1
SEX	0	0	1

NOTES: Data based on 67,812 children in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling. Definitions of acronyms are given in Appendix Three.

Regardless of which of the three populations was being considered, the relative ranking of the variables was similar - the interval since the mother's last medical care, size of the family in the household, education of the parents, the interval since the father's last medical care, location in an SMSA, whether there was a phone available, and whether there was an outpatient department in the county were important. The relative importance of the remaining variables was similar also. Sex of the child, sex of the household head, whether the child had two, one, or no parents, whether there was an emergency room in the county, proportion of families in the county with women as head of household, or proportion of families on AFDC were all of lesser importance.

Other checks of the data revealed that there were no significant differences in the age of the children for whom the interval since the mother's last medical care was known and those for whom it was not. The chi-square for the unweighted sample assuming simple random sampling was only 1.98 with 2 degrees of freedom. Nor was there a difference in family size. The chi-square for the unweighted sample was 3.23 with 3 degrees of freedom. There were no significant differences between the 137 children for whom the information on the timing of the mother's last visit for medical care was unknown and those for whom it was not except in one respect. That respect is important and led to the decision to select

the third, and most restrictive of the proposed populations, for analysis.

Other information was more likely to be missing when information on the mother's last visit for medical care was missing than it was when the information on the mother was recorded. For example, information on family income was missing for 23 percent of these 137 children in contrast with 7 percent of the other 67,812 children. Information on whether the child was covered by health insurance was missing for 8 percent of the former group and 1 percent of the latter (Table 3.4).

The records which were excluded were records with data obtained from poor interviews. There is no evidence of bias; the children for whom the data are available on the timing of the mother's last medical care are not statistically distinguishable from the children for whom the data are available. There is evidence, based on the proportion of missing information on other items, of poor interviewing or of accepting as a respondent someone who did not know the information.

Table 3.4. Proportion of items with missing information for selected variables among children with mothers according to whether interval since last medical care for child and mother was known.

	Child's Interval Not Known	Child's Interval Known	
		Mother's Interval	
		Not Known	Known
	Percent of records with missing information		
HEALTH	3.85	1.46	0.50
FAMINC	17.74	23.36	6.86
EDHEAD	4.27	5.84	0.88
PHONE	1.71	1.46	0.44
HEALTHIN	2.78	8.03	0.94
EDMA	4.49	6.57	0.70
HEALTHMA	2.35	1.46	0.34
INCMA	8.33	18.25	3.12
ACTMA	1.71	5.84	0.11

NOTES: Data based on 68,417 children with matched records for mothers in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data. Definitions of acronyms are given in Appendix Three.

In summary:

Because the purpose of the study is to draw inferences about the total population of noninstitutionalized children, there are theoretical reasons for using the sample of all children for investigating the factors influencing the child's receipt of medical care rather than restricting the population to those children with records matched to those of mothers and for whom the information on the time of both mother's and child's last visit for medical care is known. The evidence available does not support those hypothetical considerations. The relative ranking of variables which would be selected for

study are similar and the associations of these variables with the dependent variable are also similar.

There is no evidence of bias. There is evidence of poor interviewing, and the lack of information on multiple items obscures relationships. Therefore, the records of children who could not be matched with records of mothers or who did not have the timing of medical care were eliminated. The population used for further analysis was that of children with mothers for whom information on the timing of both the mother's and the child's last medical care was available.

Supplemental Data Available for One Year Only

In addition to information not being available because the child's record could not be matched with that of the mother, there were certain conditions under which questions were not asked for all children. The major reason was due to the use of supplements during single years to collect information that is not routinely collected; some questions were asked only in 1975 while others were asked only in 1976.

The supplement of interest in 1975 was one containing questions on the person's usual source of care. The supplements of interest in 1976 contained questions on the number of siblings each person in the household had and questions on coverage for health insurance under private plans and public

programs. While these were of great analytic interest for the present study, their use in the variable selection process would have resulted in selecting two sets of variables - one set for each year - and would have precluded defining a set based on data from the two years combined. The decision was to obtain the values for these variables in the variable selection process to determine whether they were important but not to include them in the ranking. They were available for inclusion for analyses using a single year of data if that should later prove desirable but the variables for the first (two year) analysis were selected as if these single-year variables did not exist.

The Study Population

The Amercian Academy of Pediatrics recommendations referred to in Chapter One are for the care of well children who receive competent parenting, who have not manifested any important health problems, and who are growing and developing satisfactorily.

The comparison of the populations available for study did not provide any evidence of differences in the receipt of medical care among children with two, one, or no parents living in the household whose records could be matched. That difference, if it had existed, might have been interpreted as evidence of a difference in competent parenting. Other than

that, there is no information in the NHIS data which could have been used to restrict the study population to those who were receiving competent parenting.

There is no information at all which could be used to select those children who were growing and developing satisfactorily.

There is, however, a series of questions about whether a person is limited in activity because of a chronic condition and a question on the respondent's perception of the child's health as compared with other children the same age. These questions could be used to define a population of well children who have not manifested any important health problems.

Children who are limited in activity, that is, children who are unable to participate in the activities usual for their age because of a chronic condition are obviously not well children. Only 3.9 percent of the children living in the community were limited in this way in 1975-76 and they were distributed throughout the population rather uniformly (Kovar 1982). They did, as would be expected, receive more, and more frequent, medical care than other children. These children were excluded from the study population.

Children who were reported to be in "fair or poor" health (as

opposed to "excellent or good" health) were not excluded because this perception of the child's health apparently measures something other than physical health. Investigation revealed that the best predictor of the child's perceived health status was the mother's reporting of her own health - mothers reported their children's health status as being the same as their own. However, while the reported health status of adults is correlated with health behavior, that is, people in poorer health had more disability and medical care use, (Kovar and Wilson 1976), the reported health status of children is not associated with health behavior in all population groups. Children in "fair or poor" health have more short-term limitation or visits for medical care than children in "excellent or good" health in some population subgroups but not in others. The perceived health status shows wide variation among socioeconomic groups and it is not a measure which is invariant across groups. For example, at any given income level a higher proportion of black than of white children were reported as being in "fair or poor" health but the black children were less likely than the white ones to be limited in activity, have days when they were kept in bed or lost time from school because of illness, or to have visits for medical care (Kovar 1982). Discussions with black physicians and others working with black children offered no clues as to what is actually being reported by the mothers except perhaps a feeling of hopelessness. While such a feeling is important, it does not indicate that the child

is not well and is in need of special medical care (it may be an indicator that both mother and child are in need of social support and care).

Thus, the study population chosen to investigate the factors influencing whether children receive care within the interval recommended by the Academy is comprised of children who were living with their mothers and who were not limited in activity by a chronic condition in 1975-76.

There were 65,235 of these children in the 1975-76 National Health Interview Survey. The selection comparable to that for the other two populations is shown in Table 3.5 for comparison with Tables 3.1 - 3.3. This is not the selection that was actually used as the number of levels was reduced in the study population for several of the variables to eliminate categories where the data were too sparse.

Table 3.5. Relative ranking of variables for children under age 18 without limitation of activity, with matched records for mothers, and with the interval since the mother's and child's last visit for medical care (MDMA and MDADE2) known.

Variable	No Control	MDMA	MDMA by FAMSIZE
	n=65,235		
	X^2/D.F.	Sum 2/D.F.	
MDMA	960	--	--
FAMSIZE	453	125	--
EDHEAD	308	89	17
EDPA	295	84	16
EDMA	286	82	15
SMSA	267	85	20
OUTCO	267	86	23
PHONE	256	75	18
MDPA	264	63	17
INCPA	210	64	16
CAPINC	186	60	15
HEALTHIN	178	52	23
AGEGROUP	150	40	12
FAMINC	130	42	13
MDCODE	121	40	10
OUTAREA	115	39	10
RACE	114	44	9
AGEMA	114	24	8
REGION	107	40	13
ADJACENCY	101	32	9
AGEPA	90	20	6
MEDSCH	73	22	6
HEALTHPA	69	20	5
PCHNEW	66	22	6
EMERCO	61	22	6
AGEBIRTH	35	8	4
USUALACT	30	20	2
ACTPA	23	8	4
HEALTHMA	22	14	3
HEALTH	19	6	2
INCMA	18	6	2
FEMFAM	11	5	2
ACTMA	6	2	3
PARENTS	2	3	5
AFDCCODE	1	3	1
SEX	0	0	1

NOTES: Data based on 65,235 children in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling. Definitions of acronyms are given in Appendix Three.

The problem that remained, then, was to select from the 41 (48 if both the father and mother were present) remaining variables those to enter the model to determine which factors influenced whether the child received medical care. The summary of the data is shown in Table 3.6 and the details are shown in Tables 4-11 of Appendix Five.

Table 3.6. Variable selection for children under age 18 without limitation of activity, with matched records for mothers, and with the interval since mother's and child's last medical care known

Variable	X^2 D.F. (1)	X^2 Sum ϱ D.F. (2)		X^2 Sum ϱ D.F. (3)	
MDMA	1688	---		---	
PHONE	487	715	210	---	
FAMSIZE	453	394	169	213	89
EDHEAD	446	409	195	204	77
EDMA	426	404	183	201	72
EDPA	422	398	182	202	74
MDPA	281	448	102	234	54
SMSA	267	644	128	335	49
OUTCO	267	644	127	334	48
INCPA	210	280	97	146	36
CAPINC	186	315	88	170	35
HEALTHIN	178	414	81	214	22
AGEGROUP	150	379	59	240	38
FAMINC	130	251	57	134	22
MDCODE	121	291	59	163	25
OUTAREA	115	290	56	164	24
RACE	114	396	66	238	24
AGEMA	114	270	36	164	28
REGION	107	298	59	168	27
ADJACENCY	101	229	48	130	20
AGEPA	90	212	30	133	29
MEDSCH	73	271	34	163	12
HEALTHPA	69	267	30	155	15
USUALACT	69	352	20	204	5
PCHNEW	66	270	32	155	15
EMERCO	61	587	33	314	13
AGEBIRTH	35	199	12	119	9
HEALTH	26	344	11	200	7
ACTPA	23	347	10	205	11
HEALTHMA	22	330	21	149	5
INCMA	18	238	7	119	5
FEMFAM	11	248	7	147	5
ACTMA	6	245	3	156	3
AFDCCODE	4	244	2	145	2
PARENTS	2	563	1	316	13
SEX	0	563	0	307	1

NOTES: Data based on 65,235 children in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling. The number of levels has been reduced for MDMA, MDPA, PHONE, EDHEAD, EDPA, EDMA, and HEALTH. Definitions of acronyms are given in Appendix Three.

Table 3.6. Variable selection for children under age 18 without limitation of activity, with matched records for mothers, and with the interval since mother's and child's last medical care known (cont)

Variable	Sum Q D.F. (4)	M-H Q (4)	M-H Q D.F. (5)	M-H Q D.F. (6)
MDMA		---	---	---
PHONE		---	---	---
FAMSIZE		---	---	---
EDHEAD	15	217	---	---
EDMA	14	197	42	---
EDPA	15	209	5	---
MDPA	14	176	151	---
SMSA	13	161	99	105
OUTCO	11	166	112	107
INCPA	10	124	48	41
CAPINC	8	113	72	69
HEALTHIN	5	53	12	---
AGEGROUP	10	69	56	49
FAMINC	8	83	21	---
MDCODE	7	81	58	51
OUTAREA	7	88	81	70
RACE	6	52	19	---
AGEMA	7	56	40	---
REGION	9	105	103	90
ADJACENCY	7	66	44	---
AGEPA	5	34	---	---
MEDSCH	3	34	---	---
HEALTHPA	4	21	---	---
USUALACT	3	3	---	---
PCHNEW	4	45	---	---
EMERCO	4	30	---	---
AGEBIRTH	3	20	---	---
HEALTH	3	30	---	---
ACTPA	4	6	---	---
HEALTHMA	2	12	---	---
INCMA	2	6	---	---
FEMFAM	2	14	---	---
ACTMA	2	15	---	---
AFDCCODE	1	2	---	---
PARENTS	4	3	---	---
SEX	1	0	---	---

NOTES: Data based on 65,235 children in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling. The number of levels has been reduced for MDMA, MDPA, PHONE, EDHEAD, EDPA, EDMA, and HEALTH. Definitions of acronyms are given in Appendix Three.

For the study population (children and youths under the age of 18 who were not limited in activity by a chronic condition), the uncorrected chi-square divided by its degrees of freedom was 1688 for the association between the interval since the mother's last contact for medical care (MDMA) and whether the child had received care within the interval deemed adequate for routine preventive care (MDADEQ). This was more than three times the value for any other variable. Thus, it was the first variable selected for inclusion.

When the interval since the mother's last contact for care was entered, the presence or absence of a telephone (PHONE) had the highest chi-square divided by the degrees of freedom regardless of whether the Higgins-Koch or the Pearson statistic was used.

When both of these variables had been entered, the third variable selected was family size if the Pearson chi-square summed over the tables was used as the selection statistic; it would have been residence in a Standard Metropolitan Statistical Area (SMSA) if the Higgins-Koch statistic had been used. Although the decision to use the Pearson statistic had been made before the data were analyzed, it is reassuring to note that family size also had a higher ranking in the univariate list.

With three variables entered there were 16 tables in the next level of variable selection and the data were becoming thin even with this large sample. There are relatively few children whose mothers have not had a medical contact within a year, who live in households without telephones, or who live in families with six or more members - the numbers are sparse in any of these categories. In addition, these variables are correlated; children in families where the mother has not seen a doctor are more likely than other children to be in large families or households where there is no phone. Because of the sparseness, the switch to the Mantel-Haenszel statistic was made at this point.

The Mantel-Haenszel statistic divided by its degrees of freedom was 217 for education of the family head (EDHEAD) and so that was the next variable selected. It would also have been selected if the Pearson chi-square had been used. In this case, the choice of the selection statistic made no difference.

With four variables already selected, the data were extremely sparse - not only cells but entire tables had no data. There was no point in even computing the Pearson statistic; only the Mantel-Haenszel was computed for the remainder of the variable selection. Further, statistics were not computed for those variables which had very low Q values. Limitations on computer programs meant that no more than five or six

variables could be used in the final model. To further screen variables that had no chance of being used in the model would have been a waste of computer time.

The fifth variable chosen was the interval since the last medical contact of the father (MDPA). With a chi-square of 302.54 with two degrees of freedom, it was significant when divided by Choi's correction factor.

The sixth, and final, variable selected was the one indicating whether or not there was an outpatient department in the county of residence (OUTCO). The Mantel-Haenszel test statistic was 106.65 which was significant with one degree of freedom assuming simple random sampling. It was still significant when divided by Choi's correction factor.

Variable selection was terminated with six variables although more variables would have been selected before termination ceased if lack of significance had been used to terminate selection. There were two reasons for stopping at this point. First, the computer software currently available cannot handle a model this large; it certainly could not handle a larger one. Second, the computer time required to run the selection program increases with the number of tables and it had become too expensive to continue - especially as any additional variables selected could not be used in the model.

The selection was made on the basis of preordained criteria. Other criteria could have been used and other variables would have been chosen. Some non-statistical comments may be of interest here.

The selection of the interval since the mother last saw a doctor as the first variable seems to uphold the social scientists' belief that there is a generational transfer of behavior and the medical belief that if the mother receives regular medical care the children will also. Both may be true but it is also possible that there is some reporting of normative behavior. The mother may be reporting what she thinks she and the child should have done as if it were real behavior. In these same data with the same selection procedure, the first variable to be selected for the child's perceived health status is the mother's perceived status - she sees her child as she sees herself. In the self-respondent study conducted by the NCHS in 1972, one of the big differences between self- and proxy- respondents was that young adults (especially males) living at home would be estimated to have fewer dental visits if they were self-respondents than if proxies reported for them. Charles Cannell suggested that respondents were reporting what their grown children should be doing while the young adults reported what they did do. In both cases the proxy respondent was usually the mother (NCHS unpublished).

The second variable was the presence or absence of a telephone. This information is collected on the NHIS in case the interviewer has to telephone for information that she forgot or was unable to collect while she was in the household. It is not collected for analytic purposes and, as far as I know, has never been used in this way before. This serendipitous discovery would not have been made if the study had been restricted to hypotheses formulated before the analysis or those based on published data from this survey.

The third variable, family size, is one that is often found to be associated with medical care or lack of it (Anderson 1968). It would be useful to investigate a more refined measure, to examine the components of the family, to try to determine whether it is the size of the family per se, the spacing of the children, or the distribution of family members that's important. For example, the number of children per adult or the number of adults of working age might be determining factors rather than simply the number of people in the family.

Education of the family head has been used as a control variable in many of the NHIS publications. The author felt that education of the mother would be preferable for conditions affecting children. Certainly in the family planning literature the woman's education is more important than the

man's. In this analysis, however, the education of the family head consistently ranked higher than the education of the mother. It is possible that a combination of the mother's and father's education, using the higher of the two for example, would have higher predictive power but that has not been investigated.

It is interesting that the interval since the father's last medical care is selected among the most important variables even when the interval since the mother's last care is controlled for. There are a number of explanations possible but none were investigated and trying to give a reason would be pure speculation.

Throughout the selection process it was a question of whether residence or non-residence in a Standard Metropolitan Statistical Area or presence or absence of an outpatient department in the county would be selected. They are highly associated. A cross-tabulation revealed that the child was more likely to have received the recommended care if the child lived in a county that had an outpatient department and was in an SMSA than any of the other three possibilities. Also, the ratio of the area of the county to the number of outpatient departments in the county was a significant variable which was eligible for selection. It is possible that the presence of an outpatient department is important when the distance is not too great, but to investigate such a

hypothesis well would take measures of distance and transportation that are not available from these data. Further investigation using a variable constructed from all the variables in these data would also be warranted.

It is noteworthy that neither race nor family income were among the variables selected. Both are universally assumed to be among the most important factors influencing whether children receive medical care. On the univariate level (Appendix Five Table 4), race was relatively important but it declined in the relative rankings once the first variables had been included. It is likely that its high univariate rank is due to the high correlation of race with other important variables.

It is likely that the availability of Medicaid to low-income families is responsible for the relative unimportance of family income. Also, a cross-classification of education by family income revealed that children in families with low educational levels were unlikely to receive care regardless of the family income, while the income of the family made no difference once the educational level was taken into account. This finding is of substantive analytical importance as education and income are frequently treated interchangeably as "measures of socioeconomic status". They are not measures of the same thing. Although highly correlated, they have independent effects. It is also of programmatic importance

because so many public programs are directed towards low-income families without considering the importance of the mother's education to her ability to care for her child. Education is recorded on the birth certificates in most states. The information is available, at least for the newborn, if providers of public services would take advantage of it.

References

Anderson,R. A Behavioral Model of Families Use of Health Services. Research Series No 25. Center for Health Administration Studies,University of Chicago,1968.

Choi,J.W. and R.B.McHugh. The Analysis of Categorical Data in Weighted Cluster Sample Surveys. Proc. Soc.Stat.Sec. of the American Statistical Association 1980.

Choi,J.W. Analysis of Categorical Data in Weighted Cluster Sample Surveys. Submitted to JASA.

Choi,J.W. A Further Study on the Analysis of Categorical Data from Weighted Cluster Sample Survey. Proc.Soc.Stat.Sec. of the American Statistical Association 1981.

Cochran,W.G. The X^2 Test of Goodness of Fit. Annals of Math.Stat. (23)3, 315-345, 1952.

Cochran,W.G. Some Methods for Strengthening the Common X^2 Tests. Biometrics (10), 417-451,December 1954.

Cox,D.R. Note on Grouping. JASA (52)280, 543-547,Dec., 1957.

Freeman,D.H. and J.F.Jekel. Table Selection and Log-linear Models. J.Chronic Dis. (33), 513-524, 1980.

Grossman,M., D.Coate, L.N.Edwards, R.A. Shakotko, D.Chernichovsky. Determinants of Children's Health. NCHSR Research Summary Series. DHHS Publication Number (PHS) 81-3309. Hyattsville, Md. July 1981.

Hidiroglou,M.A. and J.N.K.Rao. Chisquare Tests for the Analysis of Categorical Data from the Canada Health Survey. Presented at the International Statistical Institute, Buenos Aires,Argentina, December 1981.

Higgins,J.E. and G.G.Koch. Variable Selection and Generalized Chi-Square Analysis of Catagorical Data applied to a Large Cross-sectional Occupational Health Survey. Int.Stat.Rev. (45), 51-62,1977.

Hoffman,L.W. Social Change and its Effects on Parents and Children. Presented at Women: A Developmental Perspective, National Institutes of Health, Bethesda, Md. November 1980.

Koch,G.G., The Interface between Statistical Method-ology and Statistical Practice. Proc.Soc.Stat. Sec. of the American Statistical Association 1977.

Kovar.M.G. Children's Health Status and Use of Medical Care. Public Health Reports Jan-Feb,1982.

Kovar,M.G. and R.W.Wilson. Perceived Health Status - How Good is Proxy Reporting? Proc.Soc.Stat.Sec. of the American Statistical Association 1976.

Lilienfeld,A. Some Aspects of Cancer Epidemiology. Current Topics in Biostatistics and Epidemiology: A Memorial Symposium in Honor of Jerome Cornfield. Biometrics 38(Suppl.) March 1982.

Mantel,N.and W.Haenszel. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute 22(4):719-748,April 1959.

Rao,J.N.K. and A.J.Scott. The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. JASA 76(374): 221-230,June 1981.

Rosenzweig,M.R. and T.P.Shultz. Birthweight, the Production of Child Health, and Input Demand. Presented at a conference "Economic Aspects of Health" sponsored by the National Bureau of Economic Research, 1980.

CHAPTER FOUR

The Analysis

Categorical Data Analysis

The approach to categorical data analysis proposed by Bhapkar and Koch (1968) provides the framework for the analysis presented in this chapter. In particular, the data used for this analysis can be considered as coming from a multi-factor, one-response experimental design in which the term "factor" denotes a classification describing some aspect of the population to which the individual belongs and the term "response' denotes a classification describing some aspect of what happens to him. The marginal frequencies that depend only on the factor combinations are fixed by sample design or by a conditional distribution argument; all other frequencies are random variables.

In this framework, the independent variables are cross-classified into a set of mutually exclusive subclasses or subpopulations. Similarly, the dependent variables are cross-classified to determine the levels of the (potentially) multivariate response profiles. Thus, the entire distribu-

tion) are estimated within each subpopulation.

Consider a finite population of size N from which a probability sample is to be selected. Suppose that this population is divided into "s" subpopulations, and that individuals in the subpopulations can be classified into one and only one response category. Let Nij denote the number of individuals in the population in the i-th subpopulation classified into the j-th category of the response. Let Ni. denote the total number of individuals in the i-th subpopulation. Analysis of categorical data from this finite population is concerned with the proportions Pij = Nij/Ni., that is, the proportion of the individuals in the i-th subpopulation who are in the j-th response category.

Suppose that a sample of size n is selected from the population and ni. individuals are in the i-th subpopulation. Let nij denote the number of sample individuals who are in the j-th category of the response and in the i-th subpopulation. The sample proportion will then be pij = nij/ni..

The contingency table from the sample is shown in Table 4.1.

Table 4.1. Contingency Table from a Sample Survey.

Subpopulations	Response Categories				
	1	2	. . .	r	Total
1	n11	n12	. . .	n1r	n1.
2	n21	n22	. . .	n2r	n2.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
s	ns1	ns2	. . .	nsr	ns.

Corresponding to the frequency table is a table of probabilities where pij is the probability that an individual in the i-th subpopulation has the j-th response (Table 4.2). The probabilites in each subpopulation sum to one. This results from viewing each level of the factor combinations as a distinct subpopulation with the ni. fixed in advance. Thus, the probabilities are calculated within each subpopulation. The pij are estimates of the true probabilites with pij=nij/ni. .

Table 4.2. Probability Table from a Sample Survey.

Subpopulations	Response Categories				
	1	2	. . .	r	Total
1	p11	p12	. . .	p1r	1
2	p21	p22	. . .	p2r	1
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
s	ps1	ps2	. . .	psr	1

Suppose that a complex sample design has been used to select the sample of size n. The population has been divided into

elements which are classified into strata; within each stratum there are clusters. From each stratum a sample of clusters has been selected. This is, in essence, the sample design of the National Health Interview Survey.

In a sample such as that used for the NHIS, the row totals, except for the 60 age-sex-color classes used for adjustment to the U.S. population, are not fixed in advance. They are random variables. They had been fixed, however, before this study was begun because this is a secondary analysis of a completed survey.

Following this argument, the frequency counts are also fixed constants in this study because the data were obtained from a completed survey. Koch has suggested that, assuming that the cell frequencies follow a product multinomial distribution, the analytic framework is appropriate for data derived from a sample survey even when the survey has a complex design although the selection in that case is indeterminate. That is, the stratification variables are not the factors used for analysis. The stratification variables for the NHIS that are given in Appendix One are generally not variables of analytic interest. They were selected to be correlated with health variables, but are obviously more likely to be correlated with employment than with health. In any case, the theoretical basis is certainly not as firm when an estimate is derived from summing over strata which are not the domains of

interest. Nevertheless, the framework is useful provided a consistent estimator of the variance-covariance matrix is available (Koch et al 1975, Freeman et al 1976).

Given a consistent estimate of the variance-covariance matrix, the Grizzle-Starmer-Koch (1969) linearized model can be used to analyze the data from this survey.

The vector of proportions, together with their variances and covariances were computed within SAS using SURREGR. Any other computing algorithm designed to compute a vector and a consistent estimate of its covariance structure could have been used instead. The vector of sample proportions and its covariance matrix were then entered directly into the weighted least squares program. Fortunately, a computer program to implement this approach, GENCAT, is available (Landis et al 1976).

Using this approach, the goodness of fit of models can be evaluated using the Wald statistic

$Q = (F-Xb)'V^{-1}(F-Xb)$.

When the model fits and the data are derived from a simple random sample, Q asymptotically follows the chi-square distribution. Even when the covariance matrix is estimated from a sample with a complex design, the same result should hold for large samples (see, for example, Shuster and Dowling

1976).

The weighted least squares estimator for the parameter vector, B, can be obtained from the survey data as $b = (X'V^{-1}X)^{-1}X'V^{-1}F$.

When the fit of the model is adequate by the Wald statistic criteria, it is possible to examine the individual parameters to reduce the model. This test statistic is also asymptotically distributed as a chi-square random variable when the null hypothesis is true. The predicted values will adequately describe the statistically important sources of variation in the data and will have smaller sampling errors than the observed estimates. The smaller sampling errors result from using the entire sample to esimate the variances for the predicted funtions while the errors for the observed values are estimated from the sample in each subclass.

This approach to data analysis has been discussed in a number of papers and is the subject of a monograph by Landis et al published by the National Center for Health Statistics (NCHS 1982). However, the complete approach has not been applied before to a survey as large as the National Health Interview Survey - large both in terms of number of people in the sample and number of primary sampling units.

Variances and Covariances

Obtaining a consistent variance-covariance matrix from a survey with a complex sample design is not always simple. The development of computer programs to produce estimates of the variance-covariance matrixes from complex survey designs has been slow and they are still not in wide use. There are three approaches that have been developed of which only two, the balanced half-sample replication and the Taylor series approximation, have software packages that are available for analytic use. The BRR program of the National Center for Health Statistics (NCHS 1977) and the OSIRIS package of the University of Michigan (Survey Research Center Support Group 1979) depend on a balanced half-sample replication method. The SURREGR program of the Research Triangle Institute (Holt 1977, Shah 1982) and the SUPER CARP program of the Iowa State University (1980) depend on a Taylor Series approximation. If the sample size is large, they all give approximately the same estimates of the variances and covariances (Maurer et al 1978, Freeman et al 1976, Choi 1981).

It should be noted that, if the sample size is small or if the sample size in many strata or primary sampling units is zero, the variance programs will not give the same estimates. The estimates produced by the different programs may vary by a factor of two or more (Lemeshow undated). Before beginning the analysis, the analyst should examine the distribution of the sample by PSU's and, if necessary, create different

pairings of the PSU's. For example, if many pairs have observations in only one of the PSU's, PSU's with sample observations could be paired with one another. The method of pairing apparently has little influence on the variance estimates in the National Health and Nutrition Examination Survey (Stanek and Lemeshow 1978). Whether this holds for other surveys has not been examined.

There were 75 or more children under age 18 in each PSU of the 1975-76 NHIS. The sample was large enough for any of the available programs to be used.

The NCHS BRR program is the one used to produce the variance curves in the Series 10 publications. It was not used for this study for two reasons. First, the balanced half-sample replications make the tape very long and increase running time on the computer. Second, although the program is available to users outside NCHS, it is not documented for public use and cannot be used without advice and help from NCHS. It seemed preferable to use a program that was commercially available so that users of the NCHS public-use tapes could profit by the examples in this study. SURREGR was used for this study because it runs under SAS and the data tape had been converted to a SAS file.

The need for incorporating the complex survey design in the analysis is apparent when one considers the design effects

which were shown in table 1.1. The overall design effect for the response variable used in this study was 2.76. Any approach that failed to take that design effect into account would obviously produce test statistics that were not appropriate for this survey. Moreover, the design effects for the control variables included here varied from 1.12 through 3.41. An approach that failed to incorporate that variability, such as using a constant design effect to adjust the test statistics, would also be inappropriate.

Although the effect of using only the estimated variances and assuming that the covariances are zero was not investigated as part of this study, it has been by others (see, for example, Freeman and Brock 1978, Jordan et al 1982). The data are too extensive to show here, but the covariance terms were rarely zero, and they were negative in about one-third of the elements in all of the variance-covariance matrixes calculated for this study. Such deviations from the assumption of zero covariances will obviously affect the results to some extent. Therefore, the full matrix should be used, and it was used for this study.

Models

The univariate table for each population variable, or each subdomain, to be included in the analysis is shown in table 4.3. Ideally the entire table would be included in one

categorical data analysis. Such an overall analysis was precluded for two reasons.

First, the proportion of children in households without telephones (8.4 percent) was so small and the other characteristics of the families of these children were so different from those with telephones, that the two groups could not be included in the same analysis (tables 4.4 and 4.5). For example, only 2 percent of the children in households without phones compared with 21 percent of those in households with phones had household heads with 16 or more years of education. If education were crossed with any other variable, there would be empty cells for the children without phones. Many of the other cells in a cross-classification table would be empty also.

Second, both software and hardware limitations precluded such an approach. The full table would have a vector of 480 estimates (2 by 2 by 4 by 5 by 3 by 2) and one triangle of the variance-covariance matrix would contain 115,440 ((480 by 481) /2) elements. Neither the SURREGR nor the GENCAT program can handle such large tables.

The software problems were partially dealt with by modifications in the programs.

The SURREGR program was modified by Babu Shah to produce a

machine-readable output variance-covariance matrix of individual records for each element of the variance-covariance matrix. This eliminated the size restrictions in the 1977 version of the program that precluded reading a large matrix onto a disk file (Holt 1977, Shah 1982).

A version of the GENCAT program was modified by Ron Forthofer to accept an r by s table of 150 cells to surmount the limitation of 80 cells on the original. He also modified the program to accept the lower triangular matrix produced by SURREGR. This modification avoided the necessity of writing a separate program to use the SURREGR output as Diane Makuc had done (1981). The modified program was named LARGEN by Dr. Forthofer, and it will be referred to as LARGEN in the remainder of this paper. (After the SURREGR program was modified in 1982, the modification to allow LARGEN to accept the lower triangle of the matrix was no longer required as the new version of SURREGR produced the full matrix in the output file and all rows were read into LARGEN.)

These modifications made it possible to calculate the matrix and complete the categorical analysis in a single pass through the computer. They also made it possible to read a large matrix onto a disk file and use it for reanalysis of the data. The consequent saving in computer time is enormous.

Table 4.3. National estimates for the children in the study and the percentage with adequate medical care according to specified characteristics, United States 1975-76.

Characteristic	Number in Thousands	Percent of Population	Percent Yes MDADEQ
MDMA			
Yes	47,546	83.5	88.0
No	9,408	16.5	72.9
PHONE			
Yes	51,083	89.7	86.6
No	5,870	10.3	76.2
FAMSIZE			
3 or Fewer	9,477	16.6	91.0
4	16,861	29.6	89.6
5	13,578	23.8	85.7
6 or More	17,037	29.9	78.2
EDHEAD			
8 or Less	8,601	15.1	75.3
9-11	9,838	17.3	81.7
12	20,627	36.2	86.4
13-15	8,115	14.3	90.1
16 or More	9,773	17.2	92.6
MDPA			
Yes	32,389	56.9	87.9
No	15,280	26.8	80.3
Absent	9,285	16.3	86.0
Total	56,954	100.0	85.5

Table 4.4. National estimates for children with phones and the percentage with adequate medical care according to specified characteristics, United States 1975-76.

Characteristic	Number in Thousands	Percent of Population	Percent Yes MDADEQ
MDMA			
Yes	42,913	84.0	88.8
No	8,170	16.0	74.9
FAMSIZE			
3 or Fewer	8,313	16.3	91.5
4	15,541	30.4	90.4
5	12,450	24.4	86.4
6 or More	14,780	28.9	80.0
EDHEAD			
8 or Less	6,407	12.5	77.3
9-11	7,847	15.4	82.7
12	19,256	37.7	86.7
13-15	7,882	15.4	90.3
16 or More	9,692	19.0	92.7
MDPA			
Yes	29,943	58.6	88.7
No	13,706	26.8	81.8
Absent	7,435	14.6	87.0
Total	51,084	100.0	86.6

Table 4.5. National estimates for children without phones and the percentage with adequate medical care according to specified characteristics, United States 1975-76.

Characteristic	Number in Thousands	Percent of Population	Percent Yes MDADEQ
MDMA			
Yes	4,632	78.9	80.7
No	1,238	21.1	59.2
FAMSIZE			
3 or Fewer	1,164	19.8	87.7
4	1,320	22.5	80.3
5	1,128	19.2	78.1
6 or More	2,258	38.5	66.8
EDHEAD			
8 or Less	2,194	37.4	69.5
9-11	1,991	33.9	78.1
12	1,371	23.4	82.0
13-15	233	4.0	84.9
16 or More	81	1.4	86.3
MDPA			
Yes	2,446	41.7	78.0
No	1,574	26.8	66.8
Absent	1,850	31.5	81.8
Total	5,870	100.0	76.2

Variable selection had been terminated before all of the statistically significant variables had been selected. Even though those selected had been chosen using the simple random sampling assumption, it was likely that they - and more - would be significant even when the complex sample design was incorporated in the analysis. Therefore, all of these variables were wanted for the final analysis.

The software limitations made retention of all variables in a single table impossible. The characteristics of the children

in households without phones suggested that the data about them should be examined separately. However, it was still necessary to split the table. Freeman et al (1977) had proposed a method of modularizing tables using the Q statistic as the basis for dividing large tables into smaller ones. Although their suggestion was not used, the logic used for this study was similar. The data were divided into three tables on the basis of the two variables with the highest Q values after testing for interaction between the two variables.

As shown in the last column of table 4.6, the largest table for the total study population that could be analyzed within the limitations in SURREGR before the 1982 modification, the interaction between MDMA and PHONE in the logit model using the variance-covariance matrix from the complex sample design was not significant. With one degree of freedom, the probability of a chi-square this large or larger was 0.355. Therefore, the four population groups shown in table 4.7 could be considered as separate tables and analyzed as distinct populations.

Three tables were used in the analysis rather than four. Data on all children without telephones were in one table as there was such a small number of these children. Two tables were created for children with telephones - one for children whose mother had made a visit within a year and one for those

whose mothers had not.

Finally, the last variable selected, OUTCO, was excluded from the analysis because it could not be handled within the software limitations. (SURREGR was modified in 1982, and the modified program was actually used after this decision was made. See Appendix Six). The inclusion of this variable would also have resulted in empty cells in the tables.

The logit was used for the analysis rather than the probability because the probability of a 'yes' to adequate medical care was 0.8566 overall and was above 0.9 in some of the subdomains (Tables 4.4 and 4.5). This is in the range where the logit is considered preferable so that predicted probability values from model fitting are restricted to remain in the (0,1) interval.

The effect of using the full variance-covariance matrix (and of using a logit rather than a proportion) on the values of the chi-square statistics is also shown in table 4.6. Except for the error term, which is the goodness of fit in this model, the chi-square statistic is smaller when the variance-covariance matrix from the complex survey design is used rather than the variance assuming simple random sampling. This is consistent with findings from other surveys (Cohen and Kalsbeek 1978, Freeman and Brock 1978, NCHS 1982).

Table 4.6. Comparison of chi-square statistics from simple-random and complex survey designs and of proportion and logit models.

	D.F.	Proportion SRS	Proportion Complex	Logit SRS	Logit Complex
Error	4	3.63	3.78	1.51	0.79
MDMA	1	428.45	276.37	674.41	498.61
PHONE	1	181.40	87.67	235.64	122.10
FAMSIZE	3	340.93	187.86	401.57	232.79
MDMAxPHONE	1	16.87	12.17	0.78	0.99
MDMAxFAMSIZE	3	6.10	3.96	25.31	18.53
PHONExFAMSIZE	3	36.11	17.62	13.07	9.34

Table 4.7. National estimates for children in the study and the percentage with adequate medical care according to specified characteristics, United States 1975-76.

Characteristic		Number in Thousands	Percent of Population	Percent Yes
MDMA	PHONE			
Yes	Yes	42,913	75.3	88.8
No	Yes	8,170	14.3	74.9
Yes	No	4,632	8.1	80.7
No	No	1,238	2.2	59.2

All analyses in the remainder of this paper were carried out using the variance-covariance matrix from the complex sample design as input to the weighted least-squares procedure. The response variable is the logit.

Analysis

Before beginning the analysis of the three individual tables, it is worth looking at the total study population. The test

statistics were given in table 4.6 for comparison with the statistics based on a simple random sampling assumption and a probability rather than a logit to show the effects of the four approaches. More detailed information is in the following tables.

The non-significant interaction terms were not eliminated from this model because the purpose was to describe the data in full rather than to produce the most parsimonious description. The test statistics for the factorial model are given in table 4.8. The estimated parameters, their standard deviations, and the odds ratios are in table 4.9. The observed and fitted values of the estimates and their standard deviations are in table 4.10.

The odds ratios give the odds that the child will have seen a doctor within the specified interval. For a specified subpopulation, the odds are obtained by multiplication. For example, the overall odds that a child will have seen a doctor are approximately four to one (3.89). However, if the child's mother has seen a doctor within a year, the odds are (3.89 x 1.66) to one that the child has.

In view of the results from the analysis of the three separate tables that is given later, it is interesting to note that the contrast between FAMSIZE 1 and 2 is significant; the contrast between FAMSIZE 3 and 4 is not.

The b vector for table 4.9 is:

Overall mean
Differential effect of mother's seeing the doctor (MDMA 1)
Differential effect of telephone (PHONE 1)
Differential effect of 0-3 persons in family (FAMSIZE 1)
Differential effect of 4 persons in family (FAMSIZE 2)
Differential effect of 5 persons in family (FAMSIZE 3)
Interaction of MDMA and PHONE
Interaction of MDMA and FAMSIZE (3 terms)
Interaction of PHONE and FAMSIZE (3 terms)

Table 4.8. Test statistics for children in the study.

	X^2	D.F.	p
Goodness of Fit	0.79	3	0.852
Main Effects			
MDMA	498.61	1	<0.001
PHONE	122.10	1	<0.001
FAMSIZE	232.79	3	<0.001
Interactions			
MDMA*PHONE	0.99	1	0.321
MDMA*FAMSIZE	18.53	3	<0.001
PHONE*FAMSIZE	9.34	3	0.025
Contrasts			
FAMSIZE 1 vs 2	15.22	1	<0.001
FAMSIZE 2 vs 3	10.58	1	0.001
FAMSIZE 3 vs 4	2.25	1	0.134

Table 4.9. Estimated parameters for children in the study.

	Beta Values	Standard Deviations	Odds Ratios
b1	1.359	0.028	3.890
b2	0.507	0.023	1.661
b3	0.311	0.028	1.364
b4	0.426	0.047	1.531
b5	0.154	0.037	1.167
b6	-0.060	0.040	0.941
b7	-0.024	0.025	0.976
b8	0.085	0.028	1.089
b9	0.019	0.029	1.020
b10	-0.016	0.027	0.984
b11	-0.098	0.043	0.906
b12	0.094	0.037	1.099
b13	-0.029	0.042	0.972

Table 4.10. Observed and fitted proportions and standard errors for children in the study, United States 1975-76.

SUBDOMAIN	PHONE	MDMA	FAMSIZE	Observed		Fitted	
				Proportion	Standard Deviation	Proportion	Standard Deviation
1	1	1	1	.929	.00269	.929	.00264
2	1	1	2	.918	.00263	.918	.00261
3	1	1	3	.886	.00322	.886	.00316
4	1	1	4	.829	.00542	.829	.00537
5	1	2	1	.900	.01097	.899	.00910
6	1	2	2	.843	.01127	.840	.01029
7	1	2	3	.822	.01566	.822	.01479
8	1	2	4	.715	.01655	.719	.01536
9	2	1	1	.806	.01123	.807	.01036
10	2	1	2	.807	.01289	.805	.01005
11	2	1	3	.751	.01176	.753	.01069
12	2	1	4	.687	.01174	.687	.01160
13	2	2	1	.721	.03545	.722	.02604
14	2	2	2	.622	.02978	.636	.02308
15	2	2	3	.622	.04110	.623	.03001
16	2	2	4	.532	.03414	.513	.02587

Children in households with phones whose mothers had made a visit for medical care within a year were the largest group of children. They were also the most likely to have had a contact with a doctor themselves within the interval deemed necessary for minimal preventive care. There were 43 million such children in 1975-76 comprising three-quarters of the children in the study population. About nine out of ten of them (88.8 percent) had had adequate medical care within the definitions of this study.

Even among these children, however, there was a great deal of variation among population subgroups in the proportion who had received care. Those most likely to have received care were children in small families with high levels of education. Those least likely were in large families with low levels of education (Table 4.13, Column 1).

A factorial model was first considered for analyzing these data. However, two of the two-way interactions (FAMSIZE by MDPA and EDHEAD by MDPA) were not statistically significant while the three-way interaction was highly significant; the chi-square was 51.45 with 24 degrees of freedom ($p = 0.001$). This is the model shown as Example One in Appendix Six.

Therefore, a model with EDHEAD and MDPA nested in FAMSIZE was used. In the saturated model, there were four within-FAMSIZE

interactions, only one of which was significant at the 0.05 level. The statistics for the four interactions were as follows;

FAMSIZE	X^2	D.F.	p
1	14.51	8	0.069
2	14.02	8	0.081
3	20.94	8	0.007
4	3.55	8	0.895

Working from the saturated model then, the model was reduced by successively removing those interaction terms with the lowest chi-square values. With no interaction in FAMSIZE 4, the statistics for the three interaction terms were as follows;

FAMSIZE	X^2	D.F.	p
1	15.03	8	0.059
2	16.15	8	0.040
3	23.45	8	0.003

and the goodness of fit statistic was 3.55 with 8 degrees of freedom (p= 0.895). It seemed possible to reduce the model further.

With no interaction in FAMSIZE 1 or 4, the statistics for the remaining interactions were as follows;

FAMSIZE	X^2	D.F.	p
2	20.25	8	0.009
3	25.73	8	0.001

and the goodness of fit statistic was 18.58 with 16 degrees of freedom (p= 0.291). Both interaction terms were significant and the model provided a satisfactory fit.

All of the main effects within each family size were also significant. However, the difference between FAMSIZE 1 and FAMSIZE 2 was not significant. It might have been possible to reduce the model further by combining these two categories except that there was a significant interaction in FAMSIZE 2 but not in FAMSIZE 1. There were other possibilities also. The interaction could have been retained in both FAMSIZE 1 and FAMSIZE 2 and the two categories combined.

It can be argued that, with a sample size this large and the very large chi-sqare statistics that result, relaxing the level at which a statistic is judged to be statistically significant would be appropriate. There is no magic in the 0.05 level. However, a main-effects model really did not fit. The goodness of fit chi square was 92.57 with 32 degrees of freedom (p= <0.001). Even the model with only the interaction in FAMSIZE 3 retained did not fit. For that model the goodness of fit chi-square was 38.83 with 24 degrees of freedom (p= 0.028). I preferred to retain the information on the interactions in two of the family sizes.

Therefore, the model with two interaction terms was accepted as the final model. The program for this model is shown as

Example Two in Appendix Six. The contrast matrixes are given in the program. The test statistics are in table 4.11.

The estimated parameters, their standard deviations, and the odds ratios are shown in table 4.12. The observed and fitted values of the estimates and their standard deviations are in table 4.13.

The b vector for table 4.12 and for table 4.15 is:

Overall mean
Differential effect of education 0-8 years (EDHEAD 1)
Differential effect of education 9-11 years (EDHEAD 2)
Differential effect of education 12 years (EDHEAD 3)
Differential effect of education 11-15 years (EDHEAD 4)
Differential effect of father's seeing doctor (MDPA 1)
Differential effect of father's not seeing doctor (MDPA 2)
Interaction of EDHEAD and MDPA (8 terms)

in families with three or fewer, four, five, and six or more members (FAMSIZE 1, 2, 3, and 4).

Table 4.11. Test statistics for final model for children in families with phones whose mothers had a visit for medical care within a year.

	X^2	D.F.	p
Goodness of Fit	18.58	16	0.291
Main Effects			
EDHEAD	391.27	16	<0.001
Within FAMSIZE			
1	101.76	4	<0.001
2	64.32	4	<0.001
3	42.47	4	<0.001
4	126.57	4	<0.001
MDPA	210.57	8	<0.001
Within FAMSIZE			
1	22.26	2	<0.001
2	57.37	2	<0.001
3	63.22	2	<0.001
4	48.09	2	<0.001
Interactions			
Within FAMSIZE			
2	20.25	8	0.009
3	25.73	8	0.001
Contrasts			
FAMSIZE 1 vs 2	0.65	1	0.419
2 vs 3	51.45	1	<0.001
3 vs 4	20.11	1	<0.001

Table 4.12. Estimated parameters for children in families with phones whose mothers had a visit for medical care within a year.

	FAMSIZE			
i =	1	2	3	4
	Beta Values			
Bi1	2.499	2.440	1.928	1.617
Bi2	-0.583	-0.571	-0.465	-0.534
Bi3	-0.360	-0.506	-0.274	-0.197
Bi4	0.094	-0.224	-0.015	-0.014
Bi5	0.233	0.142	0.427	0.367
Bi6	0.203	0.208	0.345	0.164
Bi7	-0.221	-0.417	-0.258	-0.270
Bi8		-0.118	-0.166	
Bi9		0.207	0.001	
Bi10		0.066	-0.117	
Bi11		0.293	-0.049	
Bi12		0.097	0.042	
Bi13		0.239	0.243	
Bi14		0.095	0.067	
Bi15		-0.052	-0.515	
	Standard Deviations of Beta Values			
Bi1	0.045	0.056	0.047	0.048
Bi2	0.096	0.097	0.098	0.056
Bi3	0.074	0.096	0.084	0.071
Bi4	0.073	0.065	0.063	0.053
Bi5	0.083	0.099	0.142	0.092
Bi6	0.049	0.063	0.057	0.044
Bi7	0.057	0.074	0.072	0.048
Bi8		0.119	0.109	
Bi9		0.102	0.099	
Bi10		0.076	0.076	
Bi11		0.108	0.161	
Bi12		0.117	0.123	
Bi13		0.115	0.110	
Bi14		0.097	0.083	
Bi15		0.127	0.143	

Table 4.12. Estimated parameters for children in families with phones whose mothers had a visit for medical care within a year. (continued)

	FAMSIZE			
i =	1	2	3	4
	Odds Ratios			
Bi1	12.170	11.477	6.874	5.036
Bi2	0.558	0.565	0.628	0.586
Bi3	0.697	0.603	0.760	0.821
Bi4	1.099	0.799	0.985	0.986
Bi5	1.262	1.153	1.533	1.444
Bi6	1.225	1.232	1.412	1.178
Bi7	0.801	0.659	0.773	0.763
Bi8		0.889	0.847	
Bi9		1.230	1.001	
Bi10		1.068	0.890	
Bi11		1.341	0.952	
Bi12		1.102	0.959	
Bi13		1.270	1.275	
Bi14		1.100	1.069	
Bi15		0.950	0.597	

Table 4.13. Observed and fitted proportions and standard errors for children with phones whose mothers had a visit for medical care within a year, United States 1975-76.

SUBDOMAIN	FAMSIZE	EDHEAD	MDPA	Observed		Fitted	
				Proportion	Standard Deviation	Proportion	Standard Deviation
1	1	1	1	.877	.01994	.893	.01125
2	1	1	2	.834	.03070	.845	.01671
3	1	1	3	.910	.01898	.874	.01380
4	1	2	1	.927	.01224	.912	.00786
5	1	2	2	.824	.02605	.872	.01162
6	1	2	3	.908	.01481	.896	.00876
7	1	3	1	.936	.00592	.943	.00405
8	1	3	2	.922	.01019	.915	.00728
9	1	3	3	.923	.00840	.932	.00512
10	1	4	1	.947	.00721	.950	.00431
11	1	4	2	.951	.01260	.925	.00805
12	1	4	3	.933	.01248	.940	.00664
13	1	5	1	.968	.00638	.965	.00462
14	1	5	2	.948	.01328	.966	.00482
15	1	5	3	.951	.01556	.972	.00439
16	2	1	1	.878	.01357	.877	.01271
17	2	1	2	.811	.02648	.825	.02306
18	2	1	3	.876	.02537	.891	.02109
19	2	2	1	.912	.00972	.913	.00887
20	2	2	2	.863	.01925	.853	.01856
21	2	2	3	.836	.02571	.845	.02253
22	2	3	1	.925	.00473	.924	.00423
23	2	3	2	.872	.01051	.869	.01002
24	2	3	3	.906	.01379	.906	.01171
25	2	4	1	.954	.00512	.956	.00461
26	2	4	2	.894	.01400	.892	.01268
27	2	4	3	.924	.02053	.933	.01046
28	2	5	1	.964	.00391	.966	.00340
29	2	5	2	.945	.00880	.974	.00412
30	2	5	3	.983	.01250	.994	.00383

Table 4.13. Observed and fitted proportions and standard errors for children with phones whose mothers had a visit for medical care within a year, United States 1975-76. (continued)

SUBDOMAIN	FAMSIZE	EDHEAD	MDPA	Observed		Fitted	
				Proportion	Standard Deviation	Proportion	Standard Deviation
31	3	1	1	.841	.01670	.838	.01582
32	3	1	2	.768	.02628	.762	.02446
33	3	1	3	.845	.04001	.830	.03779
34	3	2	1	.880	.01298	.881	.01224
35	3	2	2	.824	.02513	.837	.02158
36	3	2	3	.805	.02803	.790	.02711
37	3	3	1	.895	.00678	.895	.00621
38	3	3	2	.844	.01180	.848	.01081
39	3	3	3	.872	.01830	.867	.01573
40	3	4	1	.936	.00841	.934	.00757
41	3	4	2	.838	.01839	.829	.01652
42	3	4	3	.950	.02094	.944	.02158
43	3	5	1	.946	.00740	.949	.00616
44	3	5	2	.897	.01477	.904	.01111
45	3	5	3	.788	.09553	.830	.07712
46	4	1	1	.770	.01790	.777	.01090
47	4	1	2	.702	.02245	.693	.01348
48	4	1	3	.779	.03715	.767	.02050
49	4	2	1	.834	.01544	.830	.01231
50	4	2	2	.783	.02364	.759	.01607
51	4	2	3	.815	.03205	.821	.01904
52	4	3	1	.850	.01076	.854	.00767
53	4	3	2	.797	.01580	.791	.01129
54	4	3	3	.867	.02240	.847	.01303
55	4	4	1	.896	.01281	.895	.01041
56	4	4	2	.837	.03293	.847	.01474
57	4	4	3	.840	.06033	.890	.01499
58	4	5	1	.906	.01008	.896	.00746
59	4	5	2	.831	.02917	.849	.01239
60	4	5	3	.882	.05988	.891	.01224

The second largest group of children were those in households with phones whose mothers had not made a visit for medical care within a year. There were approximately eight million such children in the United States in 1975-76 - 14.3 percent of the study population. About three-quarters (74.9 percent) had had medical care within the interval deemed adequate for minimal preventive care.

A factorial model could have been used to analyze the data about these children. The three-way interaction was not significant. However, a model nesting EDHEAD and MDPA within FAMSIZE was used instead for consistency with the first table. Thus the saturated models and main-effects models were identical for the two populations.

For this table, three of the four within-FAMSIZE interactions were not significant in the saturated model. The statistics for the four interaction terms were as follows;

FAMSIZE	X^2	D.F.	p
1	7.18	8	0.518
2	4.61	8	0.799
3	8.71	8	0.367
4	13.50	8	0.096

With an interaction term in FAMSIZE 4 only, the statistics were as follows;

FAMSIZE	X^2	D.F.	p
4	19.46	8	0.013

The goodness of fit chi-square was 23.23 with 24 degrees of freedom (p= 0.506).

A main-effects model with no interaction terms had a goodness of fit chi-square of 42.69 with 32 degrees of freedom (p= 0.098). At a conventional alpha level of 0.05, this model would have fit. Also, in the main-effects model, the contrast between FAMSIZE 1 and FAMSIZE 2 was not significant; the model could have been reparameterized to combine them.

The two levels were not combined because of the desire to keep the family sizes comparable with the first table and, as will become evident later, the third table. The reduced model was selected over the main-effets model primarily because of a wish to retain all of the useful information. An alpha level of 0.25 is sometimes recommended to avoid the possibly unwarrented oversimplification of the model (see, for example, Forthofer and Lehnen 1981). As far as I know, there is no "correct" answer. It is a matter of the analyst's judgement and I was interested in these children in large families.

The test statistics for the reduced final model are shown in table 4.14. The estimated parameters, their standard deviations, and the odds ratios are in table 4.15 and the observed and fitted proportions with their standard devia-

tions are in table 4.16. The vector for the b values in table 4.15 is the same as that for table 4.12. The computer program for this model differs from Example Two in Appendix Five only in that the interaction is in FAMSIZE 4 rather than in FAMSIZE 2 and FAMSIZE 3.

Table 4.14. Test statistics for final model for children in families with phones whose mothers had no visit for medical care within a year.

		X^2	D.F.	p
Goodness of Fit		23.23	24	0.506
Main Effects				
FAMSIZE		1257.59	4	<0.001
EDHEAD		131.95	16	<0.001
Within FAMSIZE				
	1	32.99	4	<0.001
	2	31.84	4	<0.001
	3	36.96	4	<0.001
	4	27.78	4	<0.001
MDPA		49.30	8	<0.001
Within FAMSIZE				
	1	5.44	2	0.066
	2	17.21	2	<0.001
	3	12.67	2	0.002
	4	9.76	2	0.008
Interaction				
Within FAMSIZE				
	4	19.46	8	0.013
Contrasts				
FAMSIZE	1 vs 2	0.68	1	0.407
	2 vs 3	14.73	1	<0.001
	3 vs 4	6.70	1	0.010

Table 4.15. Estimated parameters for children in families with phones whose mothers had no visit for medical care within a year.

	FAMSIZE i = 1	2	3	4
		Beta Values		
Bi1	1.498	1.410	1.028	0.798
Bi2	-0.644	-0.438	-0.355	-0.345
Bi3	-0.040	-0.278	-0.311	-0.299
Bi4	-0.170	-0.057	0.243	0.017
Bi5	0.205	0.263	-0.108	0.295
Bi6	0.160	0.312	0.208	0.214
Bi7	-0.233	-0.169	-0.221	-0.077
Bi8				0.124
Bi9				-0.169
Bi10				-0.261
Bi11				-0.071
Bi12				-0.211
Bi13				-0.136
Bi14				0.292
Bi15				0.157
	Standard Deviations of Beta Values			
Bi1	0.075	0.075	0.065	0.052
Bi2	0.149	0.121	0.106	0.099
Bi3	0.149	0.108	0.116	0.110
Bi4	0.125	0.087	0.093	0.105
Bi5	0.164	0.116	0.128	0.119
Bi6	0.094	0.083	0.084	0.069
Bi7	0.102	0.090	0.082	0.081
Bi8				0.131
Bi9				0.158
Bi10				0.149
Bi11				0.166
Bi12				0.155
Bi13				0.120
Bi14				0.147
Bi15				0.168

Table 4.15. Estimated parameters for children in families with phones whose mothers had no visit for medical care within a year. (continued)

	FAMSIZE			
i =	1	2	3	4
	Odds Ratios			
Bi1	4.474	4.095	2.796	2.220
Bi2	0.525	0.646	0.701	0.708
Bi3	0.961	0.757	0.732	0.742
Bi4	0.844	0.945	1.275	1.018
Bi5	1.228	1.301	0.898	1.343
Bi6	1.173	1.366	1.231	1.239
Bi7	0.792	0.844	0.802	0.925
Bi8				1.132
Bi9				0.845
Bi10				0.770
Bi11				0.932
Bi12				0.810
Bi13				0.873
Bi14				1.339
Bi15				1.171

Table 4.16. Observed and fitted proportions and standard errors for children with phones whose mothers had no visit for medical care within a year, United States 1975-76.

SUBDOMAIN	FAMSIZE	EDHEAD	MDPA	Observed		Fitted	
				Proportion	Standard Deviation	Proportion	Standard Deviation
1	1	1	1	.825	.06406	.727	.03915
2	1	1	2	.614	.06428	.631	.03967
3	1	1	3	.722	.08813	.709	.04335
4	1	2	1	.816	.04247	.832	.02342
5	1	2	2	.742	.05768	.760	.03234
6	1	2	3	.772	.06441	.819	.02428
7	1	3	1	.781	.03241	.816	.02069
8	1	3	2	.781	.03811	.741	.02867
9	1	3	3	.830	.02849	.803	.01963
10	1	4	1	.884	.03582	.873	.01941
11	1	4	2	.794	.06102	.815	.03047
12	1	4	3	.868	.05540	.863	.02536
13	1	5	1	.921	.02630	.902	.01574
14	1	5	2	.899	.04099	.856	.02716
15	1	5	3	.863	.07679	.894	.02037
16	2	1	1	.748	.04166	.787	.02334
17	2	1	2	.703	.05126	.675	.02787
18	2	1	3	.756	.11603	.712	.04489
19	2	2	1	.825	.03471	.817	.02006
20	2	2	2	.738	.04529	.716	.02934
21	2	2	3	.717	.09389	.750	.03458
22	2	3	1	.852	.01646	.838	.01389
23	2	3	2	.738	.03175	.745	.02277
24	2	3	3	.774	.04805	.776	.02800
25	2	4	1	.876	.02468	.879	.01552
26	2	4	2	.801	.04182	.804	.01971
27	2	4	3	.908	.04914	.829	.02833
28	2	5	1	.911	.02057	.910	.01093
29	2	5	2	.841	.03551	.851	.02027
30	2	5	3	.930	.06000	.871	.02419

Table 4.16. Observed and fitted proportions and standard errors for children with phones whose mothers had no visit for medical care within a year, United States 1975-76. (continued)

SUBDOMAIN	FAMSIZE	EDHEAD	MDPA	Observed		Fitted	
				Proportion	Standard Deviation	Proportion	Standard Deviation
31	3	1	1	.673	.04477	.656	.02705
32	3	1	2	.612	.04279	.610	.02912
33	3	1	3	.790	.05773	.670	.03881
34	3	2	1	.687	.05048	.698	.03087
35	3	2	2	.628	.04170	.613	.02921
36	3	2	3	.707	.06889	.672	.03606
37	3	3	1	.811	.02295	.818	.01521
38	3	3	2	.741	.03079	.754	.01918
39	3	3	3	.746	.06968	.799	.02610
40	3	4	1	.768	.04301	.765	.02695
41	3	4	2	.751	.05648	.690	.03657
42	3	4	3	.716	.10112	.742	.03953
43	3	5	1	.864	.02710	.856	.01901
44	3	5	2	.816	.04403	.803	.02665
45	3	5	3	.661	.10733	.841	.02497
46	4	1	1	.645	.03294	.654	.02042
47	4	1	2	.573	.04636	.612	.02391
48	4	1	3	.633	.08830	.518	.03162
49	4	2	1	.652	.04787	.651	.02514
50	4	2	2	.609	.03800	.608	.02732
51	4	2	3	.636	.07900	.514	.03353
52	4	3	1	.691	.02683	.732	.01553
53	4	3	2	.725	.02558	.694	.01786
54	4	3	3	.734	.09419	.607	.02679
55	4	4	1	.761	.04215	.802	.01863
56	4	4	2	.768	.04527	.771	.02228
57	4	4	3	.776	.07908	.696	.03062
58	4	5	1	.848	.02573	.783	.01656
59	4	5	2	.741	.06566	.751	.02269
60	4	5	3	.673	.00224	.672	.00185

The final group of children were those living in households where there was no telephone. There were relatively few such children. They comprised only 10.3 percent of the study population. They may be the most important group from a policy point of view however, because they were the least likely to have received care, and 10.3 percent of the study population still represents six million children. The one million of them whose mothers had not received care within a year were especially unlikely to have received care themselves; only 59 percent had seen a doctor within the interval recommended for their age group.

Because there were relatively so few of them and because cross-classification would have produced so many domains with no data, it was necessary to reduce the number of levels for some of the variables. Education of the head (EDHEAD) was reduced to two levels - 8 years or less and 9 years or more. Family size (FAMSIZE) was also reduced to two levels - 3 or fewer persons and 4 or more persons. With these reductions there were 24 subpopulations that were each of sufficient size for analysis.

To maintain as much comparability as possible with the other two analyses, a model with FAMSIZE, EDHEAD, and MDPA nested in MDMA was used. With no three-way interaction, the statistics for the two-way interactions were as follows;

MDMA	X^2	D.F.	p
1	7.62	5	0.179
2	14.00	5	0.016

and the goodness of fit chi-square was 2.47 with 4 degrees of freedom (p= 0.651).

Since the interaction in MDMA 1 was not significant, the first step in reducing the model was to delete that interaction. When that was done the interaction in MDMA 2 was as follows;

MDMA 2	X^2	D.F.	p
Total	16.74	5	0.005
FAMSIZE*EDHEAD	1.69	1	0.194
FAMSIZE*MDPA	1.02	2	0.598
EDHEAD*MDPA	13.35	2	0.001

and the goodness of fit chi-square was 10.09 with 9 degrees of freedom (p= 0.344). The model fit. However, the only significant component of the total interaction in MDMA 2 was that between the education of the head (EDHEAD) and the interval since the father had last received care (MDPA). With the other two components deleted from the model, the statistics for the remaining interaction were as follows;

MDMA 2	X^2	D.F.	p
EDHEAD*MDPA	13.79	2	0.001

and the goodness of fit chi-square was 13.03 with 12 degrees of freedom (p= 0.367). The model fit. Given that only one component of the interaction was significant, one might have

expected that a main-effects model would have fit. It didn't. The goodness of fit chi-square was 26.83 with 14 degrees of freedom (p= 0.020). Therefore, the model with the EDHEAD*MDPA interaction in MDMA 2 was the final model. The program is given as Example Four in Appendix Six.

The test statistics for the model are given in table 4.17. The estimated parameters and odds ratios are in table 4.18 and the observed and estimated proportions with their standard deviations in table 4.19.

The b vector for table 4.18 is:

Overall mean
Differential effect of family of 0-3 persons (FAMSIZE 1)
Differential effect of education 0-8 years (EDHEAD 1)
Differential effect of father's seeing a doctor (MDPA 1)
Differential effect of father's not seeing a doctor (MDPA 2)
Interaction of FAMSIZE and MDPA (2 terms)

in families where the mother had seen a doctor and families where she had not (MDMA 1 and 2).

Table 4.17. Test statistics for final model for children in families without phones.

	X^2	D.F.	p
Goodness of Fit	13.03	12	0.367
Main Effects			
MDMA	627.93	2	<0.001
FAMSIZE	36.41	2	<0.001
Within MDMA			
1	21.63	1	<0.001
2	12.97	1	<0.001
EDHEAD	18.69	2	<0.001
Within MDMA			
1	18.02	1	<0.001
2	2.45	1	0.118
MDPA	53.41	4	<0.001
Within MDMA			
1	52.62	2	<0.001
2	0.18	2	0.912
Interaction (EDHEAD*MDPA)			
Within MDMA 2	13.79	2	0.001
Contrasts			
MDMA 1 vs 2	76.75	1	<0.001
Within MDMA 1			
MDPA 1 vs 2	8.54	1	0.004
MDPA 2 vs 3	52.58	1	<0.001
Within MDMA 2			
MDPA 1 vs 2	0.14	1	0.707
MDPA 2 vs 3	0.11	1	0.740

Table 4.18. Estimated parameters for children in familes without phones.

	MDMA	
	i = 1	i = 2
	Beta Values	
Bi1	1.530	0.549
Bi2	0.306	0.335
Bi3	-0.235	-0.111
Bi4	0.163	0.036
Bi5	-0.485	-0.033
Bi6		-0.265
Bi7		-0.059
	Standard Deviations of Beta Values	
Bi1	0.063	0.093
Bi2	0.066	0.093
Bi3	0.055	0.071
Bi4	0.056	0.095
Bi5	0.067	0.100
Bi6		0.100
Bi7		0.106
	Odds Ratios	
Bi1	4.619	1.731
Bi2	1.358	1.398
Bi3	0.791	0.895
Bi4	1.177	1.037
Bi5	0.616	0.968
Bi6		0.767
Bi7		0.943

Table 4.19. Observed and fitted proportions and standard errors for children in families without phones, United States 1975-76.

SUBDOMAIN	MDMA	FAMSIZE	EDHEAD	MDPA	Observed		Fitted	
					Proportion	Standard Deviation	Proportion	Standard Deviation
1	1	1	1	1	.884	.04128	.854	.01852
2	1	1	1	2	.769	.05963	.753	.02753
3	1	1	1	3	.872	.03400	.872	.01644
4	1	1	2	1	.924	.01601	.903	.01258
5	1	1	2	2	.901	.02923	.830	.02130
6	1	1	2	3	.908	.01445	.916	.00926
7	1	2	1	1	.754	.02201	.760	.01697
8	1	2	1	2	.616	.03447	.624	.02318
9	1	2	1	3	.810	.03258	.788	.02217
10	1	2	2	1	.836	.01590	.835	.01118
11	1	2	2	2	.720	.02544	.726	.01971
12	1	2	2	3	.866	.01718	.856	.01206
13	2	1	1	1	.571	.12589	.632	.05332
14	2	1	1	2	.759	.10828	.664	.05513
15	2	1	1	3	.643	.10282	.749	.04199
16	2	1	2	1	.855	.07047	.785	.04091
17	2	1	2	2	.788	.06926	.735	.04181
18	2	1	2	3	.682	.05121	.661	.04516
19	2	2	1	1	.490	.04816	.468	.04144
20	2	2	1	2	.523	.05331	.503	.04974
21	2	2	1	3	.632	.05528	.604	.05192
22	2	2	2	1	.669	.04200	.652	.03888
23	2	2	2	2	.597	.04827	.587	.04033
24	2	2	2	3	.527	.06585	.500	.04546

Discussion

The probability of a child's having seen a doctor within an interval short enough to have received the recommended preventive care is highly dependent on whether the child's mother has seen a doctor within a year, on whether there is a

telephone available to the household, and on the size of the family. This was clear from the process used to select the variables for this study. Once the data were arrayed as they are in tables 4.3 - 4.5, it was also clear from observation.

When these three variables are included in the same model as they are in table 4.8, the probability of a chi-square that large or larger was less than 0.001 for each of the three main effects.

Thus, considering the main effects only, the overall probability of 0.779 was increased by 0.080 if the mother had seen a doctor within a year and decreased by 0.080 if she had not. It was increased by 0.051 if there was a telephone and decreased by the same amount if there was not. It was increased by 0.062 if the child was in a family of three or fewer persons, or by 0.028 in a family of four persons, and decreased by 0.004 in a family of five persons or by 0.086 in a family of six or more persons.

Put another way, an estimated 77.9 percent of the children had received care within the recommended interval. In families where the mother had seen a doctor within a year, 85.9 percent of the children had received care but in families where the mother had not seen a doctor, only 69.9 percent of the children had. The difference of 16 percentage points is statistically significant. If the mother had

received care and the family had a telephone, 91.0 percent of the children had received care; if there was no telephone, only 80.8 percent of the children had received care. This difference of 10 percentage points is also statistically significant.

Another way of considering this is to consider the odds ratio interpretation from the logit model. Overall, the odds were four to one that a child had had a contact with a doctor within the recommended interval. Again considering the main effects only, the odds on the child's having care were almost seven to one (4.01 x 1.66 = 6.65) if the mother had seen a doctor and nine to one (4.01 x 1.66 x 1.36 = 9.07) if the mother had seen a doctor and there was a telephone in the house.

There was considerably more information available about these children, however, than just whether the mother had seen a doctor and the availability of a telephone. The size of the family, the education of the family head, the presence or absence of a father and whether he had seen a doctor if present were all important information to consider.

The largest group of children, three-quarters of the study population, were those in households with telephones whose mothers had had a contact with a doctor within a year. These were the children most likely to have received care them-

selves. However, even among these children the probability of having seen a doctor within the interval ranged from 0.693 for children in large families where the family head had not gone beyond eighth grade and the father had not seen a doctor to 0.994 for children in four-person families where the family head had graduated from college and there was no father in the household (Table 4.13).

The proportion of the children who had received care was lower in five-person families than in smaller families and was lower still in families of six or more persons. Within each family size, the proportion of the children with care was inversely associated with the education of the family head. Within each family size, children with fathers in the household were more likely to have had care if the father had seen a doctor within a year than if he had not. In addition, children whose fathers were not in the household were more likely to have received care than children with fathers who had not seen a doctor. All of these main effects are statistically significant.

The situation is complicated in middle-sized families, that is, families of four or five persons. The interaction between the education of the family head is such that children in four person families with family heads who had graduated from college were more likely to have received care if the father was absent than if he was present even if he

had had medical care. This was also true for children in five-person families where the head had some education beyond high school but was not a college graduate.

These statements are based on statistically significant differences based on data from a very large sample. Do they mean anything in real life? The answer is probably yes, although not in such detail that it matters whether the mother has actually graduated from college or whether there are exactly four or five persons in the family. It matters because it throws some light on questions of whether the child in a single-parent family is at a disadvantage compared to one in a two-parent family. The answer is sometimes yes and sometimes no depending on circumstances. We are interested only in whether the child had seen a doctor within a specified interval. In that very restricted sense, some children without fathers were more likely to have received care than some children with fathers especially if the mother had more than the average education. These children were not at a disadvantage for medical care.

In selecting variables for this model, presence or absence of a father was not selected. However, absence of a father was the third level of the medical care for the father variable (MDPA) that was selected. The crude 'yes' vs 'no' was not sensitive, while the one that incorporated the medical care of the father was. Thus it is not presence or absence per se

that is associated with the child's receiving care; it the presence of a father who goes to the doctor that gives the child an advantage - except for children whose mothers have an above-averge education. The child without a father has an advantage over the child whose father has not seen a doctor.

The second group, children in households with phones whose mothers had not seen a doctor within a year, were less likely, as a group, to have received care within the recommended interval than those whose mothers had seen a doctor. However, the range in the proportions was large - from 0.514 through 0.910 (Table 4.13). The education of the head was so important that children in this group with highly educated parents were more likely to have received care than children in the former group whose parents had little education. Again, children with absent fathers were more likely to have received care than children whose fathers were present but who had not seen a doctor. All of the main effects (Table 4.11) were significant.

For children in large families, six or more persons, there was an interaction between the education of the head and the father's status. It was an interaction of degree, however, not a cross-over. Children without fathers were less likely to have received care than children with fathers (regardless of whether the father had seen a doctor) but the difference was not as great at the upper levels of education as at the

lower ones.

Finally there were the children living in households without telephones. If their mothers had seen a doctor, they were more likely to have received care than children with phones whose mothers had not seen a doctor. If their mothers had not seen a doctor, they were much less likely. Those children in small families (three or fewer persons) were especially likely to have received care if the mother had. If, in addition, she had gone beyond the eighth grade, the proportions were very high (Table 4.19). The absent father appeared to be an advantage to these children also - perhaps because of Medicaid regulations.

The odds ratios provide another way of looking at the data. They give the odds that the child will have seen a doctor within the recommended interval. The odds for the specified subpopulations can be obtained by multiplication. For example, the odds are that a child in a family of three or fewer persons that has a telephone are twelve to one if the mother has seen a doctor within a year (Table 4.12) but are only four to one if she has not (Table 4.15). If, however, the family head has not gone beyond eighth grade, the odds that the child has received care are reduced to less than seven to one in the former group (12.170 x 0.558 = 6.791) and just over two to one in the latter (4.474 x 0.525 = 2.349). Of course the odds are increased by the family head having

more education as can be seen by looking at the odds in tables 4.12, 4.15 and 4.18.

The odds on the child having seen a doctor are, as would be expected, especially low for the children in households without telephones (Table 4.18). However, the odds are greatly increased if the child is in a small family and if the father has seen a doctor.

The odds ratio interpretation is that the odds that a child in a household with no telephone and whose mother had seen a doctor within a year had had a contact with a physician within the interval judged adequate for preventive care were almost five to one (4.619). If this child was in a small family the odds were six to one (4.619 x 1.358 = 6.273). If the child was in a small family and the father had seen a doctor, the odds on the child having received care were increased to better than seven to one (4.619 x 1.358 x 1.177 = 7.383). This child had a better chance of receiving care than would first appear. Note that if the child's mother had not seen a doctor within a year, the odds that the child had were less than two to one (1.731)

References

Bhapkar,V.P. and Koch.G.G. Hypotheses of 'No Interaction' in Multidimensional Contingency Tables. Biometrics . 567-594,September,1968.

Choi,J. Equivalence of Two Methods in Balanced Repeated

Replications. Memorandum dated Jan. 14,1981.

Cohen,S.B. and W.D.Kalsbeek. Some statistical implications on the analysis of the design of the National Medical Care Expenditure Survey. Presented at the annual meeting of the American Public Health Association, October 1978.

Forthofer,R.N. and R.G.Lehnen. Public Program Analysis: A New Categorical Data Approach. Lifetime Learning Publications, Belmont CA, 1981.

Freeman,D.H., J.L.Freeman, D.B.Brock, and G.G.Koch. Strategies in the Multivariate Analysis of Data from Complex Surveys II: An Application to the United States National Health Interview Survey. Int.Stat.Rev. 44(3):317-330,1976.

Freeman,D.H., J.L.Freeman, and D.B.Brock. Modularisation for the Analysis of Complex Sample Survey Data. Bull.Int. Stat.Inst. 47(3):3-20,1977.

Freeman,D.H. and D.B.Brock. The Role of the Covariance Matrix Estimation in the Analysis of Complex Sample Survey Data. in Survey Sampling and Measurement,edited by N.K.Namboodiri. Academic Press,New York,1978.

Grizzle,J.E., C.F.Starmer, and G.G.Koch. Analysis of Categorical Data by Linear Methods. Biometrics 25(3):489-504, September 1969.

Holt,M.M. SURREGR: Standard errors of regression coefficients from sample survey data. Research Triangle Institute. May 1977.

Jordan,L.A., S.D.Hanes,K.M.Maurer, and J.Wellisch. Design effects for a National Study of School Children. Presented at the Annual Meeting of the American Educational Research Association, March 20,1982.

Koch,G.G., D.H.Freeman, and J.L.Freeman. Strategies in the Multivariate Analysis of Data from Complex Surveys. Int.Stat.Rev. 43(1):59-78,1975.

Landis,J.R. et al. A Computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT).Comput.Programs Biomed. 6:196-231,1976.

Lemeshow,S. Estimating the Variance of the Slope of a Linear Regression in Complex Sample Surveys. Undated.

Makuc,D. Interfacing SURREGR and GENCAT to Analyze Complex Surveys Proc.of the Soc.Stat.Sec. of the American Statisticsl Association 1981.

Maurer,K., G.Jones, and E.Bryant. Relative Computational Efficiency of the Linearized and Balanced Repeated Replication Procedures for Computing Sampling Variances. Proc. Surv.Res.Meth. American Statistical Association.1978.

National Center for Health Statistics. HES Variance and Cross Tabulation Program, Version 2 by G.Jones. March 10,1977. Unpublished.

National Center for Health Statistics. A General Statistical Methodology for the Analysis of Data From a Complex Survey: NHANES I by Landis,J.R., J.M.Lepkowski, S.A.Eklund, and S.A.Stehouwer.

Vital and Health Statistics. Series 2, Number 92. DHHS Publication Number(PHS) 82-1366, Public Health Service Hyattsville,Md.:U.S.Government Printing Office, September 1982.

Shah,B. modification of Holt,M.M. SURREGR: Standard errors of regression coefficients from sample survey data. Research Triangle Institute. April 1982.

Shuster,J.J. and D.J.Downing. Two-way contingency tables for complex sampling schemes. Biometrika,(63)271-276,1976.

Stanek,E.J. and S.Lemeshow. The Behavior of Balanced Half-Sample Variance Estimates for Linear and Combined Ratio Estimates when Strata are Paired to Form Pseudo-Strata. Estadistica XXXII(119):71-79.November 1978.

Survey Research Center Support Group. Osiris IV User's Manual. Institute for Social Research. 1979.

CHAPTER FIVE

Discussion

Survey methodology has come a long way since the National Health Interview Survey first went into the field a quarter of a century ago. However, it seems that much of the effort has gone into the design of samples. Relatively little has gone into the analysis of data from those surveys. Of course, one result of the emphasis on sample design without considering the analysis is that the analytic possibilities have had little impact on survey design.

The focus of the research on sample design has been to produce the "best" national estimates at the least cost. Best is defined as the estimates having the smallest variances. Least cost is defined for the NHIS in terms of cost when the survey is conducted in conjunction with the Current Population Survey.

There are implications for sample design from this analysis.

From a statistical point of view, one of the most remarkable things about the fitted estimates is the reduction in the standard errors. In some cases the standard deviation of the fitted estimate is close to half that for the observed

estimate. Differences, at least statistical differences, are detectable that would not have been from the observed proportions. The reduction in standard errors is particularly notable where the sample size is small. The correlation between the number of sample children and the reduction in the standard deviation in table 4.10 is -0.674. (The relationship is obvious if you compare table 4.10 with table 4.19). The proportional reduction is greatest in the small population groups. Those are the groups where precision is a problem and where, in a self-weighting sample, it costs most to increase precision by increasing the sample size.

The reduction in standard errors through the weighted least squares approach to fitting models has been noted by Freeman and Brock (1978) and in the 1982 NCHS monograph by Landis et al. However, they did not consider the implications for reducing the sample size.

The National Center for Health Statistics and the other Federal statistical agencies conduct surveys with large samples because large samples improve the precision of the estimates, that is, the variance is smaller in a large sample than in a small one with the same design. Collecting data from such large samples costs millions of dollars. In contrast, fitting models costs hundreds, or at most thousands, of dollars. There is a distinct possibility that

estimates of equal precision could be obtained from smaller samples if the currently available analytic power were fully utilized.

The second implication is that, if a weighted least squares approach is to be used, there should be fewer primary sampling units. This comes about because there are primary sampling units in which there are no observations. These "empty" PSU's are used to compute the variance-covariance matrix. When that matrix is inverted for the weighted least squares approach, there are fewer degrees of freedom in actuality than are assumed.

Furthermore, the best national estimate is not necessarily the best estimate for population subgroups of interest in analysis. If the subgroup is small, the estimate will be unreliable even if the subgroup is uniformly distributed throughout the population. This is a result of the self-weighting sample design. If the subgroup is not uniformly distributed, all estimates including the variance will be unreliable. We do not even know how to measure the reliability or variability of the estimates.

Perhaps this can be explained by examples. Large sample theory tells us that if we have simple random sampling of a normally distributed variable, a sample size of 30 may be adequate for a given purpose. However, what is an adequate

sample size if we do not have simple random sampling? One way of planning for analysis is to multiply 30 by the design effect and take that as an adequate sample size for each cell. However, as we have seen, the design effect is not the same for all characteristics. In this study of children there was clustering in the household. In a study of acute conditions among adult males, there might well be clustering in segments but not in households because the adult males might work or travel together, but there would seldom be more than one adult male in a single household.

If the subgroup is not uniformly distributed, it is very likely that there will not be any sample persons in some or many of the PSU's. There may be pairs of PSU's (paired PSU's are used to compute the variances) where both members of the pair have observations, where neither does, or where one does and the other does not. If there are pairs where only one PSU has observations or neither does, when does the estimated variance become unreliable? There appears to be no theory to answer that question.

Robert Casady has proposed some working guidelines for minimum sample sizes and minimum number of paired PSU's with observations that are now being used for the National Health and Nutrition Examination Survey (Casady 1982, NCHS in press). This author has cooperated in the development of those guidelines, and has approved them for use, but we are

well aware that we are beyond really firm theoretical grounds.

There is need, then, for moving theoretical research into the area of analysis of these large national surveys with complex sample designs if we are to take advantage of the data that are collected at such great expense. Theoretical research could also make it possible to reduce the expense. If we move into design-based models for the presentation and analysis of data, it is highly likely that we will find that smaller samples in fewer PSU's will provide estimates that are just as precise as the straight data presentation from large samples and many PSU's that is the present norm. Furthermore, some of the money could be spent in oversampling of small populations or of populations of special interest. A self-weighting sample had advantages when variances were computed using a desk calculator and the only purpose of the survey was to produce national estimates. There is no advantage when a variance-covariance matrix is calculated using a high-speed computer or the populations of interest are subnational populations.

The lack of theory is, however, only one of the problems in analysis.

As must be apparent to anyone who had read this far, the software technology has not been developed to the point where

the very large samples of these national surveys can be used to full advantage. The NHIS includes approximately 42,000 households and 110,000 people each year. Years can be combined to double or triple the sample size. There are several hundred variables inherent in the questions on the questionnaire that are of potential interest to someone. Very large tables can be generated. The software to analyze these large tables simply does not exist. One must make numerous small decisions, only some of which have been described here, in order to proceed with the analysis. As a result, few people attempt it, and of those who do, many make unwarranted assumptions in order to produce results.

The analysis when the variance-covariance matrix from the complex sample design is incorporated is expensive in computer time. Whether hundreds of dollars spent on computer time should be considered too expensive when the data cost millions of dollars to collect is an important question aside from the implications for sample design. It will not be discussed here. Nevertheless, it is considered too expensive as people applying for grants well know. Thus it is necessary to develop strategies for data analysis that will decrease computer costs.

The best strategy seems to be to use the widely available software and the assumption of simple random sampling to screen the data and to make preliminary runs on the models.

When one has gone as far as possible, the complex sample design should be incorporated before proceeding with the final analysis.

There are risks in this procedure. We have encountered cases where the design effect is less than one. If that happens, a variable could be excluded that shouldn't have been. To avoid that possibility, one can change the rejection level to reject under less stringent conditions than will be used in the final analysis.

There are risks of overfitting and there is the great risk that the analyst will interfere and bias the results. It was obvious when selecting variables in this study that changing the number of levels could change the order in which the variables were selected. If I had firmly believed that income, say, or race were of overwhelming importance, I could have changed the levels to increase the chance that they would be included. This is always a risk when categories are created by the analyst, but the risk is increased when there are preliminary analyses. Namboodiri and West (1978) have commented on the risks of overfitting. I think that I have been careful.

The methodology presented here is not the only approach that could have been taken. Fay, for example, has developed an approach to hierarchical log-linear models that relies on

jackknifing (Fay 1982a 1982b). Nevertheless, it gives the analyst the opportunity to address questions that could not have been readily addressed by other means and the results when compared with information from other studies seem reasonable.

Some General Comments on Survey Data

It is interesting and humbling to recognize that the most sophisticated analysis is useless if the data collection is flawed or imposes unanticipated relationships.

It is logical and fits the majority belief that the best predictor of a child's use of medical services is the mother's use of services. Social scientists and physicians have been saying so for years. The three measures - interval since mother's, father's, and child's last contact for medical care - have been accepted for this study as if there were no measurement error. But is it true that the mother's and father's use of medical care are really among the most important predictors of the child's use, or is it an artifact of the way the survey is conducted?

In the NHIS children cannot respond for themselves and an adult, usually the mother, answers the questions about the children. The rationale is that she is the best respondent. But if she has biases, those biases will affect her answers

about her children as well as those about herself. If she denies illness, no one in the household will be ill. If she believes they should see a dentist twice a year, everyone may be reported as having seen a dentist. If she believes in routine preventive care, everyone may be reported as having received such care. The point is that the responses for everyone are influenced by the biases of the one respondent.

I cannot prove that mothers report their childrens' health behavior to be like their own more than is really the case, but I suspect that they do. It is remarkable how many children in the NHIS have a perceived health status precisely the same as the mother's. The clustering in households (Table 1.1) for the response variable for this study is extreme. Perhaps all children do have the same health status and the same use of medical care as their mothers and siblings, but I would like to see research on the subject.

In a study like this where there are data by more than one level of aggregation, there is an additional matter of practical applied statistics to keep in mind. There is relatively little "noise" or within-PSU variation in the data for areas, in this case, counties. There is a great deal of variation among children and families in the counties but none in the county variables. As a result, the variables about characteristics of individuals will have more impact than the variables about characteristics of areas in a

multivariate analysis whether it is categorical or regression based.

Because of this, the geographic variables should not really be used as they were here. Rather they should be treated as primary controls and the data analyzed within them. If I were doing this again, I would ask different questions and approach the analysis differently. For example; What are the medical care characteristics of an area that influence childrens' receipt of medical care? Within those areas, what are the characteristics of children and their families that influence the receipt of care? Do the individual characteristics differ among areas?

I would also be satisfied with data from one year of the NHIS rather than combining data from two years to obtain the large sample that I was told that I would need. The sample is quite large enough; the problem is not sample size for national estimates, it is lack of sample persons in the rarer subpopulations. Merely increasing the size of a self-weighting sample does not solve that.

References

Casady,R.J. Memoranda to Kovar and Johnson dated April 16, June 24, and July 19,1982.

Fay,R.E. Further Results in Jackknifing Chi-Squared Test Statistics Presented at the annual meeting of the American Statistical Association, August 1982.

Fay,R.E. Contingency Table Analysis for Complex Sample

Designs (CPLX). July 1982, Unpublished.

Freeman,D.H. and D.B.Brock. The Role of the Covariance Matrix Estimation in the Analysis of Complex Sample Survey Data. in Survey Sampling and Measurement,edited by N.K.Namboodiri. Academic Press,New York,1978.

Namboodiri,N.K. and K.West. On the Danger of "Overfitting" the Data Set in the Application of the GSK Method in the Analysis of Categorical Data. Proc.Soc.Stat.Sec. of the American Statistical Association,1978.

National Center for Health Statistics. Hematological and Nutritional Biochemistry Referene Data for Persons 6 Months-74 Years of Age: United States,1976-80 by C.Johnson and R.Fulwood. Vital and Health Statistics. Series 11, Number 232. DHHS (PHS) Publication Number 83-1682. In press.

APPENDIX ONE

NHIS Sample Design

The first stage of the sample design consists of drawing a sample of primary sampling units (PSU's) from the approximately 1,900 geographically defined PSU's into which the United States is divided by the Bureau of the Census. A PSU consists of a county, a small group of contiguous counties, or a Standard Metropolitan Statistical Area. The PSU's collectively cover the 50 States and the District of Columbia.

Formation of such PSU's is an art rather than a science although several principles and rules were used. Prominent among them were the following (DHEW 1958b):

1. PSU's should be units for which a wide variety of descriptive statistics is available, since this permits the PSU's to be stratified or classified in an efficient manner.

2. When the PSU is used by a large surveying organization, there are distinct economies in using the same set of PSU's for more than 1 survey. Consequently there are advantages in having the PSU conform to administrative structure in the field, and in having the unit adaptable to many social and economic objectives.

3. For technical sampling reasons, the greater the internal heterogeneity of the PSU, the more efficient it is. This principle tends to produce physically large units.

4. Contrastingly, costs per ultimate sample unit (i.e., cluster of sample households) tend to increase with transportation distances between ultimate units within a PSU, and thus to increase with the size of the PSU. This factor has limited the size of a PSU to not more than a few neighboring counties.

The rules used by the Bureau of the Census to define the PSU's are as follows (Census 1978):

1. Each standard metropolitan statistical area (SMSA) is defined as a separate PSU. However, the two standard consolidated areas (SCA's) are each counted as one PSU, although both SCA's consist of more than one SMSA.

2. Each PSU comprises one county or two or more contiguous counties. This rule does not apply for certain states in New England where minor civil divisions (towns or townships) are used to define the PSU's. In some states, county equivalents are used: Cities, independent of any county organizations, appear in Maryland, Missouri, Nevada, and Virginia; in Louisiana, the county equivalents are known as parishes; and, in Alaska, where there are no counties, census divisions are used.

3. The PSU's are defined within the boundries of the four census regions; that is, Northeast, North Central, South, and West. The States in the four census regions are listed in Appendix Three. Exceptions to this rule have been made for SMSA's which cross regional boundries. Those SMSA's which lie in more than one State are counted as single PSU's for sampling purposes, but the parts within the individual States are separately identified for convenience in tabulation and control. There are 30 of these SMSA's, 7 of which lie in more than one census region.

4. The area of a PSU does not exceed 2,000 square miles in the West and 1,500 square miles in the other regions except in cases where a single county exceeds the maximum area.

5. The 1970 census population of each PSU is at least 7,500 in the West and 10,000 in the other regions except where this would have required exceeding the maximum area specified in rule 4.

6. In addition to meeting the limitations on total area, PSU's are formed to avoid extreme length in any direction.

The PSU's currently in use are closely related to those defined after the 1950 census. The steps employed in combining counties are roughly as follows (Census 1978):

1. The counties involved in SMSA's that are new or redefined since the previous revision are incorporated into the PSU definitions. Parts of PSU's remaining after such an operation are reviewed according to the criteria to see whether further combination of counties into PSU's is necessary.

2. Any single county exceeding the maximum area limitations is automatically classified as a separate PSU, regardless of population.

3. The characteristics of all other counties within the same State are then examined to determine whether they might advantageously be combined with contiguous counties, keeping in mind the objective of hetero-geneity and the population and area limitations.

Combinations are made for -

Counties with less than the minimum population that are combined with other counties, providing this does not conflict with other requirements.

Counties with more than the minimum population that are combined with other counties only if the maximum area is not exceeded and if the combination substantially increases the heterogeneity within the PSU. For example, a combination involving counties in different economic areas is generally considered desirable even if both counties exceed the maximum population. However, the maximum area limitation is adhered to as closely as possible since excessively large PSU's add to operating costs through increased interviewer travel expense.

4. The characteristics of counties in contiguous States within regions are then compared to determine whether better combinations can be obtained by combining counties from neighboring States.

5. The proposed combinations are reviewed to make the final determinations of PSU's. The reviews of this operation (and of the operation grouping PSU's into strata) represent an effort to apply the competent judgments of several persons in these processes. Personal judgment can

> have no place in the actual selection of sample units as known probabilites of selection for all units in the population can only be achieved by the application of random processes of selection. However, there is a large number of ways in which a given population can be structured and arranged prior to the application of the random selection process, and personal judgment can legitimately play an important role in devising an optimum arrangement, i.e., one that will minimize the variances of the sample estimates.

The above principles and rules led to the combining the 3,141 counties, county equivalents, and independent cities existing at the time of the 1970 census into 1,924 PSU's which, since the Bureau of the Census is the data-collection agency for the NHIS, are also used for the NHIS and for other surveys conducted by the Bureau of the Census.

Sampling theory makes it clear that if units to be sampled can be classified into strata whose members tend to be relatively alike within strata and unlike between strata, then sampling variances are reduced from those of samples of equivalent size drawn from an unstratified universe. The PSU's in the original design that was based on the 1950 census were stratified with the principal modes of stratification being geographic location, population density, rate of population growth (originally between 1940 and 1950), proportion of the population which is not white, and type of industry in predominantly urban areas. The details of the stratification are given in Appendix IV of the NCHS publication on the statistical design of the survey (DHEW 1958b).

The sample design adopted for the CPS and the NHIS calls for the combination of PSU's into strata with the selection of 1 PSU from each stratum. For samples like this, it is desirable to form strata with approximately equal populations. With a self-weighting sample, equal stratum sizes also have the advantage of equalizing the interviewer workloads if there is one interviewer for each stratum. The objective, therefore, is to group PSU's with similar characteristics into homogeneous strata with approximately the same 1970 population in each stratum. Because the final counts were not available, the preliminary counts were used for stratification in the redesign based on the 1970 census.

Theory also requires that the PSU's with the largest populations should be in the sample i.e., they should have a high probability of selection. PSU's with the largest populations are designated as self-representing (SR), that is, each of the SR PSU's is treated as a separate stratum and included in the sample with certainty.

If the redesign of the sample could be considered independently of the design already in use, it would be relatively simple to form the strata. However, the number of strata is determined by variance and by cost considerations and it is important to minimize the selection of a new set of PSU's when a sample is redesigned. The interviewers on the CPS and the NHIS often work on the surveys for a long time. A

substantial investment has been made in training them and experience has shown that an interviewer's efficiency and accuracy increase during the first months after she begins (all interviewers for the NHIS are women). For each PSU dropped from the sample and replaced by a new one, a new interviewer will probably have to be hired and trained and there is the additional expense of the temporary loss of accuracy in data produced by the new and relatively untrained interviewer. The rules for reviewing and modifying existing strata for use in a redesign reflect concern for these factors.

The review of the stratification of PSU's was done primarily by searching for "misfit" PSU's in the existing strata rather than undertaking an entirely new stratification for the redesign that was implemented in 1973.

The most important of the characteristics which were the basis of the original stratification and were considered for the restratification are the following;

SMSA or not
Rate of population change
Percent of population living in urban areas
Percent of population employed in manufacturing
Principal industries
Average per capita value of retail trade
Proportion of population who were not white

Limits are established for each of the characteristics and

when a stratum contained a PSU with characteristics outside those limits the PSU was considered for removal. Ordinarily, however, the total number of revisions was really small because a change in several characteristics beyond the established limits was required before a PSU was removed.

In general, the stratification derived for the earlier design was maintained; exceptions for PSU's and strata were governed by the following rules;

1. The largest of the 1,924 PSU's in the United States as of the 1970 census were designated as self-representing PSU's. Except in special cases, all PSU's with a population greater than about 250,000 were made SR. The remaining PSU's were combined with other PSU's with similar characteristics to form the NSR strata. The strata follow regional boundries (Table 1).
2. Strata with NSR PSU's were made to have approximately equal populations.
3. NSR SMSA PSU's were grouped into strata which had only SMSA PSU's so that all strata were either SMSA or non-SMSA strata.
4. The restratification of the PSU's into NSR strata was done strictly on the basis of the characteristics of the PSU's within the stratum without regard to the sample PSU's of the previous design.

Each self-representing PSU is in the sample by definition. As shown in table 1, there were 156 SR PSU's. In each of the remaining strata, 1 PSU was selected for the sample with a probability equal to its proportion of the total U.S. population in 1970. The actual selection was made in 2 stages. In the first, a computer was used to designate controlled selection patterns of PSU's and a probability was assigned to each of the admissible patterns which were

generated by the computer. In the second, one of the computer-generated patterns was randomly selected to represent the new 376 PSU sample.

Table 1. Number and Population of Strata for 376 PSU Design by Region.

Region	Strata	Population	Stratum size
Self-representing strata (SR)			
			Minimum
Total	156	130,495,743	---
Northeast	50	41,854,236	259,754
North Central	32	32,824,092	256,843
South	48	30,049,022	227,222
West	26	25,768,393	213,358
Non self-representing strata (NSR)			
			Average
Total	220	72,681,943	---
Northeast	20	7,143,875	357,194
North Central	72	23,752,566	329,897
South	102	32,745,594	321,035
West	26	9,039,908	347,689

Separate estimates were also desired for each of 41 sub-universes called Tab Areas. These specifications resulted in classification of the PSU's into 376 strata for the original design.

The object of the subsampling procedure within the sample PSU's is to obtain a self-weighting probability sample where every sample unit (person or household) has the same probability of selection so that estimates of population totals can be made by applying a single multiplier to the sample

numbers.

Sampling considerations led to the decision that the variance would be close to the minimum possible for a fixed cost if the sample was selected in compact clusters comprising about 4 adjacent housing units (or the equivalent) on the average. Therefore, ultimate sampling units (USU's) were defined in such a way that each contained an expected 4 housing units. These housing units represented the 70.6 million living quarters presumed to exist as of the 1970 census. The 70.6 million is the sum of the 1970 census count of 68.6 plus one additional unit for every 3 persons not residing in housing units.

As the sample is self-weighting, the overall probability of selection is the same for all USU's. The sampling fractions at each stage of selection must be determined to meet this requirement. The probability of selection of the PSU times the probability of selection of the USU within the PSU must equal this overall probability of selection.

In practice, a segment is selected at an intermediate sampling stage prior to designating the USU. The character of a segment depends on the source it was selected from - it may be a small area, a place, or an enumeration district for example. Segments do not have a common size but may always be expressed as an integral number of USU's. Each USU

belongs to a specific segment and, for the purpose of administrative control, the segments are numbered and the USU's are identified by their segment numbers. It should be noted that in the NCHS publications in Series 10, segments are called the ultimate stage units and many Series 10 give 6 as the expected number of households in a segment.

The selection of the actual sample within each PSU depends on the resources available within that PSU. As part of the 1970 census, the country was divided into 229,000 enumeration districts (ED's). About 121,000 of the ED's were in the mail census which covered most of the larger urban centers; the remainder were in the conventional census. Each ED is identified with a state-county code and an ED number and, for each, an address register was produced as part of the census enumeration. The completeness of the address information determines the way the census information is used in sampling. In general, ED's are classified into 4 categories by the responses to the following 2 questions;

1. Does the enumeration district (ED) have at least 90 percent of the units recorded at addresses in the register with complete street name and house number?
2. Is the ED within the jurisdiction of an authority issuing building permits?

A "Yes" answer to both questions defines a "list" ED. A "No" answer to the second question determines that an area sample will be used, and a "No" to the first with a "Yes" to the

second question determines an area sample using building permits as an update. Thus, there are three classes used to select the sample. In general, the list frame includes the urban areas from which about two-thirds of the NHIS sample is selected.

The CPS has a selection probability of 1 in 1,968. The initial NHIS design had a selection probability of 1 in 1,509.7446 or 1.3035 times the selection probability for the CPS A-design - the design used to produce national estimates. Thus, to select the NHIS sample, the initial selection of, say A29, is matched by the selection of a matching adjacent USU coded Y73 which is the initial NHIS selection, one midway to the following A29 selection, and one NHIS USU designated for each of the remaining proportion (1.0000-0.3035=0.6965) of the A29 selections. This means that the NHIS selections are not evenly spaced selections, which results in a slightly increased sampling error. At the same time, a block of 12 adjacent USU's is set aside for each NHIS selection to maintain the sample over the life of the sample design, that is, until the sample is redesigned following the 1980 census. Thus, the sample design implemented in 1973 can be maintained through 1984 before the sample is exhausted.

However, for some of the smaller PSU's, the number of USU's needed for all surveys planned for the life of the sample design could turn out to be more than the number in the PSU.

This is not a problem in the SR PSU's where the basic sampling interval of 1 in 1,312 means that there are 1,312 different systematic samples of USU's within the PSU. It is a problem in some of the smaller NSR PSU's, however, where the USU's would be exhausted before the end of the decade because the initial selections were separated by less than 80 USU's - the number needed for all the surveys planned at the time the CPS was redesigned.

It would be possible to return to the original USU's later in the decade and reuse them. Instead, an unbiased system of gradual replacement of sample PSU's over the life of the sample design was adopted. However, a system which moved to a different PSU only when the sample PSU was exhausted would eventually exclude the smaller PSU's. To avoid this bias, clusters of PSU's were established within each NSR stratum that contained a PSU unable to last out the decade. Such clusters, called rotation clusters, were determined before the sample PSU was designated. Each PSU in the rotation cluster that contained a sample PSU was assigned the portion of the decade that it was to be in the sample so that each PSU would have its proper expected sampling rate. The sample is shifted from one PSU to another within the cluster according to these assigned portions. Thus, a stratum containing only PSU's with sufficient USU's has only one PSU in the stratum (that PSU is considered a rotation cluster all by itself). A stratum containing both sufficient and

insufficient PSU's may have one or more than one PSU in the sample depending upon whether a sufficient or insufficient PSU was selected. A stratum with only insufficient PSU's will have more than one PSU (all from the same rotation cluster) in the sample.

The number of PSU's that is replaced is actually quite small. There were 14 NSR PSU's in the design designated for replacement during the decade. For the purpose of calculating variances, these replacement PSU's are given the same number as the original PSU.

References

Bureau of the Census. The Current Population Survey -- A Report on Methodology. Technical Paper No. 7. U.S. Government Printing Office, Washington,D.C. 1963.

Bureau of the Census. The Current Population Survey -- Design and Methodology. Technical Paper No. 40. U.S. Government Printing Office, Washington, D.C. January 1978.

National Center for Health Statistics. The Statistical Design of the Health Household-Interview Survey by staff of the U.S.National Health Survey and the Bureau of the Census. Health Statistics. PHS Publication Number 584-A2. Public Health Service. Washington,D.C. U.S.Government Printing Office, July 1958.

APPENDIX TWO

Creation of the Dataset

In order to create the dataset used for this analysis, it was necessary to create a file of children in the households included in the survey and to match the records of the children to the files of conditions and medical care, to the files of all persons to locate the parents or household heads, and to the Area Resource File to obtain characteristics of the communities in which the children lived.

Data from the National Health Interview Survey are coded into hierarchical files in which there is a file of records for households, a file for persons in the households, a file for conditions (illnesses, injuries, or impairments) that the people in the households have, a file for visits to physicians within the two weeks prior to the interview, a file for hospital episodes within the year, and separate files for each of the supplements that may have been used to ask additional questions on a topic of interest - such as health insurance. Because some of the household members may not have had conditions, physician visits, or hospital episodes there is not necessarily an additional record beyond

the person record for each person. On the other hand, because some of the household members may have had multiple conditions, physician visits, or hospital episodes, there may be more than one of each of these records for some people. There is no code in the person record indicating whether there are records of the specified type or, if so, how many there are. Nor is there a code in the person record indicating that records for other specified household members are contained in the person file.

Each record does contain a field that uniquely identifies the person, however, and each record does contain certain demographic information about the person and the household. The identifying and demographic information were used to create the file of children used for this analysis and to match this file with the others.

Each step was carried out separately for the two years, 1975 and 1976 to create two independent files which were then merged.

1. All files were first ordered on columns 3-13 and 19 to identify uniquely the person by primary sampling unit, week of data collection, segment, household, person, and processing quarter (The ordering was done by fields from right to left).

2. All people coded as ages 00-17 in columns 53-54 were identified as children and their person records were copied into a new file. This new file is the child file.

3. The child file was then matched with the physician visit file. The number of physician visits during the two-week

recall period was summed and the sum entered into the child's record. Information about the first visit during the two-week period was also entered into the record. 5. The master person file was then ordered on column 52 (sex) and columns 71-72 (family relationship)

6. The following decision path was followed to match the record of the child to the record(s) of the parent(s) or head of household.

Select on columns 71-72

	Number of Records
If; V or X in 71 and V or 0 in 72, child is not living with a family	142
If; 00, 10, 01, 02 in 71-72, child is head or wife of head	238
If; 3 in 72, child is child of head or spouse go to 143-144 for parents	70,223
If; 1, 2, 4, or 5 in 144, create child with mother file	68,745
If; 1, 3, 4, or 6 in 144, create child with father file but if 143-144=02 for mother, father record will not exist	58,328
If; 0, 7, 8, or 9 in 144, Match child record to each file transferring information on mother to cols information on father to cols if no match, cols - will be blank	
If; 4, 5, or 6 in 72, or if blank in cols - , Match child record to other family file transferring information on head of household to columns -	3,214

7. The file of adult females was searched to find the mother of each of the 68,745 children on the child-with-mother file. For 328 children no record could be found for the mother.

The file of adult males was then searched to find the father of each of the 58,328 children on the child-with-father file. For 337 children no record could be found for the father even after allowing for those children whose mothers were married to men in the Armed Forces.

8. Thus, the final tape file contains

73,817 children ages 00-17

69,062 children of head or spouse	(93.6%)
68,417 matched to mother	(92.7%)
57,983 matched to father	(78.5%)
4,755 not children of head or spouse	(6.4%)
1,161 mismatches or errors	(1.6%)
142 individuals not in families	(0.2%)
238 heads or wives of heads	(0.3%)
2,412 grandchildren of head or spouse	(3.3%)
800 other relatives of head or spouse	(1.1%)

9. For 69,062 children, data on the mother and/or the father was entered on the child file. For 4,755 children, data on the head of the household was entered on the child file.

10. The child file was matched to the supplements used in 1975 and 1976 and selected items such as health insurance coverage and number of siblings were added to the child file.

11. The file was matched to a master file to obtain the paired PSU's and the half-sample weights so that variances could be calculated (this information is not on the NHIS data tapes).

12. The file was ordered by county and matched to the Area Resource File maintained by the Bureau of Health Manpower and characteristics of the county in which the child resided were added to the record.

There are a number of problems with the National Health Interview Survey coding system. It is based on the Bureau of the Census system and is not designed to consistently identify family relationships even within the household.

Instead, it is designed to identify household relationships. The interviewing instructions are very clear: designate one person the head of the household (it must be the husband in

husband-wife families) and then record the relationship of every other individual in the household to that person. Thus the relationship coded is the relationship to the head of the household rather than the relationship of the child to a specified family member.

There are two codes on each person's record that can be used to establish relationships and living arrangement within the household. Both are based on the initial household listing on the face sheet of the questionnaire. The first which is titled "Family Relationship" is a two column code. The type of family - primary or secondary individual and primary or secondary family - is coded in column 71. The relationship is coded in column 72. All relationships are to the head or the spouse of the head. Thus a child can be the child of the head or spouse, grandchild of the head or spouse, or another relative. A few older children are themselves heads or wives. If a child lives in a three-generation family with one of the grandparents listed as head, the child will be coded as a grandchild even if both parents are present. If, however, the father were to be listed as the head, the child would be coded as a child of the head even though the grandparents were present.

The other code which is titled "Family Structure" is designed to identify whether the child lives with the parents. The possible codes are given in the summary of columns 143-144

later in this appendix. Essentially, they specify whether a child lives with one or both parents with or without other adult relatives.

In the so-called normal household consisting of husband and wife with unmarried children there is no problem. Husband, wife, and unmarried minor children are listed in order. The relationship that is coded is the relationship of the child to the husband who is assumed to be father. The wife is assumed to be the mother of the child. According to the "Family Structure" code in cols 143-144, 54,418 of the 73,817 children or 74 percent of the children on the 1975-76 children's file were in this category. For these children, the relationship to the household head was the relationship to the father. The remaining 26 percent lived with one parent only, one or both parents and other adults, with non-relatives, alone, or had been married. Even for these children, there was no real problem if the child lived with one or both parents but no other adult. Together these categories comprised 87 percent of the children.

However, attempting to match children to parents on the basis of the Family Structure code revealed that it was unedited. The most common problem was that although the child supposedly lived with one or both parents, no record could be found for a parent. In other cases, the record could be found for one parent but not the other either because there

was only one adult in the household or because all the adults were of the same sex. In some cases the record for the mother was present but no record could be located for the father - who may have been present but the record had been deleted because he was in the Armed Forces. (Only the civilian noninstitutionalized population is in scope for this survey.)

There were 70,223 children who were coded child of head or spouse in column 72 but it was possible to locate, identify, and verify the data for only 69,062 children from the Family Structure code in cols 143-144. Conversely, there were 71,431 children who were identified in cols 143-44 as living with one or both parents. Thus one could assume that there were 1208 children who were living with parents who were not heads of households.

Therefore, the record counts should not be interpreted to mean that 4,755 children were not living with one or both parents. According to columns 143-144, there were 71,431 children who were identified as living with one or both parents. According to column 72, there were 70,223 children who were coded as the child of the head or spouse of head. Of these 70,223 children, 69,062 (98.3%) could be located and identified in columns 143-144. Thus, as many as 2,369 of the 4,755 children who are on the final data tape as not the children of the head or spouse may have been living with one

or both parents if columns 143-144 were correct. This could happen if their parents were not heads of households. Because the data were unedited, however, there is no way of knowing whether columns 71-72 or 143-144 were correct. The best estimate is that of the 2,369 children and youths, 1,208 were living with parents who were not heads of households and that 1,161 were errors in one direction or the other.

Summary of columns 71-72

142	individuals
117	primary individuals
25	secondary individuals
238	heads or wives of heads of households
42	heads of households
181	wives with husbands at home not in Armed Forces
15	wives with husbands at home in Armed Forces
70,223	children of head or spouse of head of household
3,214	other relatives of head or spouse
2,412	grandchildren of head or spouse
2	parents of head or spouse
800	other relatives of head or spouse
73,817	Total children and youths under age 18

Summary of columns 143-144

64,081	unmarried with one or more parents and no other adult
54,418	two parents
9,086	mother only
577	father only
7,350	unmarried with one or more parents and other adults
5,029	two parents
2,150	mother
171	father
1,428	unmarried with no parents but other adult relatives
349	one relative
1,079	two relatives
754	unmarried living alone or with non-relatives
204	ever married
73,817	Total children and youths under age 18

Thus, from columns 143-144 there were:
59,447 unmarried with two parents in the household
11,236 unmarried with a mother only in the household
748 unmarried with a father only in the household
2,386 others

There is an additional problem that must be acknowledged and about which nothing could be done as it is inherent in the way the household listing is done. Although the child is coded as the child of the head or spouse in column 72 and as living with two parents (or one parent) in columns 143-144, those codes are created in response to the question "How is -- related to --(Head of household)? The response is entered and no additional questions are asked to ascertain whether the child is biologically related to the person or not. Thus a child coded as living with two parents could be living with two biological parents, one biological and one step parent, or two adoptive parents. This is consistent with the Bureau of the Census procedures and data from the National Health Interview Survey are comparable to those from the Current Population Survey. However, it is a misinterpretation to assume that living with two parents means living with two biological parents.

The final data file consisted of 73,817 records - 37,978 for 1975 and 35,839 for 1976. The record length for the file was 740 if the file contained the identifying, weighting, and substantive information and 1,956 if it also contained the half-sample weights for computing variances using the

balanced half-sample replication method.

The file was recoded onto an SAS file with a record length of 287 with contained 73 of the variables on the original file (including the weights and the paired PSU's) and 46 recoded variables. This SAS file, which was much more efficient in computer time than the original NHIS file, was used for all of the analysis in this study and for several publications and presentations - some of which are referred to in Chapter One.

APPENDIX THREE

Variables Used in the Study

I. National Health Interview Survey Variables

A copy of the questionnaire is included in "Current Estimates" for each data year. Those for the two years of data used for this study are "Current Estimates from the Health Interview Survey: United States-1975" and "Current Estimates from the Health Interview Survey: United States-1976" (NCHS 1977a, 1977b). Descriptions of survey procedures and of terms used in the survey are given in those two and all other reports in Series 10 of Vital and Health Statistics.

MDADEQ; Interval since Child's Last Medical Care

"ABOUT how long has it been since -- LAST saw or talked to a medical doctor?"

Eight precoded categories are on the questionnaire and the NHIS tape. They were reordered (T127 on the SAS tape) as;

1. Under 2 weeks
2. 2 weeks to less than 6 months
3. 6 months to less than 12 months
4. 1 year to less than 2 years
5. 2 years to less than 5 years
6. 5 years or more
7. Never

   Unknown

Recode	Frequency
0. No	10,490
1. Yes,	62,744

less than 6 months (codes 1,2) if under 2 years of age
less than 12 months (codes 1-3) if 2-5 years of age
less than 2 years (codes 1-4) if 6-17 years of age
Unknown 583

MDMA; Interval since Mother's Last Medical Care

Question and categories same as MDADE2 on NHIS tape

Recode	Frequency
1. less than 12 months	56,921
2. 1 year or more	11,298
Unknown(includes no mother)	5,598

MDPA; Interval since Father's Last Medical Care

Question and categories same as MDADE2 on NHIS tape

Recode	Frequency
1. less than 12 months	39,200
2. 1 year or more	18,354
Unknown(includes no father)	16,263

FAMSIZE; Size of Family

All members of the household are listed on the front of the questionnaire and their relationship to the to the head of the family is coded. This information is used to code family size on the NHIS tape (T95 on the SAS tape) in units.

Recode	Frequency
1. 3 or fewer members	12,890
2. 4 members	21,225
3. 5 members	17,181
4. 6 or more members	22,521

FAMINC; Family Income

(Hand card to respondent) "Which of those income groups represents your total combined family income for the past 12 months - that is your's, your --'s, etc.? Include income from all sources such as wages, salaries, social security or retirement benefits, help from relatives, rent from property, and so forth."

There are 11 precoded categories on the questionnaire and the NHIS tape (T68 on the SAS tape).
1. Under $1,000
2. $1,000-$1,999
3. $2,000-$2,999
4. $3,000-$3,999
5. $4,000-$4,999
6. $5,000-$5,999
7. $6,000-$6,999
8. $7,000-$9,999
9. $10,000-$14,999

10. $15,000-$24,999
11. $25,000 or more
    Unknown

Recode	Frequency
1. Under $5,000	8,528
2. $5,000-$9,999	15,263
3. $10,000-$14,999	17,181
4. $15,000 or more	22,521
Unknown	5,511

INCPA; Income of Father

Question and categories same as FAMINC

Recode	Frequency
1. Under $6,999	11,303
2. $7,000-$9,999	8,484
3. $10,000-$14,999	16,749
4. $15,000 or more	17,425
Unknown(includes no father)	19,856

INCMA; Income of Mother

Question and categories same as FAMINC

Recode	Frequency
1. None	29,815
2. $1-$4,999	21,929
3. $5,000-$9,999	10,764
4. $10,000	3,728
Unknown(includes no mother)	7,581

EDHEAD; Education of Family Head

If 17 years old or over, ask: "What was the highest grade or year -- attended in school? Did -- finish the -- grade (year)?"

Years of school are recorded in single years which are coded into 11 categories of completed education on the NHIS tape (T64 on the SAS tape).

1. Under 17 years of age
2. None
3. 1-4 years completed
4. 5-7 years completed
5. 8 years completed
6. 9-11 years completed
7. 12 years completed (high school graduate)
8. 13-14 years completed
9. 15 years completed
10. 16 years completed (college graduate)
11. 17 or more years completed (graduate school)
    Unknown

Recode	Frequency
1. 8 years or less	12,175
2. 9-11 years	12,887

3. 12 years	26,044
4. 13-15 years	10,066
5. 16 years or more	11,934
Unknown	729

EDMA; Education of Mother

Question and categories same as EDHEAD

Recode	Frequency
1. 8 years or less	8,134
2. 9-11 years	13,450
3. 12 years	30,828
4. 13-15 years	9,176
5. 16 years or more	6,322
Unknown(includes no mother)	5,907

EDPA; Education of Father

Question and categories same as EDHEAD

Recode	Frequency
1. 8 years or less	8,416
2. 9-11 years	8,753
3. 12 years	20,727
4. 13-15 years	8,438
5. 16 years or more	11,082
Unknown(includes no father)	16,401

AGEGROUP; Age of Child

"What is --'s date of birth?" (Enter date and age)

Age was recorded on the questionnaire and NHIS tape in single years (T53 on the SAS tape).

Recode	Frequency
1. Under 6 years of age	21,348
2. 6-11 years of age	24,484
3. 12-17 years of age	27,985

AGEMA; Age of Mother

Question and recording same as AGEGROUP

Recode	Frequency
1. Under 25 years of age	5,632
2. 25-34 years of age	27,973
3. 35-44 years of age	25,031
4. 45 years or older	9,781
Unknown(includes no mother)	5,400

AGEPA; Age of Father

Question and recording same as AGEGROUP

Recode	Frequency
1. Under 25 years of age	2,363
2. 25-34 years of age	18,400

3. 35-44 years of age	22,488
4. 45 years or older	14,732
Unknown(includes no father)	15,834

AGEBIRTH; Age of Mother at Child's Birth

Child's year of birth was subtracted from mother's year of birth and recorded on the SAS tape in single years.

Recode	Frequency
1. Under 20 years of age	7,771
2. 20-24 years of age	23,040
3. 25-30 years of age	19,478
4. 30-34 years of age	10,707
5. 35 years or older	7,391
Unknown(includes no mother)	5,400

RACE; Race of Child

No question asked. Race was obtained by interviewer observation (T50 on the SAS tape).

Recode	Frequency
1. White	62,356
2. Black	10,476
3. Other	985

SEX; Sex of Child

No question asked if name indicated sex (T52 on the SAS tape).

Recode	Frequency
1. Male	37,453
2. Female	36,364

ACTCHILD; Usual Activity of Child

If under 6 years of age, no question asked

If 6-16 years of age, "What was -- doing MOST OF THE PAST 12 MONTHS - going to school or doing something else?" If something else, "What was -- doing?"

If 17 years of age or older, "What was -- doing during MOST OF THE PAST 12 MONTHS -
If male, - working or doing something else?"
If female, - keeping house, working, or doing something else?"
For both, if something else, "What was -- doing?"
For both, if 45+ years, "Is -- retired?"

The precoded categories on the questionnaire are edited into 7 categories on the NHIS data tape (T74 on the SAS tape).
0. Under 6 years
1. Usually working
2. Keeping house (females only)

3. Retired, health (45+ years)
4. Going to school
5. Something else
6. Retired, other (45+ years)

Recode	Frequency
1. Under 6 years of age	21,348
2. Going to School	51,106
3. Other	1,363

ACTMA; Usual Activity of Mother

Questions and data tape same as ACTCHILD

Recode	Frequency
1. Usually working	23,844
2. Keeping house	43,662
3. Other	822
Unknown(includes no mother)	5,489

ACTPA; Usual Activity of Father

Questions and data tape same as ACTCHILD

Recode	Frequency
1. Usually working	53,932
2. Other	4,008
Unknown(includes no father)	15,877

PHONE; Telephone in the Household

"What is the telephone number here?"

The categories on the NHIS tape (T43 on the SAS tape) are;
1. Yes
2. No.
3. Yes but no number
Unknown

Recode	Frequency
0. No	7,931
1. Yes	65,553
whether or not a number was given	
Unknown	333

FAMILY; The Presence or Absence of Parents

The relationship to the family head was used to link the child's record to the parents' records (Appendix Two). The linked records were used to create the categories.

Recode	Frequency
1. Two parents	57,338
2. Mother only	11,079
3. Father only	645
4. Neither parent	4,755

PARENTS; The Presence of Absence of Fathers

For children with records linked for mothers only there were only two categories of FAMILY.

Recode	Frequency
1. Two parents	57,338
2. Mother only	11,079

HEALTH; Perceived Health Status of Child

"Compared to other persons --'s age, would you say that his health is excellent, good, fair, or poor?"

There were 4 precoded categories on the questionnaire and the NHIS data tape (T145 on the SAS tape) - Excellent, good, fair, and poor.

Recode	Frequency
1. Excellent	43,742
2. Good	26,492
3. Fair or Poor	3,181
Unknown	402

HEALTHMA; Perceived Health Status of Mother

Question and categories same as HEALTH

Recode	Frequency
1. Excellent	30,051
2. Good	29,090
3. Fair or Poor	9,033
Unknown(includes no mother)	5,643

HEALTHPA; Perceived Health Status of Father

Question and categories same as HEALTH

Recode	Frequency
1. Excellent	30,980
2. Good	20,621
3. Fair or Poor	6,181
Unknown(includes no father)	16,035

LIMITED; Limitation of Activity of Child

If under age 1, "Is -- limited in any way because of his health?"

If 1-5 years of age, "Is -- able to take part at all in ordinary play with other children?" "Is he limited in the kind of play he can do because of his health?" "Is he limited in the amount of play because of his health?"

After asking about the person's usual activity, If ages 6-16 and 'something else', "In terms of health would -- be able to go to school?"

If age 6-16 and going to school, "Does(would) -- have to go to a certain type of school because of his health?" "Is

he(would he be) limited in school attendence because of his health?" "Is he limted in the kind or abount of other activities because of his health?"

In brief, the questions for persons age 17 or older are; "Does -- now have a job?" "In terms of health, is -- NOW able able to (work-keep house) at all?" "Is he limited in the amount of (work_housework) he can do because of his health?" "is he limited in the kind or amount of other activities because of his health?" and, for people who are retired, "Does -- health now keep him from working?"

The responses are edited into 5 categores on the NHIS tape (T98 on the SAS tape);
1. Cannot perform usual activity
2. Limited in amount or kind of usual activity
3. Limited in outside activities
4. Not limited
5. Not applicable (no chronic condition)

Recode	Frequency
1. Limited	9,900
2. No limitation	63,917

SEXHEAD; Sex of Head of Family

Derived from relationship to head of family and the linked records (Appendix Two).

Recode	Frequency
1. Male	60,672
2. Female	13,145

HEALTHIN; Health Insurance Coverage of the Child

Questions from the HMO supplement in 1975
Questions from the Health Insurance supplement in 1976

"We are intereted in allkinds of health insurance plans except those which pay only for accidents. (Not counting Medicare or Medicaid) Is anyone in the family covered by hospital insurance, that is, a health insurance plan which pays any part of the hospital bill?" For each plan named, "Is -- covered under this (Name) plan?"

Recode	Frequency
0. No	18,213
1. Yes	54,875
Unknown	729

SIB; Number of Living Siblings of Child - 1976 only

Questions from the Diabetes supplement in 1976

Number of living siblings coded on the data tape in units.

Recode	Frequency

1. 1 sibling	15,187
2. 2-3 siblings	12,914
3. 4-5 siblings	4,503
4. 6 or more siblings	2,421
Unknown	814

T401 MEDICAID used for the child during the past 12 months

Recode	Frequency
0. No	32,576
1. Yes	3,263

T403 Private plan paid for care of the child during the past 12 months

Recode	Frequency
0. No	28,199
1. Yes	7,640

T141 Received AFDC during the past 12 months

Recode	Frequency
0. No	33,104
1. Yes	2,735

PLACE; Usual Place of Medical Care - 1975 only

Questions from the HMO supplement

Recode	Frequency
1. Private doctor's office	26,886
2. Outpatient Clinic	3,896
3. Emergency Room	2,204
4. Other	4,527
Unknown	465

REGION; Census Region of Residence

Coded on the NHIS tape from the address information on the questionnaire (T29 on the SAS tape).

1. Northeast; Maine,New Hampshire,Vermont,Massachusetts, Rhode Island,Connecticut,New York,New Jersey,Pennsylvania
2. North Central; Michigan,Ohio,Indiana,Illinois,Wisconsin, Minnesota,Iowa,Missouri,North Dakota,South Dakota, Kansas,Nebraska
3. South; Delaware,Maryland, District of Columbia,Virginia, West Virginia,North Carolina,South Carolina,Georgia, Florida,Kentucky,Texas,Tennessee,Alabama,Mississippi, Arkansas,Louisiana,Oklahoma
4. West; Montana,Idaho,Wyoming,Colorado,New Mexico,Arizona, Utah,Nevada,Washington,Alaska,Oregon,California,Hawaii

Recode	Frequency
1. Northeast	16,555
2. North Central	19,842
3. South	24,272
4. West	13,148

WEIGHT; The Weight Used to Obtain National Estimates

The weight is calculated as part of the adjustment for nonresponse and stratification described in Chapter Two. It is on all NHIS data tapes and was used as it appears on those tapes (T21316 is the 2-year person weight on the SAS tape).
The range was 26-4221 with a median on 854 and a mean of 888.75. The 90 percent range was 782-1143.

PSEUDO; The Paired Primary Sampling Unit

The paired primary sampling units are required for all variance calculations from the complex survey regardless of the method used to calculate the variances. They are not on the NHIS data tapes and were obtained for this study by matching to a master variance tape.

Stratum; The Stratum of the Sample Design

The strata are required for all variance calculations from the complex survey regardless of the method used to calculate variances. They are not on the NHIS data tapes but the variable PSEUDO is numbered within strata and, if available, is used to form the strata i.e., units 1 and 2 are in stratum 1, units 3 and 4 are in stratum 2.

II. Area Resource File Variables

The Area Resouce File (ARF) is maintained by the Bureau of Health Manpower. Documentation is in the references (BHM 1978,1979). ARF locations and brief citations are given below.

When there was a choice, data from 1975 were used for creation of the SAS tape for this study. When there was not a choice, data from as close to 1975 as possible were used with the limitation that nothing was later than 1976.

ADJACENCY; Human Resource Profile Code (HRPC) for County of Residence

The categories on the ARF are;
0. Core counties of SMSA's of 1,000,000 or more pop.
1. Noncore counties of SMSA's of 1,000,000 or more pop.
2. Counties of metro areas of 250,000-999,999 pop.
3. Counties of metro areas of 50,000-249,000 pop.
4. Contiguous counties of 20,000 or more urban pop.
5. Noncontiguous counties of 20,000 or more urban pop.
6. Contiguous counties of 2,500-19,999 urban pop.

7. Nonontiguous counties of 2,500-19,999 urban pop.
8. Contiguous counties of less than 2,5000 urban pop.
9. Noncontiguous counties of less than 2,500 urban pop.

Recode	Frequency
1. Core counties (code 0)	19,201
2. Fringe counties (code 1)	10,091
3. Other metro (codes 2,3)	24,454
4. Adjacent (codes 4-6)	13,237
5. Not adjacent (codes 7-9)	6,834

SMSA; Standard Metropolitan Statistical Area for County of Residence

The HRPC categories were recoded as;

Recode	Frequency
1. Inside SMSA, HRPC 0-3	53,746
2. Not in SMSA, HRPC 4-9	20,071

MDCODE; Physician to Population Ratio in County of Residence

Derived from the ARF by dividing the number of physicians in the county in 1975 by the population in 1975 (x 100,000).
Tape locations; 223-227 (AMA Dist 1975)
4321-4326 (Census P-26)

The continuous variable (MDPOP on the SAS tape) had a range of 0.00 -1148.91 with a median of 137.01 and a mean of 155.06. The 90 percent range was 32.05-305.54. Only 0.34 percent of the sample children were in counties where there was no physician.

Recode	Frequency
1. Under 70	15,363
2. 70-136	21,505
3. 137-213	17,937
4. 214 or greater	19,012

OUTCO; Whether or not there was an Outpatient Department in the County of Residence in 1975

Derived from the ARF by dividing the number of outpatient departments in the county in 1975 by the population in 1975 (x 100,000)
Tape locations; 1335-1337 (Hosp File,187)
4321-4326 (Census P-26)

The continuous variable (OUTPOP on the SAS tape) had a range of 0.00-15.63 with a median of 0.65 and a mean of 0.72. The 90 percent range was 0.02-2.08.
It was first recoded as OUTCODE;

Recode	Frequency
1. 0	23,639
2. 0.0001-0.6570	15,738

3. 0.6571-1.026	16,279
4. 1.027 or greater	18,161

Then recoded to OUTCO;

Recode	Frequency
0. No	23,639
1. Yes	50,178

OUTAREA; Number of Outpatient Departments per Square Mile of Land Area in the County of Residence

Derived from the ARF by dividing the number of outpatient departments in the county in 1975 by the land area in square miles (x 100,000)
Tape locations; 1335-1337 (Hosp File,187)
8382-8390 (1972 City County Data Book)

The continuous variable (OUTLAND on the SAS tape) had a range of 0.00-0.96 with a median of 0.002 and a mean of 0.02. 0.02. The 90 percent range was 0-0.03.

Recode	Frequency
1. 0.0	23,639
2. 0.001-0.019	39,921
3. 0.020-0.099	7,070
4. 0.100 or more	3,187

EMERCO; Whether or not there was an Emergency Room in the County of Residence in 1975

Derived from the ARF by dividing the number of emergency rooms in the county in 1975 by the population in 1975 (x 100,000)
Tape locations; 1380-1382 (Hosp File,191)
4321-4326 (Census P-26)

The continuous variable (T483 on the SAS tape) had a range of ***

Recode	Frequency
0. No	3,475
1. Yes	70,342

PCHNEW; Prevailing Charge for a Limited Office Visit for a new Patient in the County of Residence in 1976

Copied from the ARF
Tape location; 2527-2532 (HCFA Medi Prev Chge Tape) The continuous variable (T498 on the SAS tape) had a range of 0.00-35.00 with a median of 12.30 and a mean of 12.48. The 90 percent range was 0-20.

Recode	Frequency
1. $5.00 or less	9,454
2. $5.00-$10.00	15,255
3. $10.01-$15.00	33,698
4. $15.01 or more	15,410

CAPINC; Per Capita Income in the County of Residence in 1974

Copied from the ARF
Tape location; 7900-7904 (Census P-25)

The continuous variable (T557 on the SAS tape) had a range of 1827-7982 with a median of 4531 and a mean of 4538. The 90 percent range was 2925-6032.

Recode	Frequency
1. Under $2,500	1,162
2. $2,500-$3,999	17,370
3. $4,000-$5,999	51,481
4. $6,000 or more	3,804

MEDSCH; Median Years of School Completed in the County of Residence in 1970

Copied from the ARF
Tape Location; 8095-8097 (Census 4th count population)

The continuous variable (T575 on the SAS tape) had a range of 7.40-13.00 with a median of 12.10 and a mean of 11.70. The 90 percent range was 10.1-12.4.

Recode	Frequency
1. Under 9	2,678
2. 9.0-11.9	22,088
3. 12.0-12.9	48,885
4. 13.0 or more	166

AFDCCODE; Proportion of Families on Aid to Families with Dependent Children in 1975

Copied from the ARF (x 100,000)
Tape Location; 7985-7990 (1976 NCSS A-8)

The continuous variable (AFDCPOP on the SAS tape) had a range of 0.00-58,756.05 with a median of 4152.52 and a mean of 5061.86. The 90 percent range was 1009.32-10,917.30.

Recode	Frequency
1. Under 2712.0	18,523
2. 2712.0-4152.51	18,446
3. 4152.52-6150.04	18,201
4. 6150.05 or more	18,647

FEMFAM; Proportion of Families with Female Heads in the County of Residence in 1975

Copied from the ARF
Tape Location; 8190-8196 (1972 City County Data Book)

The continuous variable (T587 on the SAS tape) had a range of 3.70-25.20 with a median of 10.00 and a mean of 10.57. The 90 percent range was 6.9-14.0.

Recode	Frequency
1. Under 8.0	16,785
2. 8.0-11.9	36,544
3. 12.0-15.9	15,651
4. 16.0 or more	4,837

References

Bureau of Health Manpower. The Area Resource File: ARF A Manpower Planning and Research Tool. DHEW Publication Number(HRA) 80-4, Public Health Service. Hyattsville,Md. October 1979.

Bureau of Health Manpower. Area Resource File Technical Documentation. Unpublished. December 1978.

National Center for Health Statistics. Current Estimates from the Health Interview Survey: United States-1975 by T.F.Drury. Vital and Health Statistics. Series 10, Number 115. DHEW Publication Number(HRA) 77-1543, Public Health Service. Rockville,Md.: U.S.Government Printing Office, March 1977.

National Center for Health Statistics. Current Estimates from the Health Interview Survey: United States-1976 by E.R.Black. Vital and Health Statistics. Series 10, Number 119. DHEW Publication Number(PHS) 78-1547, Public Health Service. Hyattsville,Md.: U.S.Government Printing Office, November 1977.

APPENDIX FOUR

Comparison With Other Data

Because the National Health Interview Survey(NHIS) counts are ratio-estimated to agree with the Bureau of the Census population estimates, the NHIS estimates should agree with those from the Current Population Survey(CPS) and other surveys conducted by the Bureau of the Census. There may be differences in the national estimates due to sampling error and to differences in the time at which data are collected for any age-race-sex category but, in general, the target populations are the same for the two surveys and similar listing and interviewing procedures are used.

Data from the two surveys do, in fact, agree very well for those variables for which comparisons are possible.

The Bureau of the Census population estimate for the average of July 1, 1975 and July 1, 1976 was 65,691 children and youths under age 18 (Census, 1978). The annual average for the National Health Interview Survey in 1975-76 was 65,722 children and youths. Agreement for each of the three age groups used in this analysis was also good (Table 1).

The number of children by income category as estimated from the NHIS and from the March 1976 CPS is shown in table 2. Estimates from the two surveys are not strictly comparable in this case because the non-response data have been allocated on the CPS but not on the NHIS and because the reference periods are not the same. However, comparability appears to be reasonably good although the NHIS data are deficient in the tails of the distribution - especially the upper one.

Comparability for living arrangements, shown in table 3, is also good. The CPS data, however, exclude all ever-married persons regardless of age while the NHIS data only exclude those age 17 and over; marital status is not reported on that survey for persons under age 17.

Comparability for family size is shown in table 4 for the 1975-76 NHIS and the March 1977 CPS. By 1977 the total number of children was smaller due to the declining birth rate which may account for a smaller proportion of the children being in large families in the CPS data. The deficiency is so consistent, however, that it seems more likely that a difference in survey procedures or in definitions used for tabulation accounts for it.

Table 1. Number of Children and Youths under age 18, by Age

Age	NHIS 1975-76 Annual Average	CPS July 1, 1975-76 Average
	Number in thousands	
Total	65,722	65,691
Under 6 years	19,217	19,141
6-11 years	21,471	21,534
12-17 years	25,034	25,016

SOURCE: P-25 No-721

Table 2. Number of Children and Youths under age 18, by Income

Family Income	NHIS 1975-76 Annual Average	CPS March 1976
	Number in thousands	
Total	65,722	65,750
Under $2,000	1,519	1,811
$2,000-2,999	1,587	1,629
$3,000-3,999	2,203	1,990
$4,000-4,999	2,398	2,460
$5,000-5,999	2,782	2,214
$6,000-6,999	2,952	2,486
$7,000-9,999	7,900	8,141
$10,000-14,999	15,647	15,052
$15,000-24,999	16,683	20,504
$25,000 or more	7,093	8,662
Not Reported	4,958	---

Source: P-60 No-105.

Table 3. Number of Children and Youths under age 18, by Living Arrangements

Living Arrangement	NHIS 1975-76 Annual Average	CPS March 1976
	Number in thousands	
Total	65,532	65,129
	Percent Distribution	
Total	100.0	100.0
In Families	98.9	99.3
Two Parents	80.5	80.0
One Parent	16.5	17.1
Mother	15.5	15.8
Father	1.0	1.2
No Parent	2.0	2.3
Not In Families	1.1	0.7

NOTE: Married children and youths are excluded.
SOURCE: P-23 No-66.

Table 4. Number of Children and Youths under age 18, by Family Size

Family Size	NHIS 1975-76 Annual Average	CPS March 1977
	Number in thousands	
Total	65,578	63,729
2 Person	1,616	1,689
3 Person	9,833	9,379
4 Person	18,842	18,870
5 Person	15,228	15,017
6 Person	9,430	8,795
7 Person	4,840	4,786
8 Person	2,680	2,463
9 or More	3,108	2,730

NOTE: Family size was not reported in the NHIS for an estimated 144 thousand children and youths.
SOURCE: P-20 No-326

Another survey conducted by the Bureau of the Census was the 1976 Survey of Income and Education (SIE) conducted in

April-July, 1976. This survey was one of the largest ever undertaken. It was designed to fulfil two Congressional mandates (1) to estimate for each State the number of children 5-17 years of age living in poverty families and (2) to estimate the number of children and other persons in each State who are in need of bilingual education guidence and counseling because of limited English-speaking ability. Approximately 191,500 households were selected in a fashion to make State estimates possible. The final interview rate for the 158,000 occupied housing units was 95.4 percent.

According to the SIE, there were 64,610 thousand children and youths under age 18 in the spring of 1976. Approximately 9,867 thousand or 15.3 percent were classified as living below the poverty level; 17,814 thousand or 27.6 percent were below 150 percent of poverty. Of the 10,434 thousand children in families with female heads, 49.0 percent were living below the poverty level; 67.8 percent were below 150 percent of poverty.

Children were more likely to be living in poverty than people of any other age group tabulated (14.0 percent of the people age 65 or older were living below the poverty level). They constituted only 31 percent of the total population but 41 percent of the population in poverty.

The poverty concept is important because many public programs

have eligibility requirements based on the official poverty designation or some multiple of it. The poverty index, originally developed by Mollie Orchansky of the Social Security Administration in 1964 and revised by a Federal Interagency Committee in 1969, provides a range of income cutoffs adjusted by family size, sex of family head, number of children under age 18, and farm or non-farm residence. In 1975 the threshold ranged from a low of $2,187 for a single female living on a farm as an unrelated individual to a high of $9,792 for a nonfarm family of 7 with a male head and one related child under age 18.

Because of the way the information on family income is obtained in the NHIS (respondents are shown flash cards with income categories listed on them and are not asked to furnish exact dollar amounts), because family income was not reported for an estimated 5,910 thousand children and youths in the NHIS and is not imputed for the NHIS although it is for the CPS, and because there is not a family record giving the age and sex of each family member living on the income of this family - information which is required to relate the family to the poverty level - it was not possible to make comparable estimates of the number of children living in poverty from the NHIS. It was possible, however, to make approximations by using the available family size and income information. Using this approximation developed by Diane Makuc, 11.8 percent of the children and youths were at the lower two

poverty levels and 21.1 percent at the lower three in 1976. The results of this calculation are shown in Table 5.

Table 5. Number of Children and Youths under age 18, by Family Size and Makuc's Poverty index

Character	Number in Thousands	Poverty Index 1	2	3	4
		Percent Distribution			
Total	59,275	4.5	7.3	9.2	79.1
Race					
White	49,783	3.0	5.6	7.6	83.8
Black	8,644	13.3	16.7	17.8	52.3
Other	848	6.1	10.0	12.8	71.1
Family					
Two Parents	46,058	1.9	3.9	6.4	87.8
Mother Only	8,962	15.1	20.8	21.1	43.1
Father Only	562	3.6	7.4	8.1	80.9
Neither	3,693	12.0	16.1	15.3	56.6
Family Income					
Under $5,000	6,988	38.4	61.6	0.0	0.0
$5,000-9,999	12,951	0.0	0.0	41.9	58.1
$10,000-14,999	14,990	0.0	0.0	0.0	100.0
$15,000 or more	24,346	0.0	0.0	0.0	100.0

Although the levels are not comparable, the distributions are similar; black children and children living with no father in the household were more likely than other children to be living with very little money income available even when there is a control for family size. It must be noted that this income does not include payments in kind such as Medicaid, food stamps, or housing allowances.

APPENDIX FIVE

Variable Selection Tables

Table 1. Population Selection
All children under age 18, children with mothers, and children with mothers and with interval since mother's and child's last contact for medical care known

	All children under age 18		Children with mothers		Children with mothers and with MDMA and MDADEQ known	
	X^2	X^2 / D.F.	X^2	X^2 / D.F.	X^2	X^2 / D.F.
NEED						
LIMITEDACT	834.74	104.34	738.49	92.31	704.59	176.15
HEALTH	297.87	37.23	239.94	29.99	133.81	33.45
CHILD						
AGEGROUP	304.90	76.22	312.93	78.23	285.16	142.58
RACE	305.50	76.37	251.08	62.77	232.74	116.37
SEX	0.07	0.07	0.33	0.33	0.31	0.31
USUALACT	210.92	21.19	184.88	18.49	140.14	28.03
FAMILY						
FAMINC	730.35	91.29	668.07	83.51	522.45	130.61
FAMSIZE	1425.34	237.56	1403.29	233.88	1379.32	459.77
EDHEAD	1182.09	118.21	1073.56	107.36	1548.00	309.60
PHONE	559.04	139.76	550.43	137.61	530.60	265.30
HEALTHIN	465.94	116.49	425.56	106.39	345.30	172.65
SEXHEAD	7.22	3.61	6.27	3.13	3.65	3.65
PARENTS	21.91	10.95	6.27	3.13	3.65	3.65
PARENTS						
EDPA	1586.79	158.68	1518.27	151.83	1481.35	296.27
EDMA	1673.54	167.35	1581.23	158.12	1442.53	288.51
AGEPA	406.22	50.78	393.07	49.13	358.54	89.63
AGEMA	490.81	61.35	353.48	58.91	340.78	113.59
ACTPA	79.00	19.75	62.59	15.65	46.65	23.33
ACTMA	18.86	9.43	18.86	9.43	18.39	6.13
MDPA	941.77	156.96	861.53	143.59	828.38	267.13
MDMA	2398.82	399.80	4628.05	771.34	1963.45	981.72
HEALTHPA	215.44	35.91	188.57	31.43	185.76	61.92

HEALTHMA	209.97	34.99	109.81	18.30	31.20	15.60
INCPA	912.94	114.12	876.73	109.59	848.92	212.23
INCMA	261.21	32.65	134.68	16.83	69.36	17.34
AGEBIRTH	289.76	28.98	151.26	18.91	143.02	35.75
COUNTY						
REGION	401.84	66.97	341.96	56.99	330.14	110.05
ADJACENCY	425.60	53.20	408.80	51.10	413.30	103.32
SMSA	276.32	138.16	273.21	136.60	274.21	274.21
MDCODE	368.17	61.36	369.79	61.63	367.23	122.41
OUTCO	286.56	143.28	266.88	133.44	271.74	271.74
OUTAREA	190.97	63.66	171.83	57.28	347.75	115.92
EMERCO	33.76	16.88	41.39	20.69	64.15	64.15
PCHNEW	221.78	36.96	212.37	35.39	210.38	70.13
CAPINC	621.62	103.60	582.28	97.05	581.10	193.70
MEDSCH	248.67	41.45	229.72	38.29	229.45	76.48
AFDCCODE	11.92	1.99	6.31	1.05	3.45	1.15
FEMFAM	38.41	6.40	32.65	5.44	32.10	10.70
1975 ONLY						
USUALPLACE	560.58	140.15	423.52	105.88	837.20	837.20
PLACE	421.74	52.72	317.80	39.73	630.17	210.06
1976 ONLY						
SIB	856.90	107.11	766.63	95.83	892.57	223.14
AFDC	7.84	1.96	5.63	1.41	5.72	5.72
MEDICAID	18.02	9.01	9.13	4.57	18.03	18.03
PRIVATE	1159.50	579.75	1128.57	564.28	1025.60	1025.60

NOTES: Data based on 73,817 children under age 18, 68,417 children with matched records for mothers, and 67,817 children with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling.

Table 2. Population Selection
All children under age 18, children with mothers, and children with mothers and with interval since mother's and child's last contact for medical care known controlling for interval since mother's last contact for medical care

	All children under age 18		Children with mothers		Children with mothers and with MDMA and MDADEQ known	
	Sum Q	Sum Q / D.F.	Sum Q	Sum Q / D.F.	Sum Q	Sum Q / D.F.
NEED						
LIMITEDACT	747.81	46.74	657.47	41.09	652.44	54.37
HEALTH	150.95	9.43	132.61	8.29	130.52	10.88
CHILD						
AGEGROUP	234.39	29.30	233.54	29.19	228.82	38.14
RACE	289.90	36.23	273.16	34.15	267.71	44.62
SEX	15.44	3.86	0.76	0.19	0.68	0.23
USUALACT	122.02	6.10	101.02	5.05	95.72	6.38
FAMILY						
FAMINC	518.29	32.39	506.02	31.63	503.31	41.94
FAMSIZE	1189.82	99.15	1147.49	95.62	1140.20	126.69
EDHEAD	896.64	44.83	859.13	42.96	1338.11	89.21
PHONE	474.64	59.33	466.29	58.29	465.64	77.61
HEALTHIN	309.67	38.71	303.88	37.99	301.70	50.28
SEXHEAD	21.40	5.35	9.47	2.37	7.99	2.66
PARENTS	9.08	2.27	9.47	2.37	7.99	2.66
PARENTS						
EDPA	1344.43	67.22	1286.72	64.34	1271.82	84.79
EDMA	1278.21	63.91	1266.48	63.32	1235.50	82.37
AGEPA	253.88	15.87	237.01	14.81	234.48	19.54
AGEMA	226.10	14.13	221.45	18.45	219.48	24.39
ACTPA	63.34	7.92	48.84	6.11	46.83	7.81
ACTMA	28.70	2.39	24.38	2.03	21.70	2.41
MDPA	621.45	51.79	596.33	49.69	595.00	66.11
MDMA						

HEALTHPA	185.33	15.44	167.13	13.93	163.67	18.19
HEALTHMA	121.55	10.13	117.16	9.76	114.85	12.76
INCPA	803.76	50.24	779.13	48.70	774.69	64.56
INCMA	89.50	5.59	82.21	5.14	71.43	5.95
AGEBIRTH	104.36	5.22	99.28	6.21	95.87	7.99
COUNTY						
REGION	428.35	35.70	375.93	31.33	370.64	41.18
ADJACENCY	413.84	25.87	404.04	25.25	397.71	33.14
SMSA	268.96	67.24	262.49	65.62	302.49	100.83
MDCODE	376.53	31.38	366.55	30.55	365.54	40.62
OUTCO	276.53	69.13	266.48	66.62	261.63	87.21
OUTAREA	221.92	18.49	192.21	64.07	348.97	43.62
EMERCO	48.48	12.12	47.55	11.89	68.83	22.94
PCHNEW	227.32	18.94	218.29	18.19	212.93	23.66
CAPINC	592.09	49.34	561.22	46.77	559.57	62.17
MEDSCH	230.28	19.19	213.58	17.80	211.59	23.51
AFDCCODE	21.66	1.81	24.25	2.02	19.82	2.20
FEMFAM	54.40	4.53	51.06	4.26	42.19	4.69
1975 ONLY						
USUALPLACE	779.33	97.42	690.23	86.28	142.75	47.58
PLACE	579.09	36.19	532.89	33.31	529.84	132.46
1976 ONLY						
SIB	751.62	46.98	733.78	45.86	720.43	60.04
AFDC	17.33	2.17	14.74	1.84	19.60	6.53
MEDICAID	26.95	6.74	12.78	3.20	13.70	4.57
PRIVATE	990.72	247.68	957.45	239.36	951.92	317.31

NOTES: Data based on 73,817 children under age 18, 68,417 children with matched records for mothers, and 67,817 children with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling.

Table 3. Population Selection
All children under age 18, children with mothers, and children with mothers and with interval since mother's and child's last contact for medical care known controlling for interval since mother's last contact for medical care and family size

	All children under age 18		Children with mothers		Children with mothers and with MDMA MDADEQ known	
	Sum Q	Sum Q / D.F.	Sum Q	Sum Q / D.F.	Sum Q	Sum Q / D.F.
NEED						
LIMITEDACT	715.72	22.37	625.38	19.54	625.32	13.03
HEALTH	194.26	6.07	183.48	6.55	163.39	3.40
CHILD						
AGEGROUP	270.85	16.93	270.00	16.88	270.70	11.28
RACE	209.89	13.12	193.15	12.07	210.51	8.77
SEX	30.65	3.83	13.91	1.74	8.48	0.71
USUALACT	803.36	20.08	711.29	17.78	116.20	1.94
FAMILY						
FAMINC	592.22	18.51	579.95	18.12	619.73	12.91
FAMSIZE	---		---		---	
EDHEAD	715.77	17.89	678.26	16.96	1030.79	17.18
PHONE	421.73	26.36	413.38	25.84	433.11	18.05
HEALTHIN	269.52	16.85	263.73	16.48	272.17	11.34
SEXHEAD	87.43	10.93	86.68	12.38	53.73	4.48
PARENTS	87.43	10.93	86.68	12.38	53.73	4.48
PARENTS						
EDPA	1018.13	25.45	960.42	24.01	980.57	16.34
EDMA	945.32	23.63	933.59	23.34	916.00	15.27
AGEPA	360.32	11.26	339.49	12.12	61.42	1.28
AGEMA	353.12	11.03	341.26	16.25	283.43	7.87
ACTPA	155.46	9.72	113.40	7.09	100.41	4.18
ACTMA	149.41	6.23	153.84	9.62	60.64	1.68
MDPA	630.63	26.28	605.54	25.23	636.35	17.68
MDMA	---		---		---	

HEALTHPA	216.12	9.01	189.12	7.88	160.46	4.46
HEALTHMA	155.64	6.49	159.11	6.63	105.70	2.94
INCPA	775.07	24.22	750.44	23.45	774.36	16.13
INCMA	179.55	5.61	180.04	6.43	116.96	2.44
AGEBIRTH	220.42	6.89	214.06	6.69	275.22	5.73
COUNTY						
REGION	516.30	21.51	463.88	19.33	229.88	6.39
ADJACENCY	419.85	13.12	410.05	12.81	462.95	9.64
SMSA	248.09	31.01	241.62	30.20	251.36	20.95
MDCODE	350.00	14.58	340.02	14.17	320.46	8.90
OUTCO	255.86	31.98	245.81	30.73	247.13	20.59
OUTAREA	231.86	9.66	206.02	8.58	374.01	10.39
EMERCO	40.62	5.08	39.69	4.96	69.73	5.81
PCHNEW	183.99	7.67	172.32	8.21	240.56	6.68
CAPINC	552.67	23.03	521.80	21.74	542.07	15.06
MEDSCH	198.66	8.28	158.15	7.53	470.31	13.06
AFDCCODE	47.88	2.00	51.49	2.14	49.51	1.38
FEMFAM	60.86	2.54	51.43	2.14	105.70	2.94
1975 ONLY						
USUALPLACE	766.65	47.92	266.48	16.66	704.42	29.35
PLACE	561.51	17.55	515.31	16.10	554.95	11.56
1976 ONLY						
SIB	305.18	9.54	272.40	8.51	363.56	7.57
AFDC	51.36	3.21	48.77	3.05	26.74	2.23
MEDICAID	60.23	7.53	46.06	5.76	36.38	3.03
PRIVATE	1034.84	129.36	899.48	112.44	891.66	74.31

NOTES: Data based on 73,817 children under age 18, 68,417 children with matched records for mothers, and 67,817 children with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey. All statistics were calculated on unweighted data assuming simple random sampling.

Table 4. Variable Selection; First Level
Children under age 18 without limitation of activity, with mothers, and with interval since mother's and child's last contact for medical care known

	Levels	Chi-Square	Degrees of Freedom	X^2 D.F.
NEED				
LIMIT	2	499.22	1	499.22
HEALTH	5	76.22	4	19.06
reduced	4	73.73	3	24.58
CHILD				
AGEGROUP	3	299.52	2	149.76
RACE	3	227.81	2	113.90
SEX	2	0.03	1	0.03
USUALACT	6	147.57	5	29.51
reduced	3	138.05	2	69.03
FAMILY				
FAMINC	5	520.55	4	130.14
FAMSIZE	4	1357.77	3	452.59
EDHEAD	6	1540.62	5	308.12
reduced	4	1339.09	3	446.36
PHONE	3	511.80	2	255.90
reduced	2	486.92	1	486.92
HEALTHIN	3	356.24	2	178.12
PARENTS	2	1.93	1	1.93
PARENTS				
EDPA	6	1472.74	5	294.55
reduced	4	1266.67	3	422.22
EDMA	6	1428.30	5	285.66
reduced	4	1278.23	3	426.08
AGEPA	5	358.46	4	89.62
AGEMA	4	342.32	3	114.11
ACTPA	3	46.83	2	23.42
ACTMA	4	16.85	3	5.62
MDPA	4	793.26	3	264.42
reduced	3	562.17	2	281.09
MDMA	3	1919.85	2	959.92
reduced	2	1687.56	1	1687.56

HEALTHPA	4	206.54	3	68.85
HEALTHMA	4	64.96	3	21.65
INCPA	5	838.23	4	209.56
INCMA	5	70.83	4	17.71
AGEBIRTH	5	138.46	4	34.61
COUNTY				
REGION	4	320.61	3	106.87
ADJACENCY	5	403.73	4	100.93
SMSA	2	266.72	1	266.72
MDCODE	4	361.73	3	120.58
OUTCO	2	267.31	1	267.31
OUTAREA	4	344.01	3	114.67
EMERCO	2	61.37	1	61.37
PCHNEW	4	197.15	3	65.72
CAPINC	4	558.95	3	186.32
MEDSCH	4	217.50	3	72.50
AFDCCODE	4	3.92	3	1.31
FEMFAM	4	32.99	3	11.00
1975 ONLY				
USUALPLACE	2	833.93	1	833.93
PLACE	5	629.62	3	209.87
1976 ONLY				
AFDC	2	3.65	1	3.65
MEDICAID	2	13.59	1	13.59
PRIVATE	2	976.72	1	976.72
SIB	5	877.42	4	219.35

NOTES: Data based on 65,235 children under age 18 who are not limited in activity by a chronic condition, with matched records for mothers, and with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey.
All statistics were calculated on unweighted data assuming simple random sampling.

Table 5. Variable Selection; Second Level
Children under age 18 without limitation of activity, with mothers, and with interval since mother's and child's last contact for medical care known; controlling for interval since mother's last contact for medical care

	Chi-square	D.F.	X^2 / D.F.	Sum of ϱ	D.F.	Sum ϱ / D.F.
NEED						
LIMIT	2166.23	2	1083.11	460.82	2	230.41
HEALTH	1718.62	5	343.72	66.10	6	11.02
CHILD						
AGEGROUP	1892.97	5	378.59	237.40	4	59.35
RACE	1980.14	5	396.03	265.08	4	66.27
SEX	1687.65	3	562.55	0.06	2	0.03
USUALACT	1758.58	5	351.72	81.72	4	20.43
FAMILY						
FAMINC	2256.44	9	250.72	452.32	8	56.54
FAMSIZE	2754.77	7	393.54	1012.39	6	168.73
EDHEAD	2864.10	7	409.16	1167.71	6	194.62
PHONE	2144.17	3	714.72	419.37	2	209.69
HEALTHIN	2068.29	5	413.66	322.05	4	80.51
PARENTS	1689.69	3	563.23	1.47	2	0.74
PARENTS						
EDPA	2784.45	7	397.78	1092.35	6	182.06
EDMA	2829.73	7	404.25	1099.22	6	183.20
AGEPA	1909.95	9	212.22	236.61	8	29.58
AGEMA	1892.08	7	270.30	218.76	6	36.46
ACTPA	1732.53	5	346.51	40.96	4	10.24
ACTMA	1713.63	7	244.80	20.36	6	3.39
MDPA	896.74	2	448.37	409.79	4	102.45
MDMA						
HEALTHPA	1868.39	7	266.91	180.08	6	30.01
HEALTHMA	1652.26	5	330.45	123.80	6	20.63
INCPA	2518.03	9	279.78	772.79	8	96.60
INCMA	1662.62	7	237.52	52.26	8	6.53

AGEBIRTH	1789.12	9	198.79	92.97	8	11.62
COUNTY						
REGION	2085.83	7	297.98	355.13	6	59.19
ADJACENCY	2057.01	9	228.56	384.43	8	48.05
SMSA	1932.40	3	644.13	256.17	2	128.09
MDCODE	2036.25	7	290.89	353.57	6	58.93
OUTCO	1932.09	3	644.03	253.14	2	126.57
OUTAREA	2030.93	7	290.13	337.88	6	56.31
EMERCO	1761.01	3	587.00	66.70	2	33.35
PCHNEW	1891.18	7	270.17	193.71	6	32.29
CAPINC	2203.24	7	314.75	528.74	6	88.12
MEDSCH	1896.26	7	270.89	204.78	6	34.13
AFDCCODE	1706.89	7	243.84	14.13	6	2.36
FEMFAM	1738.51	7	248.36	40.94	6	6.82
1975 ONLY						
USUALPLACE	2097.16	4	524.29	47.43	4	11.86
PLACE	983.44	8	122.93	536.68	8	67.09
1976 ONLY						
SIB	1288.24	8	161.03	708.04	8	88.51
AFDC	823.90	3	274.63	6.94	2	3.47
MEDICAID	822.49	3	274.16	7.15	2	3.58
PRIVATE	1759.27	3	586.42	904.68	2	452.34

NOTES: Data based on 65,235 children under age 18 who are not limited in activity by a chronic condition, with matched records for mothers, and with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey.
All statistics were calculated on unweighted data assuming simple random sampling.

Table 6. Variable Selection; Third Level
Children under age 18 without limitation of activity, with mothers, and with interval since mother's and child's last contact for medical care known; controlling for interval since mother's last contact for medical care and for whether the household has a telephone

	Chi-square	D.F.	X^2 / D.F.	Sum of Q	D.F.	Sum Q / D.F.
NEED						
LIMIT	2608.21	7	372.60	445.27	4	113.32
HEALTH	2200.98	11	200.09	89.25	12	7.44
CHILD						
AGEGROUP	2640.44	11	240.04	302.99	8	37.87
RACE	2615.62	11	237.78	188.86	8	23.61
SEX	2150.32	7	307.19	5.10	4	1.28
USUALACT	2243.50	11	203.95	43.17	8	5.40
FAMILY						
FAMINC	2541.27	19	133.75	346.64	16	21.67
FAMSIZE	3197.31	15	213.15	1069.45	12	89.12
EDHEAD	3053.25	15	203.55	921.57	12	76.80
PHONE		---			---	
HEALTHIN	2354.37	11	214.03	179.10	8	22.39
PARENTS	2209.66	7	315.67	52.92	4	13.23
PARENTS						
EDPA	2037.25	15	202.48	887.98	12	74.00
EDMA	3018.85	15	201.26	862.40	12	71.87
AGEPA	2523.86	19	132.83	464.71	16	29.04
AGEMA	2462.59	15	164.17	333.47	12	27.79
ACTPA	2250.67	11	204.61	91.63	8	11.46
ACTMA	2183.97	14	156.00	31.13	12	2.59
MDPA	2576.89	11	234.26	433.37	8	54.17
MDMA		---			---	
HEALTHPA	2318.68	15	154.58	179.88	12	14.99
HEALTHMA	2237.54	15	149.17	63.33	12	5.27
INCPA	2776.44	19	146.13	581.00	16	36.31
INCMA	2255.53	19	118.71	87.34	16	5.46

AGEBIRTH	2270.02	19	119.47	147.47	16	9.22
COUNTY						
REGION	2515.64	15	167.71	319.90	12	26.66
ADJACENCY	2469.60	19	129.98	319.97	16	20.00
SMSA	2343.68	7	334.81	194.46	4	48.62
MDCODE	2441.66	15	162.78	305.59	12	25.47
OUTCO	2336.22	7	333.75	193.89	4	48.47
OUTAREA	2457.88	15	163.86	291.41	12	24.28
EMERCO	2200.14	7	314.31	52.52	4	13.13
PCHNEW	2331.88	15	155.46	176.03	12	14.67
CAPINC	2545.66	15	169.71	423.93	12	35.33
MEDSCH	2278.30	14	162.74	142.38	12	11.87
AFDCCODE	2174.88	15	144.99	28.37	12	2.36
FEMFAM	2208.09	15	147.21	53.99	12	4.50

NOTES: Data based on 65,235 children under age 18 who are not limited in activity by a chronic condition, with matched records for mothers, and with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey.
All statistics were calculated on unweighted data assuming simple random sampling.

Table 7. Variable Selection; Fourth Level
Children under age 18 without limitation of activity, with mothers, and with interval since mother's and child's last contact for medical care known; controlling for interval since mother's last contact for medical care, whether the household has a telephone, and family size

	Sum of χ^2	D.F.	Sum χ^2 / D.F.	Overall χ^2	D.F.	Overall χ^2 / D.F.
NEED						
LIMIT	411.77	16	25.74	406.21	1	406.21
HEALTH	151.35	48	3.15	88.61	3	29.54
CHILD						
AGEGROUP	331.13	32	10.35	137.69	2	68.85
RACE	178.06	32	5.56	103.58	2	51.79
SEX	17.96	16	1.12	0.02	1	0.02
USUALACT	106.78	32	3.34	6.60	2	3.30
FAMILY						
FAMINC	480.36	64	7.51	332.45	4	83.11
FAMSIZE		---			---	
EDHEAD	726.97	48	15.15	652.20	3	217.40
PHONE		---			---	
HEALTHIN	172.95	32	5.40	106.21	2	53.11
PARENTS	69.41	16	4.34	3.28	1	3.28
PARENTS						
EDPA	728.88	48	15.19	626.90	3	208.97
EDMA	677.75	48	14.12	591.06	3	197.02
AGEPA	327.47	64	5.12	137.35	4	34.34
AGEMA	312.36	48	6.51	167.68	3	55.89
ACTPA	120.85	32	3.78	12.65	2	6.33
ACTMA	90.70	48	1.89	46.03	3	15.34
MDPA	459.10	32	14.35	351.86	2	175.93
MDMA		---			---	
HEALTHPA	175.69	48	3.66	63.62	3	21.21
HEALTHMA	115.23	48	2.40	36.07	3	12.02
INCPA	644.42	64	10.07	494.51	4	123.63

	INCMA	132.84	64	2.08	24.99	4	6.25
	AGEBIRTH	182.02	64	2.84	79.28	4	19.82
COUNTY							
	REGION	424.03	48	8.83	314.07	3	104.72
	ADJACENCY	417.71	64	6.53	264.75	4	66.19
	SMSA	200.51	16	12.53	161.11	1	161.11
	MDCODE	314.87	48	6.56	241.50	3	80.50
	OUTCO	180.26	16	11.27	165.61	1	165.61
	OUTAREA	326.06	48	6.79	265.33	3	88.44
	EMERCO	67.52	16	4.22	29.96	1	29.96
	PCHNEW	215.58	48	4.49	135.13	3	45.04
	CAPINC	404.31	48	8.42	338.24	3	112.75
	MEDSCH	156.53	48	3.26	100.55	3	33.52
	AFDCCODE	68.32	48	1.42	7.36	3	2.45
	FEMFAM	116.15	48	2.42	41.74	3	13.91

NOTES: Data based on 65,235 children under age 18 who are not limited in activity by a chronic condition, with matched records for mothers, and with timing of mother's and child's last medical care known in the 1975-76 National Health Interview Survey.
All statistics were calculated on unweighted data assuming simple random sampling.

Table 8. Variable Selection; Fifth Level
Children under age 18 without limitation of activity, with mothers, and with interval since mother's and child's last contact for medical care known; controlling for interval since mother's last contact for medical care, whether the household has a telephone, family size, and education of the head of household

	Overall Q	Degrees of Freedom	Overall Q / D.F.
NEED			
LIMIT	374.94	1	374.94
CHILD			
AGEGROUP	111.32	2	55.66
RACE	38.30	2	19.15
FAMILY			
FAMINC	84.78	4	21.20
FAMSIZE		---	
EDHEAD		---	
PHONE		---	
HEALTHIN	24.27	2	12.14
PARENTS			
EDPA	16.47	3	5.49
EDMA	125.43	3	41.81
AGEMA	118.70	3	39.57
MDPA	302.54	2	151.27
MDMA		---	
INCPA	192.07	4	48.02
COUNTY			
REGION	310.10	3	103.37
ADJACENCY	175.40	4	43.84
SMSA	99.48	1	99.48
MDCODE	173.38	3	57.79
OUTCO	111.61	1	111.61
OUTAREA	241.62	3	80.54
CAPINC	215.43	3	71.81

NOTES: Data based on 64,663 children under age 18 who are not limited in activity by a chronic condition, with matched records for mothers, and with timing of mother's and child's last medical care known, and with the level of education of the family head known in the 1975-76 National Health Interview Survey.
All statistics were calculated on unweighted data assuming simple random sampling.

Table 9. Variable Selection; Sixth Level
Children under age 18 without limitation of activity, with mothers, and with interval since mother's and child's last contact for medical care known; controlling for interval since mother's last contact for medical care, whether the household has a telephone, family size, education of the head of the household, and interval since the father's last contact for medical care

	Overall χ^2	Degrees of Freedom	Overall χ^2 / D.F.
CHILD			
AGEGROUP	98.16	2	49.08
FAMILY			
FAMSIZE		---	
EDHEAD		---	
PHONE		---	
PARENTS			
MDPA		---	
MDMA		---	
INCPA	162.89	4	40.72
COUNTY			
REGION	269.28	3	89.76
SMSA	104.70	1	104.70
MDCODE	151.71	3	50.57
OUTCO	106.65	1	106.65
OUTAREA	211.01	3	70.34
CAPINC	205.86	3	68.62

NOTES: Data based on 64,663 children under age 18 who are not limited in activity by a chronic condition, with matched records for mothers, and with timing of mother's and child's last medical care known, and with the level of education of the family head known in the 1975-76 National Health Interview Survey.
All statistics were calculated on unweighted data assuming simple random sampling.

APPENDIX SIX

Analytic Set-Up Files

All computer tapes were converted to SAS tapes before any analysis was done. If the conversion had not been made, the set-up files would have to contain statements to make that conversion.

All computer runs were made using WYLBUR (On-Line Business Systems 1981). TSO users would have to make the appropriate changes in the control cards.

The program used in this study is a version of GENCAT (Landis et al 1976). Named LARGEN by Ron Forthofer who modified it, it had two modifications. A small program to convert the lower triangle of the variance-covariance matrix produced by SURREGR into an upper triangle acceptable to the program was incorporated. If this had not been done, there would have to be an intermediate step to make the conversion. The program was also modified to take a table of 200 cells instead of 80 - hence, LARGEN.

To save time and to prevent errors in typing in the design

matrix, it can be obtained using the FUNCAT procedure in SAS, transposed, and copied into the correct place in LARGEN. This was done for the factorial model in Example One. Before transposing, the extra material on the FUNCAT matrix was stripped which left 26 elements per line. The transposition is necessary because GENCAT (and LARGEN) require that the columns be input as rows - a procedure that is logical if one is using cards as input.

The file to compute the variance-covariance matrix and the categorical data analysis for a factorial model in one pass using the version of SURREGR that was in use until 1982 is as follows:

```
EXAMPLE ONE

OLD VERSION OF SURREGR AND FACTORIAL DESIGN

YES MDMA YES PHONE LOGIT

 //  JOB card(s)    TIME=15,REGION=1008K
 //STEP1 EXEC SAS,TIME.SAS=30,REGION.SAS=300K
 //SAS.WORK  DD UNIT=3350,SPACE=(TRK,(480,160),RLSE)
 //CHILD  DD DSN=STUDY,UNIT=3400-6,DISP=(OLD,KEEP),
 // VOL=SER=000000
 //VAROUT  DD DSN=&&DUMMYM,DISP=(NEW,PASS),UNIT=3350,
 // SPACE=(TRK,(15,5),RLSE),DCB=RECFM=FB
 //SYSIN DD *
   DATA;
   SET CHILD.STUDY;
   KEEP MDADE02 PHONE UNIT FAMSIZE MDMA EDHEAD MDPA T21316
        STRATUM PSEUDO;
 * ALL UNKNOWNS CHANGED TO UNIT UNKNOWN AND
   KEPT FOR VARIANCE COMPUTATIONS;
 * DEFINE POPULATION;
 * YES MDMA VISIT YES PHONE;
   IF MDMA=>2 THEN UNIT=.;
   IF PHONE=>2 THEN UNIT=.;
 * DEFINE STRATA;
```

```
  STRATUM=PSEUDO/2 + .5;
  STRATUM=INT(STRATUM);
* SORT BY STRATUM AND PSU IF TAPE NOT IN SORT;
  PROC SORT;
       BY STRATUM PSEUDO;
* CALCULATE VARIANCE-COVARIANCE MATRIX;
  PROC SURREGR MISSPSU BETA DATAOUT;
  CLASSES UNIT FAMSIZE EDHEAD MDPA;
  LEVELS 1 4 5 3;
  MODEL MDADEQ=UNIT*FAMSIZE*EDHEAD*MDPA/NOINT;
  STRATUM STRATUM;
  PSU PSEUDO;
  WEIGHT T21316;
  DATA _NULL_;
  SET DUMMYM01;
  FILE VAROUT;
  PUT (B001-B060) (4*E15.8/);
  PUT (V001-V999) (4*E15.8/);
  PUT (V1000-V1830) (4*E15.8/);
//STEP2 EXEC LARGEN,TIME=5
//FT05F001  DD DSN=&&DUMMYM,DISP=(OLD,DELETE)
//SYSIN  DD *
    5     3     5          1     0  YES MDMA,YES PHONE,LOGIT
   60    60     1                   (4E15.8)
    4                               (4E15.8)
    1     2   120     2     1       (2F2.0)
 1
-1
    4                         1     (60F1.0/60F1.0)
01010101010101010101010101010101010101010101010101010101010101
01010101010101010101010101010101010101010101010101010101010101
    2                         1
    1     2    60     1     1       (2F2.0)
 1-1
    7     1    60           1       (2(26F3.0/)8F3.0)
60 BY 60 DESIGN MATRIX FOR SATURATED MODEL ON NEXT 180 LINES
DESIGN MATRIX OBTAINED FROM FUNCAT, TRANSPOSED, AND INSERTED
       MODEL=FAMSIZE*EDHEAD*MDPA
             FAMSIZE
             EDHEAD
             MDPA
               FAMSIZE*EDHEAD
               FAMSIZE*MDPA
               EDHEAD*MDPA
                 / X ;
    8     1     3                 (60F1.0)    FAMSIZE
01
001
0001
    8     1     4                 (60F1.0)    EDHEAD
00001
000001
0000001
00000001
```

```
     8     1     2                    (60F1.0)     MDPA
000000001
0000000001
     8     1    12                    (60F1.0)     FAMSIZE*EDHEAD
00000000001
000000000001
0000000000001
00000000000001
000000000000001
0000000000000001
00000000000000001
000000000000000001
0000000000000000001
00000000000000000001
000000000000000000001
     8     1     6                    (60F1.0)     FAMSIZE*MDPA
00000000000000000000001
000000000000000000000001
0000000000000000000000001
00000000000000000000000001
000000000000000000000000001
0000000000000000000000000001
     8     1     8                    (60F1.0)     EDHEAD*MDPA
000000000000000000000000000001
0000000000000000000000000000001
00000000000000000000000000000001
000000000000000000000000000000001
0000000000000000000000000000000001
00000000000000000000000000000000001
000000000000000000000000000000000001
     8     1    24                    (60F1.0)     THREE WAY
00000000000000000000000000000000000001
000000000000000000000000000000000000001
0000000000000000000000000000000000000001
00000000000000000000000000000000000000001
000000000000000000000000000000000000000001
0000000000000000000000000000000000000000001
00000000000000000000000000000000000000000001
000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000000000001
```

```
0000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000001
    8     1     1                    (30F2.0/30F2.0)    FAMSIZE 1 VS 2
 0  1-1
    8     1     1                    (30F2.0/30F2.0)    FAMSIZE 2 VS 3
 0  0  1-1
    8     1     1                    (30F2.0/30F2.0)    FAMSIZE 3 VS 4
 0  0  0  1
    8     1     1                    (30F2.0/30F2.0)    EDHEAD 1 VS 2
 0  0  0  0  1-1
    8     1     1                    (30F2.0/30F2.0)    EDHEAD 2 VS 3
 0  0  0  0  0  1-1
    8     1     1                    (30F2.0/30F2.0)    EDHEAD 3 VS 4
 0  0  0  0  0  0  1-1
    8     1     1                    (30F2.0/30F2.0)    EDHEAD 4 VS 5
 0  0  0  0  0  0  0  1
    8     1     1                    (30F2.0/30F2.0)    MDPA 1 VS 2
 0  0  0  0  0  0  0  0  1-1
    8     1     1                    (30F2.0/30F2.0)    MDPA 2 VS 3
 0  0  0  0  0  0  0  0  0  1
//
```

While this run is perfectly feasible and might be the thing to do if the categorical analysis were to be done only once, it is not efficient in computer time because the computation of the matrix is so slow. Instead, on the first run the file VAROUT was labeled (NEW,CATLG) in step one and (OLD,KEEP) in step two and stored online on a WYLBUR disk. That made it available for future analyses which could be done rapidly and interactively. (TSO users can follow the same procedures using TSO conventions.) The following examples show the program as it was actually run to make the variance-covariance matrix available for other analyses. It also shows the changes made in the program after the new version of SURREGR was introduced in April,1982. This new version reads the entire probability vector and the variance-

covariance matrix by rows into the print file. Thus, option 1 is used in GENCAT. Users should note that this is not documented in the published SURREGR documentation.

EXAMPLE TWO

YES MDMA YES PHONE LOGIT

SATURATED NESTED MODEL

```
//  JOB card(s)   TIME=15,REGION=1008K
//STEP1 EXEC SAS,TIME.SAS=30,REGION.SAS=300K
//SAS.WORK  DD UNIT=3350,SPACE=(TRK,(480,160),RLSE)
//CHILD  DD DSN=STUDY,UNIT=3400-6,DISP=(OLD,KEEP),
// VOL=SER=000000
//VAROUT  DD DSN=WYL.ACCT.MGK,DISP=(NEW,CATLG),UNIT=3350,
// SPACE=(TRK,(15,5),RLSE),DCB=RECFM=FB
//SYSIN DD *
  DATA;
  SET CHILD.STUDY;
  KEEP MDADE2 PHONE UNIT FAMSIZE MDMA EDHEAD MDPA T21316
       STRATUM PSEUDO;
* ALL UNKNOWNS CHANGED TO UNIT UNKNOWN AND
  KEPT FOR VARIANCE COMPUTATIONS;
* DEFINE POPULATION;
* YES MDMA VISIT YES PHONE;
  IF MDMA=>2 THEN UNIT=.;
  IF PHONE=>2 THEN UNIT=.;
* DEFINE STRATA;
  STRATUM=PSEUDO/2 + .5;
  STRATUM=INT(STRATUM);
* SORT BY STRATUM AND PSU IF TAPE NOT IN SORT;
  PROC SORT;
       BY STRATUM PSEUDO;
* CALCULATE VARIANCE-COVARIANCE MATRIX;
  PROC SURREGR MISSPSU BETA DATAOUT;
  CLASSES UNIT FAMSIZE EDHEAD MDPA;
  LEVELS 1 4 5 3;
  MODEL MDADE2=UNIT*FAMSIZE*EDHEAD*MDPA/NOINT;
  STRATUM STRATUM;
  PSU PSEUDO;
  WEIGHT T21316;
  DATA _NULL_;
  SET DUMMYM01;
  FILE VAROUT;
  PUT (B001-B060) (4*E15.8/);
//STEP2 EXEC LARGEN,TIME=2
//FT05F001  DD  DSN=WYL.ACCT.MGK.MATRIX,DISP=(OLD,KEEP)
//SYSIN  DD  *
```

```
     5     3     5           1     0  YES MDMA,YES PHONE,SATURATED
    60    60     1                    (4E15.8)
     1                               (4E15.8)
     1     2   120     2     1        (2F2.0)
  1
 -1
     4                       1        (60F1.0/60F1.0)
010101010101010101010101010101010101010101010101010101010101010101010101
010101010101010101010101010101010101010101010101010101010101010101010101
     2                       1
     1     2    60     1     1        (2F2.0)
  1-1
     7     2    60    15     1     0  (15F2.0)        11 Saturated FAMSIZE
  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
  1  1  1  0  0  0  0  0  0  0  0  0-1-1-1
  0  0  0  1  1  1  0  0  0  0  0  0-1-1-1
  0  0  0  0  0  0  1  1  1  0  0  0-1-1-1
  0  0  0  0  0  0  0  0  0  1  1  1-1-1-1
  1  0-1  1  0-1  1  0-1  1  0-1  1  0-1
  0  1-1  0  1-1  0  1-1  0  1-1  0  1-1
  1  0-1  0  0  0  0  0  0  0  0  0-1  0  1
  0  0  0  1  0-1  0  0  0  0  0  0-1  0  1
  0  0  0  0  0  0  1  0-1  0  0  0-1  0  1
  0  0  0  0  0  0  0  0  0  1  0-1-1  0  1
  0  1-1  0  0  0  0  0  0  0  0  0  0-1  1
  0  0  0  0  1-1  0  0  0  0  0  0  0-1  1
  0  0  0  0  0  0  0  1-1  0  0  0  0-1  1
  0  0  0  0  0  0  0  0  0  0  1-1  0-1  1
     8     1     4                    (60F1.0)        FAMSIZE
1
0000000000000001
00000000000000000000000000000001
0000000000000000000000000000000000000000000000001
     8     1    16                    (60F1.0)        EDHEAD
01
001
0001
00001
0000000000000001
0000000000000001
0000000000000001
0000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
0000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000001
     8     1     8                    (60F1.0)        MDPA
000001
0000001
0000000000000000001
```

```
0000000000000000000001
00000000000000000000000000000000000001
00000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000001
    8    1    8                    (60F1.0)          INTERACTION 1
00000001
000000001
0000000001
00000000001
000000000001
0000000000001
00000000000001
000000000000001
    8    1    8                    (60F1.0)          INTERACTION 2
00000000000000000000001
00000000000000000000001
000000000000000000000001
0000000000000000000000001
00000000000000000000000001
000000000000000000000000001
0000000000000000000000000001
00000000000000000000000000001
    8    1    8                    (60F1.0)          INTERACTION 3
000000000000000000000000000000000000001
000000000000000000000000000000000000001
0000000000000000000000000000000000000001
00000000000000000000000000000000000000001
000000000000000000000000000000000000000001
0000000000000000000000000000000000000000001
00000000000000000000000000000000000000000001
    8    1    8                    (60F1.0)          INTERACTION 4
0000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000001
00000000000000000000000000000000000000000000000000000001
000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000001
//
```

For reanalysis of the data when the variance-covariance matrix is on disk, the first lines are the following;

```
//  JOB card(s)    TIME=2,REGION=1008K
/*SETUP        KOVAR,2,,,INP=NONE
// EXEC LARGEN,TIME=1
//FT05F0001    DD  DSN=WYL.DHES.MGK.MATRIX,DISP=(OLD,KEEP)
```

```
//SYSIN    DD  *
```

The final program for children with phones whose mothers had made a visit for medical care within a year is shown below. The logit transformation is made at the beginning of the program. The transformation from the Beta values to the odds ratios is the first reanalysis. The transformation from the Beta values to the predicted proportions is the second reanalysis.

EXAMPLE THREE

YES MDMA, YES PHONE LOGIT

FINAL NESTED MODEL

```
//  JOB card(s)  TIME=2,REGION=1008k
/*SETUP         KOVAR,2,,,INP=NONE
//STEP1 EXEC LARGEN,TIME=2
//FT05F001  DD  DSN=WYL.ACCT.MGK.MATRIX,DISP=(OLD,KEEP)
//SYSIN  DD  *
     5     3     5          1     0  YES MDMA,YES PHONE,REDUCED
    60    60     1                   (4E15.8)
     1                               (4E15.8)
     1     2   120     2     1       (2F2.0)
 1
-1
     4                      1        (60F1.0/60F1.0)
0101010101010101010101010101010101010101010101010101010101010101
0101010101010101010101010101010101010101010101010101010101010101
     2                      1
     1     2    60     1     1       (2F2.0)
 1-1
     7     1    44           1     1  (30F2.0/30F2.0)          11 FINAL
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

 1 1 1 0 0 0 0 0 0 0 0 0-1-1-1

 0 0 0 1 1 1 0 0 0 0 0 0-1-1-1

 0 0 0 0 0 0 1 1 1 0 0 0-1-1-1

 0 0 0 0 0 0 0 0 0 1 1 1-1-1-1
```

1 0-1 1 0-1 1 0-1 1 0-1 1 0-1

0 1-1 0 1-1 0 1-1 0 1-1 0 1-1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 0 0 0 0 0 0 0 0 0-1-1-1

0 0 0 1 1 1 0 0 0 0 0 0-1-1-1

0 0 0 0 0 0 1 1 1 0 0 0-1-1-1

0 0 0 0 0 0 0 0 0 1 1 1-1-1-1

1 0-1 1 0-1 1 0-1 1 0-1 1 0-1

0 1-1 0 1-1 0 1-1 0 1-1 0 1-1

1 0-1 0 0 0 0 0 0 0 0 0-1 0 1

0 0 0 1 0-1 0 0 0 0 0 0-1 0 1

0 0 0 0 0 0 1 0-1 0 0 0-1 0 1

0 0 0 0 0 0 0 0 0 1 0-1-1 0 1

0 1-1 0 0 0 0 0 0 0 0 0 0-1 1

0 0 0 0 1-1 0 0 0 0 0 0 0-1 1

0 0 0 0 0 0 0 1-1 0 0 0 0-1 1

0 0 0 0 0 0 0 0 0 0 0 1-1 0-1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 0 0 0 0 0 0 0 0 0-1-1-1

0 0 0 1 1 1 0 0 0 0 0 0-1-1-1

0 0 0 0 0 0 1 1 1 0 0 0-1-1-1

0 0 0 0 0 0 0 0 0 1 1 1-1-1-1

1 0-1 1 0-1 1 0-1 1 0-1 1 0-1

0 1-1 0 1-1 0 1-1 0 1-1 0 1-1

1 0-1 0 0 0 0 0 0 0 0 0-1 0 1

0 0 0 1 0-1 0 0 0 0 0 0-1 0 1

```
 0 0 0 0 0 0 1 0-1 0 0 0-1 0 1

 0 0 0 0 0 0 0 0 0 1 0-1-1 0 1

 0 1-1 0 0 0 0 0 0 0 0 0 0-1 1

 0 0 0 0 1-1 0 0 0 0 0 0 0-1 1

 0 0 0 0 0 0 0 1-1 0 0 0 0-1 1

 0 0 0 0 0 0 0 0 0 0 0 1-1 0-1 1

                                  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

                                  1 1 1 0 0 0 0 0 0 0 0 0-1-1-1

                                  0 0 0 1 1 1 0 0 0 0 0 0-1-1-1

                                  0 0 0 0 0 0 1 1 1 0 0 0-1-1-1

                                  0 0 0 0 0 0 0 0 0 0 1 1 1-1-1-1

                                  1 0-1 1 0-1 1 0-1 1 0-1 1 0-1

                                  0 1-1 0 1-1 0 1-1 0 1-1 0 1-1
    8     1     4                   (44F1.0)          FAMSIZE
1
00000001
00000000000000000001
00000000000000000000000000000000000001
    8     1    16                    (44F1.0)          EDHEAD
01
001
0001
00001
000000001
0000000001
00000000001
000000000001
00000000000000000001
00000000000000000001
00000000000000000001
00000000000000000001
00000000000000000000000000000000000001
00000000000000000000000000000000000001
00000000000000000000000000000000000001
00000000000000000000000000000000000001
    8     1     8                   (44F1.0)          MDPA
000001
0000001
00000000000001
00000000000001
00000000000000000000000000001
00000000000000000000000000001
```

```
00000000000000000000000000000000000000000001
00000000000000000000000000000000000000000001
    8    1    8                    (44F1.0)          INTERACTION 2
0000000000000001
0000000000000001
0000000000000001
0000000000000001
0000000000000001
0000000000000001
0000000000000001
0000000000000001
    8    1    8                    (44F1.0)          INTERACTION 3
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
00000000000000000000000000000001
    8    1    4                    (44F1.0)          EDHEAD IN 1
01
001
0001
00001
    8    1    4                    (44F1.0)          EDHEAD IN 2
000000001
0000000001
00000000001
000000000001
    8    1    4                    (44F1.0)          EDHEAD IN 3
00000000000000000000001
00000000000000000000001
00000000000000000000001
00000000000000000000001
    8    1    4                    (44F1.0)          EDHEAD IN 4
00000000000000000000000000000000000001
00000000000000000000000000000000000001
00000000000000000000000000000000000001
00000000000000000000000000000000000001
    8    1    2                    (44F1.0)          MDPA IN 1
000001
0000001
    8    1    2                    (44F1.0)          MDPA in 2
00000000000001
00000000000001
    8    1    2                    (44F1.0)          MDPA IN 3
00000000000000000000000000001
00000000000000000000000000001
    8    1    2                    (44F1.0)          MDPA IN 4
00000000000000000000000000000000000000000001
00000000000000000000000000000000000000000001
    8    1    1                    (30F2.0/14F2.0)   FAMSIZE 1 VS 2
 1  0  0  0  0  0  0-1
```

```
   8     1     1                    (30F2.0/14F2.0)  FAMSIZE 2 VS 3
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0-1

   8     1     1                    (30F2.0/14F2.0)  FAMSIZE 3 VS 4
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0-1
   6
   3                      1
   6
   1     1    60          1     0  (22F2.0/22F2.0)

1 1 0 0 0 1 0

1 1 0 0 0 0 1

1 1 0 0 0-1-1

1 0 1 0 0 1 0

1 0 1 0 0 0 1

1 0 1 0 0-1-1

1 0 0 1 0 1 0

1 0 0 1 0 0 1

1 0 0 1 0-1-1

1 0 0 0 1 1 0

1 0 0 0 1 0 1

1 0 0 0 1-1-1

1-1-1-1-1 1 0

1-1-1-1-1 0 1

1-1-1-1-1 1-1

                1 1 0 0 0 1 0 1 0 0 0 0 0 0 0

                1 1 0 0 0 0 1 0 0 0 0 1 0 0 0

                1 1 0 0 0-1-1-1 0 0 0-1 0 0 0

                1 0 1 0 0 1 0 0 1 0 0 0 0 0 0

                1 0 1 0 0 0 1 0 0 0 0 0 1 0 0

                1 0 1 0 0-1-1 0-1 0 0 0-1 0 0
```

1 0 0 1 0 1 0 0 0 1 0 0 0 0 0

1 0 0 1 0 0 1 0 0 0 0 0 0 1 0

1 0 0 1 0-1-1 0 0-1 0 0 0-1 0

1 0 0 0 1 1 0 0 0 0 1 0 0 0 0

1 0 0 0 1 0 1 0 0 0 0 0 0 0 1

1 0 0 0 1-1-1 0 0 0-1 0 0 0-1

1-1-1-1-1 1 0-1-1-1-1 0 0 0 0

1-1-1-1-1 0 1 0 0 0 0-1-1-1-1

1-1-1-1-1-1-1 1 1 1 1 1 1 1 1

1 1 0 0 0 1 0 1 0 0 0 0 0 0 0

1 1 0 0 0 0 1 0 0 0 0 1 0 0 0

1 1 0 0 0-1-1-1 0 0 0-1 0 0 0

1 0 1 0 0 1 0 0 1 0 0 0 0 0 0

1 0 1 0 0 0 1 0 0 0 0 0 1 0 0

1 0 1 0 0-1-1 0-1 0 0 0-1 0 0

1 0 0 1 0 1 0 0 0 1 0 0 0 0 0

1 0 0 1 0 0 1 0 0 0 0 0 0 1 0

1 0 0 1 0-1-1 0 0-1 0 0 0-1 0

1 0 0 0 1 1 0 0 0 0 1 0 0 0 0

1 0 0 0 1 0 1 0 0 0 0 0 0 0 1

1 0 0 0 1-1-1 0 0 0-1 0 0 0-1

1-1-1-1-1 1 0-1-1-1-1 0 0 0 0

1-1-1-1-1 0 1 0 0 0 0-1-1-1-1

1-1-1-1-1-1-1 1 1 1 1 1 1 1 1

1 1 0 0 0 1 0

1 1 0 0 0 0 1

1 1 0 0 0-1-1

```
                                        1 0 1 0 0 1 0
                                        1 0 1 0 0 0 1
                                        1 0 1 0 0-1-1
                                        1 0 0 1 0 1 0
                                        1 0 0 1 0 0 1
                                        1 0 0 1 0-1-1
                                        1 0 0 0 1 1 0
                                        1 0 0 0 1 0 1
                                        1 0 0 0 1-1-1
                                        1-1-1-1-1 1 0
                                        1-1-1-1-1 0 1
                                        1-1-1-1-1-1-1
    3                       1
    1    2   120    2       1        (2F2.0)
 1
 1
    4                       1        (60F1.0/60F1.0)
010101010101010101010101010101010101010101010101010101010101010101
010101010101010101010101010101010101010101010101010101010101010101
    2                       1
    1    2    60    1       1        (2F2.0)
 1-1
    3
//
```

The final program for children without phones is shown below. The logit transformation is made at the beginning of the program. The transformation from the Beta values to the odds ratios is the first reanalysis. The transformation from the Beta values to the predicted proportions is the second reanalysis.

It is given with the input for the variance-covariance

program so that the reader can see the specifications. The final model was actually run from the matrix on disk.

EXAMPLE FOUR

CHILDREN WITHOUT PHONES

FINAL MODEL

```
//  JOB card(s)   TIME=15,REGION=1008K
//STEP1 EXEC SAS,TIME.SAS=30,REGION.SAS=300K
//SAS.WORK  DD UNIT=3350,SPACE=(TRK,(480,160),RLSE)
//CHILD  DD DSN=STUDY,UNIT=3400-6,DISP=(OLD,KEEP),
// VOL=SER=000000
//VAROUT  DD DSN=WYL.ACCT.MGK.MATRIX,DISP=(NEW,CATLG),UNIT=3350,
// SPACE=(TRK,(15,5),RLSE),DCB=RECFM=FB,VOL=SER=WYL101
//SYSIN DD *
  DATA;
  SET CHILD.STUDY;
  KEEP MDADE2 PHONE UNIT FAMSIZE MDMA EDHEAD MDPA
       T21316 STRATUM PSEUDO;
* ALL UNKNOWNS CHANGED TO UNIT UNKNOWN AND
  KEPT FOR VARIANCE COMPUTATIONS;
* DEFINE POPULATION;
* NO PHONE;
  IF MDMA = > 2 THEN  MDMA = 2 ;
  IF PHONE < 2 THEN UNIT = . ;
* REDUCE LEVELS OF VARIABLES;
  IF EDHEAD = > 2 THEN EDHEAD = 2 ;
  IF FAMSIZE = > 2 THEN FAMSIZE = 2 ;
* DEFINE STRATA;
  STRATUM=PSEUDO/2 + .5;
  STRATUM=INT(STRATUM);
* SORT BY STRATUM AND PSU IF TAPE NOT IN SORT;
  PROC SORT;
       BY STRATUM PSEUDO;
* CALCULATE VARIANCE-COVARIANCE MATRIX;
  PROC SURREGR MISSPSU BETA DATAOUT;
  CLASSES UNIT MDMA FAMSIZE EDHEAD MDPA;
  LEVELS 1 2 2 2 3;
  MODEL MDADE2=UNIT*FAMSIZE*EDHEAD*MDPA/NOINT;
  STRATUM STRATUM;
  PSU PSEUDO;
  WEIGHT T21316;
  DATA _NULL_;
  SET DUMMYM01;
  FILE VAROUT;
  PUT (B001-B024) (4*E15.8/);
//STEP2 EXEC LARGEN,TIME=1
```

```
//FT05F001  DD DSN=WYL.ACCT.MGK.MATRIX,DISP=(OLD,KEEP)
//SYSIN  DD *
     5     3     5          1     0  NO PHONE,LOGIT,NESTED
    24    24     1                   (4E15.8)
     1                               (4E15.8)
     1     2    48     2     1       (2F2.0)
 1
-1
     4                      1        (48F1.0)
01010101010101010101010101010101010101010101010101010101010101
     2                      1
     1     2    24     1             (2F2.0)
 1-1
     7     1    12          1     1  (24F2.0)   NO PHONE,FINAL MODEL
 1  1  1  1  1  1  1  1  1  1  1  1
 1  1  1  1  1  1-1-1-1-1-1-1-1-1
 1  1  1-1-1-1  1  1  1-1-1-1-1-1
 1  0-1  1  0-1  1  0-1  1  0-1  1
 0  1-1  0  1-1  0  1-1  0  1-1  0
                          1  1  1  1  1  1  1  1  1  1  1  1  1
                          1  1  1  1  1  1-1-1-1-1-1-1-1-1-1-1
                          1  1  1-1-1-1  1  1  1-1-1-1-1-1-1-1
                          1  0-1  1  0-1  1  0-1  1  0-1  1  0-1
                          0  1-1  0  1-1  0  1-1  0  1-1  0  1-1
                          1  0-1-1  0  1  1  0-1-1  0  1  1  0-1
                          0  1-1  0-1  1  0  1-1  0-1  1  0  1-1
     8     1     2                   (12F1.0)          MDMA
1
00001
     8     1     2                   (12F1.0)          FAMSIZE
01
0000001
     8     1     2                   (12F1.0)          EDHEAD
001
00000001
     8     1     4                   (12F1.0)          MDPA
0001
00001
000000001
0000000001
     8     1     2                   (12F1.0)          EDHEAD BY MDPA
00000000001
000000000001
     8     1     1                   (12F1.0)          FAMSIZE IN MDMA 1
01
     8     1     1                   (12F1.0)          FAMSIZE IN MDMA 2
0000001
     8     1     1                   (12F1.0)          EDHEAD IN MDMA 1
001
     8     1     1                   (12F1.0)          EDHEAD IN MDMA 2
00000001
     8     1     2                   (12F1.0)          MDPA IN MDMA 1
0001
00001
```

```
     8     1     2                    (12F1.0)          MDPA IN MDMA 2
000000001
0000000001
     8     1     1                    (12F2.0)          MDMA 1 VS MDMA 2
 1 0 0 0 0-1
     6
     3
     6
     1     1    24          1     0   (12F2.0)
 1 1 1 1 0
 1 1 1 0 1
 1 1 1-1-1
 1 1-1 1 0
 1 1-1 0 1
 1 1-1-1-1
1-1 1 1 0
1-1 1 0 1
1-1 1-1-1
1-1-1 1 0
1-1-1 0 1
1-1-1-1-1
            1 1 1 1 0 1 0
            1 1 1 0 1 0 1
            1 1 1-1-1-1-1
            1 1-1 1 0-1 0
            1 1-1 0 1 0-1
            1 1-1-1-1 1 1
            1-1 1 1 0 1 0
            1-1 1 0 1 0 1
            1-1 1-1-1-1-1
            1-1-1 1 0-1 0
            1-1-1 0 1 0-1
            1-1-1-1-1 1 1
     3                        1
     1     2    48     2     1          (2F2.0)
 1
 1
     4                        1          (24F1.0/24F1.0)
010101010101010101010101
010101010101010101010101
     2
     1     2    24     1     1          (2F2.0)
 1-1
     3
//
```

Some Comments on Computer Time

The computer time needed to calculate the variance-covariance

matrix is much greater than the time required to execute the LARGEN or GENCAT programs. While the statistician who knows precisely where he or she is going may be able to anticipate the entire analysis and utilise the reanalysis option in the programs, I find it much safer and more efficient to put the matrix on disk to have it available for the reanalyses. Some indications of time on the NCHS computer are as follows;

	CPU Time	
	SURREGR	LARGEN
Table 4.8	6 min	9 sec
Table 4.11	14 min	149 sec
Table 4.14	14 min	136 sec
Table 4.17	6 min	12 sec

References

Forthofer,R.N. and R.G.Lehnen.Public Program Analysis: A New Categorical Data Approach. Lifetime Learning Publications, Belmont CA, 1981.

Holt,M.M. SURREGR: Standard errors of regression coeffficients from sample survey data. Research Triangle Institute. May 1977.

Shah,B. modification of Holt,M.M. SURREGR: Standard errors of regression coefficients from sample survey data. Research Triangle Institute. April 1982.

Landis,J.R.,J.L.Freeman,W.M.Stanish,G.G.Koch, and A.L.Lewis. GENCAT: A computer Program for the Generalized Leaast Squares Analysis of Multivariate Categorical Data. Proc.Stat.Comp.Sec. of the American Statistical Association,1976.

Landis,J.R.,W.M.Stanish, and G.G.Koch.A Computer Program for the Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares to Compute Wald Statistics (GENCAT). Documentation for the GENCAT programs obtainable from Landis, University of Michigan, School of Public Health, Department of Biostatistics, Ann Arbor, MI 48109.

On-Line Business Systems,Inc. OBS WYLBUR Reference Manual. San Francisco,CA. 1981.