# Determinants of
# Total Family Charges
# for Health Care:
# United States, 1980

Series C, Analytical Report No. 8
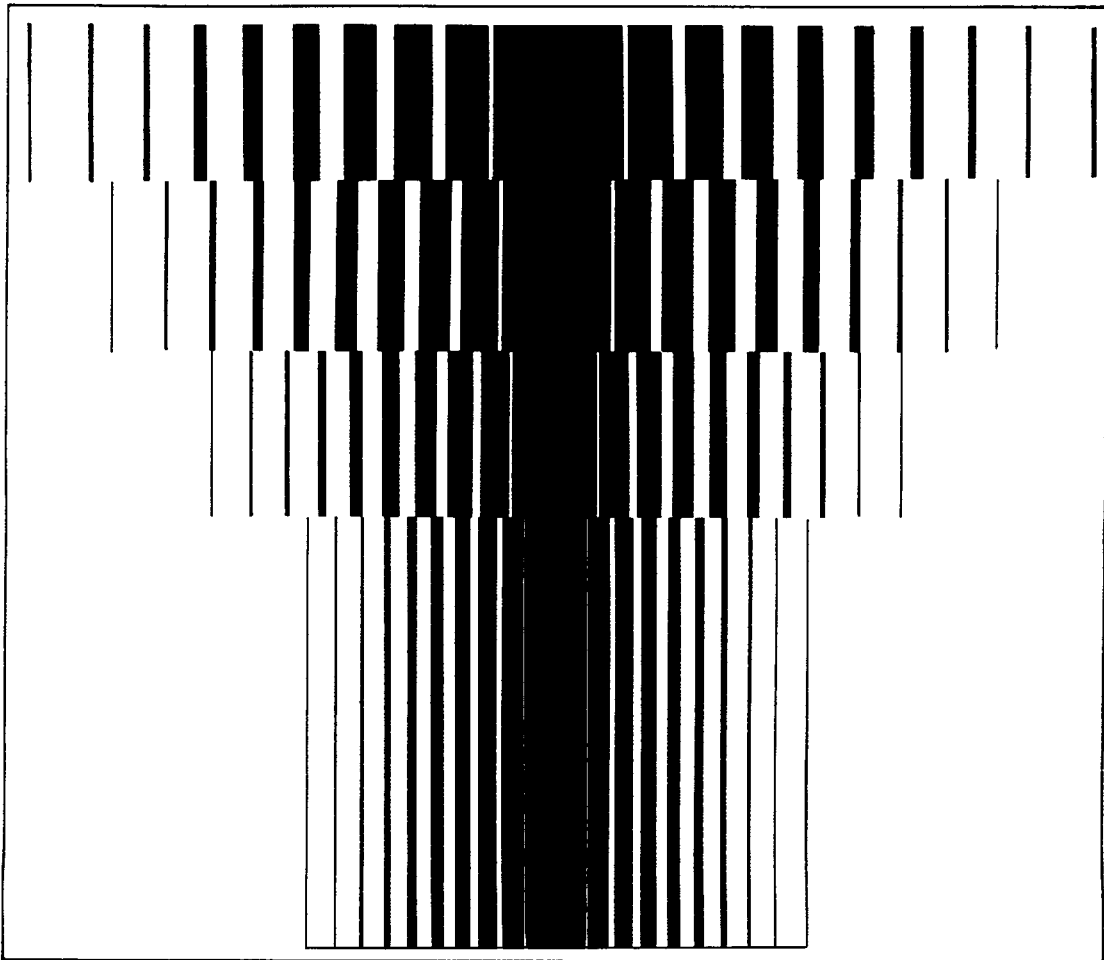
**Copyright Information**

All material appearing in this report is in the public domain
and may be reproduced or copied without permission; citation
as to source, however, is appreciated.

## National Medical Care Utilization and Expenditure Survey

The National Medical Care Utilization and Expenditure Survey (NMCUES) is a unique source of detailed national estimates on the utilization of and expenditures for various types of medical care. NMCUES is designed to be directly responsive to the continuing need for statistical information on health care expenditures associated with health services utilization for the entire U.S. population.

NMCUES will produce comparable estimates over time for evaluation of the impact of legislation and programs on health status, costs, utilization, and illness-related behavior in the medical care delivery system. In addition to national estimates for the civilian noninstitutionalized population, it will also provide separate estimates for the Medicaid-eligible populations in four States.

The first cycle of NMCUES, which covers calendar year 1980, was designed and conducted as a collaborative effort between the National Center for Health Statistics, Public Health Service, and the Office of Research and Demonstrations, Health Care Financing Administration. Data were obtained from three survey components. The first was a national household survey and the second was a survey of Medicaid enrollees in four States (California, Michigan, Texas, and New York). Both of these components involved five interviews over a period of 15 months to obtain information on medical care

utilization and expenditures and other health-related information. The third component was an administrative records survey that verified the eligibility status of respondents for the Medicare and Medicaid programs and supplemented the household data with claims data for the Medicare and Medicaid populations.

Data collection was accomplished by Research Triangle Institute, Research Triangle Park, N.C., and its subcontractors, the National Opinion Research Center of the University of Chicago, Ill., and SysteMetrics, Inc., Berkeley, Calif., under Contract No. 233–79–2032.

Co-Project Officers for the Survey were Robert R. Fuchsberg of the National Center for Health Statistics (NCHS) and Allen Dobson of the Health Care Financing Administration (HCFA). Robert A. Wright of NCHS and Larry Corder of HCFA also had major responsibilities. Daniel G. Horvitz of Research Triangle Institute was the Project Director primarily responsible for data collection, along with Associate Project Directors Esther Fleishman of the National Opinion Research Center, Robert H. Thornton of Research Triangle Institute, and James S. Lubalin of SysteMetrics, Inc. Barbara Moser of Research Triangle Institute was the Project Director primarily responsible for data processing.

# Contents

**List of Text Tables**

## Symbols

| | |
|---|---|
| - | No families with these characteristics in sample |
| * | Potential reliability problem; statistic is based on sample size of fewer than 50 or has relative standard error greater than 30 percent |
| . . . | Category not applicable |

# Determinants of Total Family Charges for Health Care: United States, 1980

By Jonathan H. Sunshine, Ph.D.,
Sunshine and Associates,
and Marvin Dicker, Ph.D.,
Division of Applied Research,
National Institute on Drug Abuse
(formerly of the National Center for Health Statistics)

## Executive Summary

This report addresses a question of importance for policymakers: "What are the determinants of the total charges for health care that U.S. families face?" Policymakers' concerns about this question have two main grounds. First, U.S. health care costs are large and growing rapidly. They now exceed 11 percent of the gross national product, and the answer to the question can shed some light on their troubling growth. Second, total family charges for health care reflect the quantity of health care received by families, and it is important to know whether the determinants of total charges are principally the need for health care, or involve other factors less related to need.

In this report, the determinants of total charges and their importance are identified principally through multiple regression analysis. Total charges are defined as the full amount charged for all types of health care for all family members regardless of whether these amounts are paid out of pocket, paid by insurance (or public health care coverage programs), or go unpaid. The data used are from the family data files of the 1980 National Medical Care Utilization and Expenditure Survey (NMCUES). This report presents data on the approximately 5,000 multiple-person families interviewed in this year-long longitudinal survey. The report provides a separate analysis for each of three socioeconomic family populations that have consistently been of interest to policymakers. These are (1) older families (defined for this report as all U.S. multiple-person families with a member 65 years of age or over); (2) younger, lower income families (all U.S. multiple-person families below 200 percent of the poverty level in 1980 and with all members under 65 years of age); and (3) younger, better off families (all U.S. multiple-person families at 200 percent of the poverty level or higher in 1980 and with all members under 65).

Multiple regression analysis was used to investigate the effect on total family charges of family demographic and sociocultural characteristics, family illnesses, special health events (such as births, deaths, and hospitalizations of family members), general family health status, family income, family health insurance characteristics, and family geographic and urbanization characteristics. Regressions were run separately for each of the three socioeconomic family populations, with total family charges as the dependent variable and approximately 45 variables measuring these family characteristics as independent variables. Because of the large number of independent variables involved, a multiple-step regression process (described in Appendix I) was used. The statistical significance of the findings was assessed by using a regression program (SURREGR) that takes into account the complex sample design of NMCUES and by using an $F$-test that was analogous to a multiple $t$-test.

The regressions had a particularly high explanatory power, with an $R^2$ (the proportion of the variance in the dependent variable explained by the independent variables) of 0.57 to 0.72, depending on the family population involved. Similar studies have previously reported $R^2$ statistics of 0.04 to 0.53.

Hospitalization emerged as the most important single determinant of total family charges. It showed a uniquely strong statistical significance in all three family populations, and, in all three, families with one hospitalization had total charges about four times as large as did otherwise similar families with no hospitalizations, even though the similar families had the same pattern of illnesses and other characteristics. In addition, each further hospitalization increased total family charges by approximately 30 percent in each of the three family populations.

In general, health variables were found to be the most important determinants of total family charges. Like hospitalization, family illness days in bed was statistically significant in all three family populations and had a broadly similar effect in all. (Each 1 percent increase in the quantity (family illness days in bed plus 1) increased total family charges by approximately 0.1 to 0.2 percent, other family characteristics being equal.) Three family illness variables—(1) a family member having cancer, (2) a family member having heart or

circulatory disease, and (3) a family member suffering from an accident, injury, or poisoning—were each significant in two of the family populations. The size of their effects sometimes differed among family populations, suggesting differing specific illnesses (each category encompasses several specific illnesses) or severity among the family populations. However, as would be expected, the effects of these illnesses were always to increase total charges. The increases were on the order of 20–60 percent after controlling for the effects of all other variables in the analysis. (In particular, note that this is the measured increase after controlling for the large effects of hospitalization described in the preceding paragraph.) Perceived health status was also found significant in two family populations, with worse health leading to increased total charges. Health variables significant for one family population were family work-loss days due to illness and limitation in major activity. For both, worse health led to higher total charges.

Some nonhealth variables were significant determinants of total family charges for health care, but they were significant for only one or sometimes two family populations. They included family income, significant in two populations, with each 1 percent increase in income leading to a 0.15–0.26 percent increase in total charges (other family characteristics being equal), and region, also significant in two populations with, however, different geographic regions atypical in total charges for different populations. Completeness of health care coverage was also significant for two family populations, with both populations, as expected, showing smaller total charges for families with incomplete coverage. The presence of a child in a family and the age, race, and education of the family head were each found significant in one family population—the population of younger, better off multiple-person families.

The predominance of health variables among the determinants of total family charges is reassuring. It suggests that overall in the United States, the need for health care is the most important factor in determining how much care families receive.

This finding contrasts strongly with results of a parallel regression analysis (Dicker and Sunshine, 1988) that examined determinants of the financial burden of health care costs on U.S. families. That analysis measured the financial burden of a family's health care costs as the ratio of its total out-of-pocket expenses for health (including family-paid insurance premiums) to its income. It found that income and health care coverage were the primary determinants of the burden. Hospitalization, in contrast, had little effect on this measure of financial burden (called the "financial burden index"), and NMCUES data provide an explanation of why this was so. The NMCUES data show that costs associated with hospitalization were extensively insured, so much so that a dollar of charges for inpatient care resulted in only one-fifth as much out-of-pocket spending (the spending included in the financial burden index) as did a dollar of charges for other kinds of health care.

Other comparisons of the determinants of the financial burden index with the determinants of total family charges are also illuminating. For example, a comparison shows that among younger, lower income families, out-of-pocket expenses increased relatively rapidly as income increased while total charges (which presumably approximate care received) did not increase significantly with increasing income. This raises questions of equity. As well, it may generate work disincentives. The work disincentives arise because increased income, which would be achieved by more work, led to no more health care and to increased out-of-pocket expenses. The increased out-of-pocket expenses cut into the increase in discretionary income available from more work, making the financial rewards of working relatively small.

The equity concern is reinforced by the finding of a reverse pattern for younger, better off families. Among these families, total charges (and presumably care) increased more rapidly in percentage terms as income rose than did out-of-pocket expenses, meaning that the percentage of costs paid out of pocket decreased with increasing income. Equity concerns raised by these findings are compounded by the equity issue raised by findings concerning income in the regression analyses of the financial burden index: Income was one of the two most important determinants of the index (the other was completeness and type of health insurance coverage), and the financial burden of health costs was greater for lower income families than for better off families. Equity concerns also arise from findings that black families had total charges about 30 percent lower than comparable white families and presumably received correspondingly less health care.

A comparison of findings of the two sets of regression analyses regarding the effects of incomplete health care coverage yields somewhat unexpected results. Incomplete coverage was associated with lower total charges, as is widely reported in the literature, where it is found to result from (1) decreased use of care because families with incomplete coverage face higher out-of-pocket costs for many health care services and (2) a tendency of those who use little health care to obtain little coverage. However, the regressions also showed incomplete coverage to be associated with lower out-of-pocket costs, which implies that the lesser use of care by families with incomplete coverage (compared with families with complete coverage) more than offset the higher out-of-pocket costs for each service faced by the former families. In contrast, the literature generally reports higher out-of-pocket costs for families with incomplete coverage and a much smaller reduction in the use of care by these families.

The finding that hospitalization, rather than some health status variable(s), was the most important single determinant of total family charges may indicate that in 1980 hospitalization was sometimes a high-cost modality of health care rather than a necessary response to severe health problems. This interpretation draws support from the finding that there was no effect on total

charges of the death of a family member, for death is often a reflection of an extremely severe illness. The interpretation is also supported by the major decline in hospitalization rates that occurred after 1980 in response to cost containment measures. The decline shows that hospitalization, as carried out in 1980, was not always necessary.

# Introduction

## Overview

A consistent goal of contemporary U.S. health policy has been adequate health care at an affordable cost. In pursuing this goal, recent emphasis has been on relieving financially burdensome health care expenses among U.S. families (Catastrophic Illness Expenses, 1986). However, there has been a lack of adequate data on which to base national policy initiatives. This report attempts to supply some of the needed data. It is the sixth in a series of reports from the 1980 longitudinal National Medical Care Utilization and Expenditure Survey (NMCUES) that have presented data on health care expenses and the use of health care by U.S. families (Dicker, 1983a; Dicker and Sunshine, 1987; Sunshine and Dicker, 1987a; Sunshine and Dicker, 1987b; Dicker and Sunshine, 1988).

Although many aspects of the U.S. health care system can be studied using the individual as the unit of analysis, the family is the more appropriate unit for studying financially burdensome expenses. This is because the financial burden of health care, and decisions concerning the use and financing of health care, usually are family responsibilities rather than individual responsibilities.

Given that the family should be the unit of analysis, the NMCUES is a particularly good data source, as it was originally conceptualized with family analysis in mind. It has a distinct and carefully conceptualized method of longitudinal family construction, a collection of specially developed longitudinal and cross-sectional family variables, and a family public use data tape that will allow other researchers to carry further the research presented here (Public Use Data Tape Documentation: Family Data, 1986).

There are multiple ways of measuring the financial burden of family health expenses. Dicker and Sunshine, in a companion report (1988), discuss the uses of six different measures of financial burden and present data on each of them. Included among the six measures were two different measures of total charges for health care. The companion report focused on a detailed regression analysis of the determinants of the level of one of the six measures, which seemed to be the best measure of a family's ability to pay health care costs. This measure was called the financial burden index in the report. It was defined as the ratio of total family out-of-pocket expenses for health (including family-paid premiums) to family income.

The present report extends the analysis begun in the previous report by presenting a detailed investigation, also using regression analysis, of the determinants of levels of total family charges for health care. Total family charges for health care are defined as the annualized total amount billed for all types of health care services and health care supplies to all family members regardless of whether these charges are paid out of pocket, paid by health care coverage, or remain unpaid. (Sunshine and Dicker, 1987b, called this measure "total family expenditures for health care" and presented extensive frequency tables on it.) Total family charges are the underlying health expenses that families and societies face. Ways of paying for total charges and of limiting them are the building blocks with which a strategy to manage financially burdensome health expenses must be constructed. This report, therefore, complements the previous report, in which actual family financial burden was measured directly as the ratio of out-of-pocket expenses to family income. Here, the focus is on the potential financial burden created by total family charges. Comparing the determinants of the actual expenses that families pay (out-of-pocket expenses) and the potential expenses they could have faced (total charges) helps in better understanding the nature of financially burdensome health expenses.

To ensure comparability between the two reports, the regressions in this report use the same data base and the same independent variables as the previous companion report on the financial burden index (Dicker and Sunshine, 1988). These independent variables include sociodemographic, health-related, socioeconomic, and geographic factors that have been suggested as possibly contributing either directly or indirectly to high levels of total health care charges, out-of-pocket expenses, and family financial burden. The independent variable in this report, as previously stated, is total family charges for health care.

Because the focus of this report is on supplying data for use in making health policy, three populations of particular interest to policymakers are examined. These are (1) older families (defined as all U.S. multiple-person families with a member 65 years of age or over);

(2) younger, lower income families (defined as all U.S. multiple-person families with all members under 65 years of age and with characteristics (income, family size, and so forth) that placed them below 200 percent of the poverty level in 1980); and (3) younger, better off families (defined as all U.S. multiple-person families with all members under 65 years of age and with characteristics (income, family size, and so forth) that placed them at 200 percent of the poverty level or higher in 1980). Data were tabulated separately on all three family populations and a separate regression analysis was done for each population. The underlying assumption of this approach was that both the causes of and remedies for financially burdensome health care expenses may be population specific.

This report, in short, attempts to answer two questions: What are the determinants of high levels of total family charges for health care among different policy-relevant populations of multiple-person families, and how (and why) do these determinants of total family charges (that is, of the total health care expenses incurred by families) differ from the determinants of the financial burden index found in the previous study? This report differs from other reports on such subjects in its focus on multiple-person families, its examination of a larger and more varied collection of explanatory variables, its consistent controlling for age and family socioeconomic status through the use of separate populations for analysis, and its use of multiple regression analysis.

## Background

Wyszewianski (1986a) makes a distinction between the size of a family's health care expenses and a family's ability to pay the bill. This distinction underlies the major distinction among studies dealing with financially burdensome or catastrophic health care expenses. Studies that focus on the size of health care expenses usually measure financially burdensome expenses by the total charges (in dollars) for health care services. If these charges are large (above a certain threshold) they are considered to be catastrophic charges. Wyszewianski suggests that this approach reflects the most common concept of catastrophic expenses as involving illnesses with very large expenses. By contrast, studies that focus on a family's ability to pay the bill usually measure financially burdensome expenses by the ratio of out-of-pocket health care expenses to family economic resources (usually income). In these types of studies, families with relatively moderate or low total charges may be found to have burdensome health care expenses if the families are both poor and have inadequate or no health care coverage.

The companion report on the financial burden index (Dicker and Sunshine, 1988) presented a detailed review of selected family studies from both orientations. These included Koretz, 1982; Berki et al., 1985; Berki, 1986;

Wyszewianski, 1986b; and Catastrophic Illness Expenses, 1986. Of these studies, only Koretz (1982) assessed family financial burden by using total charges for health care, the topic of this report, as the measure of financial burden. Therefore, from among this group of studies, only findings of Koretz' study are reported here.

Koretz did not investigate the *determinants* of levels of total family charges for health care. However, some of his findings on the *distribution* of total charges are worth noting. In particular, two of the most important are:

1. Among the population sampled, families exceeding any given catastrophic threshold in a single year are relatively rare, but they account for a sizable proportion of total health expenses. For example, only 5 percent of the examined families exceeded $5,000 in total expenses in a given year, but those families accounted for half of all expenses for all families.

2. Although only a small proportion of families have catastrophic expenses in a given year, a much larger proportion have high expenses at least once over several years. For example, over a 3-year period, 27 percent of the families exceeded a $3,000 catastrophic threshold at least once, while in a single year only 11 percent did so.

All of the studies cited above (including Koretz, 1982) have common methodological characteristics that tend to limit the inferences that can be made from their findings. They all combine multiple-person families and one-person families into a single category and present statistics only for this combined category. Thus the reader cannot determine whether any of their findings are intrinsic to either one or the other type of family. Second, in keeping with the lack of consensus as to what level of charges or expenses or ratio of expenses to family income (depending on the measure used) constitutes a financial burden, they all use multiple thresholds. These multiple thresholds for financial burden leave it up to the reader to decide which threshold and which findings are the appropriate ones. Third, all these studies present distributions of categorical independent variables in tabular form. Therefore, they can control for only a few independent variables at one time. This makes it difficult to assess whether any apparent causal relationships they identify are artifacts of independent variables likely to be important but not included in the tables. It also severely limits precision in measuring the quantitative relationship between the independent variables and the dependent variable. Finally, none of the studies compares different measures of financial burden with each other. This makes it difficult to understand differences that are found when different measures are used.

This report and the previous companion report on the financial burden index (Dicker and Sunshine, 1988)

attempt to overcome the above limitations by adopting the following design features: First, both reports present data only on multiple-person families. These are the type of units conventionally thought of as "families" both by professionals concerned with the family and by the general public. Second, both categorical and continuous dependent variables are used. This allows for the examination of the distribution of families and subcategories of families with respect to specific thresholds, as in previous research, but it also makes possible a quantitative appraisal of the factors that determine levels of the dependent variables. Third, both reports use multiple regression techniques that simultaneously control for a large number of independent variables. Because such techniques automatically allow elimination of a host of plausible alternative hypotheses, they increase confidence that any relationship found is not the result of spurious correlations. Finally, both reports compare the results found for different measures of financial burden. These comparisons highlight the weaknesses and strengths of the different measures and show something of the dynamics of the formation of family financial burden.

The previous companion report included a descriptive section that compared distributions of U.S. families when several measures of financial burden were used. Two basic types of financial burden measures were compared. One type consisted of charges (or expenses) in dollars, which may be considered a measure of absolute financial burden. The other consisted of charges (or expenses) as a percent of family income, which may be considered a measure of relative financial burden. For each basic type of measure, both total charges for health care and out-of-pocket expenses for health were examined. Moreover, two measures of out-of-pocket expenses were compared. These were "total out-of-pocket health expenses" including family-paid health insurance premiums and "out-of-pocket expenses for health care services" excluding family-paid health insurance premiums. This latter is the measure of out-of-pocket expenses usually found in the literature. Thus, in total six measures were compared—three dollar measures and three ratio measures.

When, using the same dollar threshold, total charges for health care were compared in the report with out-of-pocket expenses for health care, a much higher percentage of U.S. multiple-person families was found to have financially burdensome expenses on the basis of the total charges measure. For example, 18 percent of families had total charges of $3,000 or more in 1980, compared with 2 and 3 percent of families, respectively, that reached this threshold for the two variants of out-of-pocket expenses. Thus, the total charge measure showed six to nine times as many families experiencing financial burden in 1980 as did the out-of-pocket measures. A similar pattern was found using different dollar thresholds and when controls were used for family income and age status categories.

When total charges for health care as a percent of income were compared with out-of-pocket expenses for health care as a percent of income, using the same percent-of-income threshold, again a higher percentage of U.S. multiple-person families was found to have financially burdensome expenses using the total charges measure. For example, 29 percent of all U.S. families had total charges for health care of 10 percent or more of their income in 1980, compared with 12 percent and 6 percent of families, respectively, that had a ratio of expenses to income this high for the two variants of out-of-pocket expenses. Thus the total charges as a percent-of-income measure showed approximately 2.5 to 5 times as many families experiencing financial burden as did measures of out-of-pocket expenses as a percent of income. The differences, however, were only approximately half as great as those found for dollar measures. As before, a similar pattern was found for percent-of-income measures when various thresholds were used and when controls were used for family income and age status.

Regardless of whether dollar measures or percent-of-income measures are used to measure financial burden, measures based on total charges tend to overestimate the proportion of U.S. families actually experiencing financially burdensome health care expenses. This is because few families actually pay the full amount of their total charges; out-of-pocket expenses are a family's actual costs. Of the two out-of-pocket measures examined, the one that better measures financial burden is total out-of-pocket expenses for health including family-paid premiums. This is because family-paid premiums are as real a cost to a family as out-of-pocket payments made directly for health care services. Indeed, some families can directly trade off higher premiums and the more comprehensive coverage they buy for lower out-of-pocket costs directly for health care services.

The companion report (Dicker and Sunshine, 1988) concluded that total out-of-pocket expenses for health as a percent of family income is the best single measure of a family's financial burden for health costs. This measure not only includes all of a family's out-of-pocket expenses for health, but also takes into account a family's ability to pay. As previously stated, this measure was named the financial burden index. It had not been previously used. However, it had face validity. At any given threshold, it produced estimates of the percent of families with financial burden that were between those for total charges and those for out-of-pocket expenses for health care services (excluding family-paid premiums).

The companion report used multiple regression analyses on three family populations to investigate and identify the factors that contributed to a high level of family financial burden as measured by this index. These regressions examined the effect of approximately 45 factors involving family social, economic, and demographic characteristics, family health and illness characteristics, family health insurance characteristics, and family

geographic and urbanization characteristics. The overall finding was that in 1980 the major determinants of financially burdensome family health expenses (as measured by the index) were family income and the completeness and type of health insurance coverage (or public health care coverage program) the family had. This was the case for each of the three family populations examined. In contrast, health status variables such as the general health status of family members, special health events (such as death or institutionalization of a family member), and illnesses that family members had (including major illnesses) were much less important as determinants of a family's health-related financial burden, as measured by the index. This relatively negative finding regarding the effect of general health status, special health events, and illnesses as determinants of health-related financial burden was surprising.

This present report, as previously stated, examines the effect in the same three populations of the same approximately 45 factors on the level of total family charges for health care. A brief, preliminary summary of findings on this subject was presented in Dicker and Sunshine (1988). This summary indicated that health-related variables were more important as determinants of total family charges for health care than as determinants of financially burdensome family health expenses (as these were measured by the financial burden index). This report presents the findings on total family charges for health care in complete form and at full length so that they may be more precisely compared with the detailed findings on the financial burden index presented in the previous report. Hopefully, this comparison will increase understanding of the relationship between, and value of, two of the major measures of health-related financial burden: out-of-pocket expenses as a percent of income and total charges for health care.

# Data and Methods

## Source of the Data

As previously pointed out, the data presented in this report are from the National Medical Care Utilization and Expenditure Survey (NMCUES). In NMCUES, information was collected on health problems, health care received, expenditures for care, health insurance, and related topics. Data were obtained throughout calendar year 1980 from a sample of the U.S. civilian noninstitutionalized population. NMCUES included both a national household sample encompassing approximately 6,800 families and four State Medicaid samples. All information in this report is based on the national household sample. Detailed technical information on the sample, on estimation procedures, and on measurement procedures can be found in Appendix II.

NMCUES differs from most surveys of health in that it was a panel (or longitudinal) survey. Altogether, there were either four or five interviews, approximately 3 months apart, that were conducted with each family in the sample from early 1980 to early 1981. In each interview, information on all family members was gathered, usually from a single family respondent.

## Definition of the Family

Because NMCUES is a longitudinal survey, covering an entire year, the important concept of *longitudinal family* was developed to deal with the facts that the composition of a family can change over time and that families come into and go out of existence over time. The concept of longitudinal family used in this report is presented in detail in Appendix II. Simplified, it is as follows:

At a point in time, a family is defined as a group of persons sharing a common household and related by blood, marriage, adoption, or a formal foster care relationship. An unmarried student 17–22 years of age living away from home is also considered a part of a family.

When an initially sampled family had a change in membership during 1980, the prechange and postchange groups were considered the same family if and only if the "majority" of members of the prechange group became members of the postchange group, and the "majority" of members of the postchange group had

previously been members of the prechange group. For the purpose of counting a "majority," persons moving into or out of the sample universe—namely, the universe of civilian noninstitutionalized persons residing in the United States—were omitted from the count. For example, persons who were born, or who had died, or who had moved into or out of an institution, or into or out of the military were omitted from the count.

## Standardization for Part-Year Families

One problem with analyzing data from a longitudinal survey is that some families enter and leave the survey universe during the time covered by the survey. This has two consequences. First, the number of different families in the longitudinal universe is larger than the number of families that would be found in a cross-sectional survey. Second, a number of families (about 12 percent in NMCUES) did not exist for the full survey year (Dicker and Casady, 1984).

If each family that ever existed during the year were treated equally as one unit, the count of families, which would be equal to the gross total number of distinct families that ever existed during the year, would be larger than the average number of families that existed at a single point in time (the average cross-sectional estimate). Also, if each family that ever existed during the year were treated as one unit, measures of health spending and use of health care by families would not be comparable, as some counts of family spending and use of care would be for a whole year and some for less than a whole year.

Consequently, the following standardizing procedures were chosen. The population of families was time-adjusted so that, for example, a family that had existed for only half a year was counted as only half a unit. Therefore, in this report the total number of families in any category represents the total number of family years for that category. (Alternatively, this can be considered the average number of families in that category at a point in time during the year 1980.) Moreover, the counts for any health behavior event were adjusted to represent annual rates for that event. For example, a family in the survey for half the year with $150 in total charges is represented as a half family year unit with total charges at an annual rate of $300 per year.

Because these concepts are awkward to use in writing, families will be generally discussed in the following text as if they represented one unit each, and the expenses will be discussed as if they were actual expenses rather than annual rates. It should be noted, however, that the term "family," as used in the text, means *family years* and that all health expenses are rates per family year.

## Definition of a Health Expense

Annualized total charges for a family's health care are the health expenses used in this report. In principle, this measure would include expenses for all types of health care. However, the actual analysis is limited by the type of expense data collected in NMCUES, which did not cover all types of health expenses. The data used in this report include charges for the following types of health services: inpatient hospital care, inpatient physician care, outpatient hospital and emergency room care, ambulatory physician care, dental care, acquisitions of prescription medicines, care from other independent medical providers (such as chiropractors, speech therapists, faith healers, and psychologists), and the acquisition of health care supplies and services (such as eyeglasses, orthopedic items, hearing aids, ambulance services, and diabetic items). In previous reports (Sunshine and Dicker, 1987a; Sunshine and Dicker, 1987b; Dicker and Sunshine, 1988), the measure used in this report has been labeled charges for "all health care combined." However, this measure does not include charges for nonprescription medicines, nursing homes, and other types of long-term care institutions.

## Adjustments to the Sample

As previously pointed out, this report covers only multiple-person families (defined as families with an average family size of 1.5 members or more during the survey year). This is the type of family that the general public and most social scientists mean when they use the concept "family." (See the discussion in Dicker and Casady, 1985.) Moreover, as the review of the literature indicates, this social unit has not been treated separately by most previous researchers examining financially burdensome health expenses. Also, to have included one-person families in the analysis would have meant having a separate analysis for that type of social unit. This would have excessively increased both the size of the report and the amount of time needed to complete it. Thus, one-person families (defined as families with an average size of less than 1.5 members during the survey year) were excluded from the analysis in this report.

The NMCUES family sample consisted of 4,888 responding multiple-person families. Of these, 43 (or 0.9 percent) were families with military heads. Because

NMCUES was a survey of the noninstitutionalized civilian population, another family member (usually the spouse) was imputed as the head of these families. This imputation produced many anomalies in the data (see Public Use Data Tape Documentation: Family Data, 1986, pp. 22–23). Consequently, it was decided to exclude these families from the analysis. This gave a basic sample of 4,845 multiple-person families that were unequivocally representative of the civilian family population of the United States.

Another adjustment to the sample was made because of the use of family income as an explanatory variable in this report and in the companion report on the financial burden index (Dicker and Sunshine, 1988).

Some families reported either a zero income or a very low income (defined as an income less than $1,000 or under 20 percent of the poverty level). For these families, reported income is probably not a good measure of the actual financial resources available, and some adjustment was necessary. Two types of adjustments have been used in the literature. The first imputes a minimum income for such families (Duan et al., 1982). The second leaves such families out of the analysis (Berki et al., 1985). Each has its advantages and disadvantages. This report follows Berki and leaves these families out of the analysis. Hence, 21 multiple-person families with reported zero or very low incomes (0.4 percent of all families) are excluded, giving a basic sample of 4,824 civilian families. Almost all these zero-income or very-low-income families are younger, lower income families; approximately 2 percent of such families were excluded. It is believed that this exclusion does not fundamentally distort the analysis presented here.

## The Two-Part Model

Another adjustment to the sample is the result of using a two-part model recommended by Duan et al. (1982) for analyzing the determinants of health spending (or of use of health services). The first part of this model identifies what distinguishes populations with health spending (or use of health services) from populations without health spending (or use of health services). The second part of the model identifies the determinants of the amount of spending or use in populations with spending or use. This model has been used in previous NMCUES reports on family health care spending and use of health services (see Dicker and Sunshine, 1987; Sunshine and Dicker, 1987a; Sunshine and Dicker, 1987b; Dicker and Sunshine, 1988).

The regression analyses in this report are directed at answering the question addressed by the second part of the model: "What are the causes of financially burdensome health expenses among families with expenses?" Following Duan et al. (pp. 20–24), the population for the regression analyses was, therefore, limited to families with positive (nonzero) health expenses. This procedure

also eliminated the need to impute dollar amounts for families with zero expenses in order to avoid having to calculate the logarithm of zero, which is undefined. Of the 4,824 families remaining in the sample after the adjustments discussed previously had been made, 91 families had zero total charges. These 91 families (about 1.9 percent of the adjusted sample) were excluded, leaving a total of 4,733 multiple-person, civilian families for the regression part of the study.

The exclusion of families with zero total charges from the regression analyses affected the three socioeconomic populations similarly, with 2–3 percent of each population being excluded. It is again believed that this exclusion does not seriously distort the analysis presented here.

## Regression Methods

Multiple regression analysis was used to identify the statistically significant determinants of total family charges and to estimate the effect of different family characteristics on these charges. Appendix I presents a technical description of the analytic procedures followed. For the reader's convenience, a summary of these procedures follows.

Multiple regression analysis is a statistical technique for estimating the effect on a single dependent variable of each of a set of independent (or causal) variables. The effect of each independent variable is estimated while controlling for the effect of all of the other independent variables in the set. Multiple regression analysis readily incorporates a large number of independent variables, including both continuous and categorical variables, but requires assumptions to be made about the functional form of the relationship among the variables. When multiple regression analysis using many independent variables shows a statistically significant association between the dependent variable and a particular independent variable, the analyst may assume that a relationship exists between the two variables and that it possibly is a causal one, at least in the population sampled. One reason for this is that the analysis controls for the effects of all the other independent variables in the variable set. However, misleading results can still occur, particularly if causally important variables are omitted from the analysis.

### The Regression Model

The set of independent variables assumed to cause changes in the dependent variable and the functional form of this hypothesized relationship are usually referred to as the regression model. Note that a regression analysis examines how a particular set of independent variables organized into a particular model affects the value of the dependent variable for a particular population. The analysis is both model specific and population specific, although inferences are often made to a broader population and to other models.

*Variables used*—The model used in this report to analyze the relationship between family characteristics and total family charges for health care was derived from a general conceptualization of the health care system. This conceptualization suggests that generalized health status, specific health conditions (illnesses), and special health events (births, deaths, hospitalizations, and so forth) interact with family demographic factors to produce a family potential for the use of health care services. The actual use of care (the final level of total charges) results from a further interaction of the above health factors with economic and social factors such as sociocultural use patterns, family economic status, prices of health care, general economic conditions, and family health care coverage. Therefore, variables were selected for the model that were representative of the above types of health, economic, and social factors. It was assumed that a properly selected set of such variables would include several variables that significantly affect levels of total family charges for health care.

Besides this general conceptualization of the health care system, the review of the literature reported above suggested two hypotheses to be tested. First, Berki et al. (1985) and Wyszewianski (1986) suggested that the level of family income should be an important determinant of financially burdensome health expenses. Second, Berki suggested that high-cost illnesses should also be a determinant. To test these hypotheses, the model included variables for family income and types of illnesses that family members could have.

Testing hypotheses depends on controlling for variables that under alternative hypotheses could be the cause of the outcome actually found. A number of the independent variables in the model are of this type. Finally, the extensive literature on the importance of health insurance in affecting both total expenditures and out-of-pocket payments suggests that health insurance variables be included in the model. (For a brief review of this literature and recent findings, see Manning, Newhouse, Duan, et al., 1987.)

The independent variables selected could easily be arranged in categories of an Andersen-Newman model as presented in Buczko (1986). An Andersen-Newman model calls for health status variables (such as perceived health status), enabling variables (such as income), and predisposing variables (such as age). The model used here includes these types of variables, and one of the strengths of NMCUES is that it allowed for the inclusion of all these types of variables in the model. The specifics of operationalizing the model are found in Table I. This table gives the actual operational form of the dependent variable and of each of the 47 independent variables used in the regression analyses reported here.

It should be noted that many of the variables in Table I are imperfect indicators of the underlying

concepts that they represent. As a consequence, an actual variable can fit into the underlying conceptual scheme in more than one way. For example, the variable D9, which identifies families with a head of black race, may fit into the scheme in at least three ways. For one, it may be a demographic factor affecting health status. (Black persons, for example, have particularly high rates of hypertension.) Second, if racially based discrimination exists, D9 would denote a smaller supply of care available. Third, it may mark sociocultural differences in habits and preferences in the use of health care. Note that D9 does *not* represent overall economic differences because such differences are controlled for by the use of income as an independent variable in the regression.

The functional form of the relationship hypothesized to exist between the dependent variable, total family charges, and the independent variables is multiplicative. That is, it was assumed that a specified change in an independent variable will multiply total charges by a constant amount. For example, having a family member with heart disease might multiply a family's total charges by 1.3—that is, increase them by 30 percent—as compared with what they would be if no member had such an illness. (The multiplicative model was chosen in preference to an additive model for reasons detailed in Appendix I, which also describes how an additive model would work.) This hypothesized form of the functional relationship between the dependent variable and the independent variables calls for the dependent variable and several of the continuous independent variables to be used in logarithmic form. (Again, Appendix I explains why this is so.) Finally, the model does not take into account interaction effects between variables. In order to take such effects into account, special variables to measure interactions would have to be included in the model.

### Regression Procedures Followed

Regression analysis was carried out separately for the three socioeconomic multiple-person family populations focused upon in this study. As previously stated, these are older families (families with a member 65 years of age or over); younger, lower income families (families with no member 65 years of age or over and with family income below 200 percent of the poverty level); and younger, better off families (families with no member 65 years of age or over and with family income equal to or greater than 200 percent of the poverty level).

There were several steps in the analysis, as detailed in Appendix I. In brief, the steps were as follows. First, a small number of the initial 47 independent variables were excluded from each regression as not relevant. For example, a variable primarily used to distinguish between families with all members 65 years of age or over and families with only some members 65 years

of age or over was omitted from the regression for the two younger family populations. The initial exclusion left 43 to 45 independent variables in the regressions, with the number depending on the family population involved.

Next, stepwise regression was carried out using PC SAS (SAS Institute, 1985). A major reason for using stepwise regression was to eliminate possible multicollinearity (strong correlation) among variables, as several variables were sometimes used to operationalize a single basic concept. For example, four different sets of variables were used separately to operationalize the concept of family general health status. These were (1) total family bed days due to illness, (2) total family work loss days due to illness, (3) a family-level scale of reported health status, and (4) a family-level scale of limitations in main activity. The result of the stepwise regressions was a much smaller preferred regression model for each of the three socioeconomic family populations. These preferred models contained 23 to 31 independent variables.

However, PC SAS does not properly estimate variances of regression coefficients for samples with a complex survey design, such as that found in NMCUES. Therefore, the three preferred models were rerun as ordinary, nonstepwise regressions using SURREGR (Holt and Shah, 1982). SURREGR is a regression program that appropriately estimates variances in a sample with a complex design, but it cannot carry out stepwise regression analysis.

The results of the SURREGR regressions were used to identify which independent variables were statistically significant. These results are shown in detail in Tables II, III, and IV. Tables A, C, and E (one for each of the three family populations) show the statistically significant independent variables in each preferred model and the estimated effect of each significant variable on total family charges. Only about one-fourth to one-half of the independent variables in the preferred models were found to be statistically significant.

A SURREGR multiple regression on the full 43–45 variable models was carried out for each family population in order to check that the PC SAS stepwise regressions did not omit a statistically significant variable because of their deficiencies in variance estimation. Results are shown in Tables V, VI, and VII. No omissions were found by this procedure.

### Sampling Error

Because the statistics shown in this report are based on a sample of families rather than on information from all families, they are subject to sampling error. The standard error is a statistic that measures such errors. Standard errors for most estimates in this report have been computed and are presented along with the estimates.

## Nonsampling Error

In addition, estimates presented in this report are subject to nonsampling errors such as biased interviewing and reporting, misrecording of responses, undercoverage, and nonresponse. Extensive efforts were made to minimize these errors in the data collection and data processing for the survey (see Bonham, 1983).

In terms of nonsampling error, it should be noted that data in this report are derived from information furnished by a survey of households—that is, "consumers" of health care. Data reported by providers of care, for example, in surveys of physicians, hospitals, and nursing homes, are generally different from those reported by households (Sunshine, 1984). Anderson and Thorne (1985) specifically compared use of health care and expenditures on health care, as reported by families in NMCUES, with estimates underlying the national health accounts, which are generally provider based. They reported good agreement on total U.S. use of health care and on out-of-pocket expenditures for health care services after coverage differences—such as the omission of military and institutionalized persons from NMCUES—are taken into account. However, they found an approximate 10 percent difference between the national health accounts and NMCUES in total charges for health care services. It is likely that total charges, as estimated in this report, underestimate the true amount. Sunshine and Dicker (1987b, pp. 7–8) discuss in detail possible sources of problems in a family's reporting of total charges in NMCUES. These include (1) a family's limited knowledge of payments made on its behalf by insurance and government programs (a problem likely to be most severe when the family's own payments are zero or a small portion of total charges, as for inpatient hospital care or care under Medicaid) and (2) imputation of charges when care was provided free, for a nominal charge, or on a prepaid basis.

## Statistical Significance and Hypothesis Testing

Frequency tables in this report show not only estimates of mean total charges for health care for various family categories and estimates of the percent of families in each category with various high levels of total charges, but also an estimate of the standard error of each of these statistics. Where the text indicates that two estimates differ, the difference has been tested by a multiple $t$-test using the Bonferroni inequality (see Levy and Lemeshow, 1980, p. 296) and found significant at the 0.05 level. Standard errors were computed by the SESUDAAN computer software package (Shah, 1981), which takes into account the effect of the NMCUES complex sample design upon the standard errors of statistics estimated from its data.

This report uses multiple regression analysis to examine the relationship between total family charges for health care and approximately 45 independent variables that characterize families. Even after stepwise regression produced smaller preferred models, the preferred models still had a large number of variables, and an adjustment was made in estimating significance at the 0.05 level that was analogous to the adjustment made with a multiple $t$-test using the Bonferroni inequality. A variable was considered significant at the 0.05 level only if the simple estimate of its significance level was less than or equal to $0.05/n$, where $n$ is the number of independent variables in the preferred model. Using this adjustment, significance at the 0.05 level of probability was the equivalent of significance at the 0.0022 or lower level of probability as estimated by the SURREGR computer software package (Holt and Shah, 1982), which was used in the regression analyses. For more details, see Appendix I.

# Findings

This section presents findings on the determinants of total family charges for health care for multiple-person families in the civilian noninstitutionalized population of the United States during the year 1980. It discusses, in turn, each of the three family populations covered by this report.

Readers should note that the effect of a variable, when measured in a regression analysis, is estimated *after* controlling for the effect of all other variables in the regression. This is an important quantitative feature of the findings reported in this section and is noted at several points.

## Older Families

Older multiple-person families are families with a member age 65 or older and with average size of 1.5 members or greater. Statistically significant results from the regression analysis for these families are shown in Table A, with more details of the regression analysis found in Table II. The squared multiple correlation coefficient ($R^2$) for the regression equation for this family population was 0.72, which means the independent variables shown in Table II explained 72 percent of the variance in the dependent variable (which was the natural

### Table A
### Significant regression findings for total family charges for health care for older multiple-person families

| Significant factor | Effect (all other factors assumed constant) |
|---|---|
| **Special health event:**<br>Hospitalization | The regression coefficient for families with one or more hospitalizations (variable H13) was 1.179 and the regression coefficient for number of discharges (variable H14) was 0.244. Together, these imply a multiplication of total family charges by 4.15 for families with one discharge. In addition, the latter coefficient implies a multiplication of total family charges by 1.28 for each additional discharge. Thus, families with one discharge had total family charges about 4.15 times as high as families with no hospitalization and each additional discharge was associated with a further increase of approximately 28 percent in total family charges. |
| **Type of illness:**<br>Cancer | The regression coefficient for families with member(s) having cancer or other neoplasms was 0.377. This implies a multiplication by 1.46. Thus, total family charges for these families were approximately 46 percent higher than for families with no members having these diseases. |
| Heart and circulatory diseases | The regression coefficient for families with member(s) having heart or circulatory disease was 0.480. This implies a multiplication by 1.62. Thus, total family charges for these families were approximately 62 percent higher than for families with no members having these diseases. |
| Accidents, injuries, and poisonings | The regression coefficient for families with member(s) suffering accidents, injuries, or poisonings was 0.268. This implies a multiplication by 1.31. Thus, total family charges for these families were approximately 31 percent higher than for families with no members affected by such incidents. |
| **General health status:**<br>Family illness days in bed | The regression coefficient for family illness days in bed was 0.190. This means that each 1 percent increase in the quantity (family illness days in bed + 1) was associated with an increase of approximately 0.19 percent in total family charges. |
| Family income | The regression coefficient for family income was 0.148. This means that each 1 percent increase in family income was associated with approximately a 0.15 percent increase in total family charges. |

NOTES: For further details of the regression, see Appendix Table II. The probability for the 0.05 level of significance for the preferred model for older families using the multiple $F$-test discussed in Appendix I is 0.0020.

For an explanation of the above interpretations of the regression coefficients, see Appendix I.

logarithm of annualized total family charges for health care). Seven of the 25 variables in the preferred model were found to be statistically significant determinants of total family charges for older multiple-person families.

A particularly strong association was found between total family charges and hospitalization of a family member. The partial $F$-statistic, which measures the statistical significance of each independent variable in the regression, is far larger for hospitalization than for any other independent variable. Moreover, the numerical values of the regression coefficients involved are such that hospitalization had a large effect on total family charges. (Conceptually, hospitalization is a single variable. However, it was operationalized with two variables: (1) a dummy variable recording whether or not a family experienced any hospitalizations and (2) a continuous variable recording how many hospitalizations it experienced. This was done to permit the analysis to distinguish the effect of a family having any hospitalization at all from the effects of subsequent hospitalizations; the analysis finds large quantitative differences between these two effects. The very large partial $F$-statistic is found for the first of these two variables. Both variables are considered in estimating quantitatively the effect of hospitalization on total charges.) A family with one hospitalization of a family member in 1980 typically had total charges more than four times as large as those of an otherwise identical family with no hospitalizations. "Otherwise identical" here means (among other things) that both families had the same broad pattern of illnesses (such, for example, as having or not having a member with heart or circulatory disease) and had the same general family health status.

Additional hospitalizations were associated with a further increase of total family charges of approximately 28 percent for each hospitalization. Thus a family with two hospitalizations typically had total charges about 28 percent higher than an equivalent family with only one discharge. The former family typically had total charges more than five times as high as those of an equivalent family with no hospitalizations.

Interestingly, these estimated effects of hospitalization are smaller than those that appear in a frequency table for the same population of older, multiple-person families (Table B). This table shows families with hospitalizations experiencing mean total charges about 10 times as large as those of families with no hospitalizations. The difference between the two estimates is

## Table B

### Total family charges for health care for older multiple-person families, by hospitalization status and income: United States, 1980

| Characteristic | Sample size | Number of families in thousands | Mean total charges | Percent of families with total charges of— | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | | | | $3,000 or more | $5,000 or more | $10,000 or more | $20,000 or more |
| All | 840 | 10,538 | $3,312 | 27.6 | 18.2 | 7.9 | 2.3 |
| | | | (264) | (1.7) | (1.5) | (1.0) | (0.6) |
| **Hospitalization** | | | | | | | |
| No discharges | 507 | 6,342 | 747 | 2.0 | *0.2 | – | – |
| | | | (34) | (0.6) | (0.2) | | |
| One or more discharges | 333 | 4,197 | 7,188 | 66.2 | 45.4 | 19.8 | 5.9 |
| | | | (545) | (3.2) | (3.2) | (2.2) | (1.3) |
| **Annualized family income in 1980** | | | | | | | |
| Less than $15,000 | 436 | 5,561 | 2,984 | 25.5 | 15.6 | 6.7 | 2.2 |
| | | | (300) | (2.1) | (1.8) | (1.3) | (0.7) |
| $15,000 or more | 404 | 4,977 | 3,679 | 29.9 | 21.1 | 9.2 | 2.5 |
| | | | (384) | (2.3) | (2.2) | (1.3) | (0.8) |
| **Hospitalization and income** | | | | | | | |
| No discharges: | | | | | | | |
| Income less than $15,000 | 264 | 3,339 | 625 | *1.0 | – | – | – |
| | | | (39) | (0.6) | | | |
| Income $15,000 or more | 243 | 3,003 | 884 | *3.0 | *0.4 | – | – |
| | | | (52) | (1.1) | (0.4) | | |
| One or more discharges: | | | | | | | |
| Income less than $15,000 | 172 | 2,223 | 6,529 | 62.4 | 39.0 | 16.8 | *5.5 |
| | | | (645) | (4.4) | (4.3) | (2.9) | (1.8) |
| Income $15,000 or more | 161 | 1,974 | 7,931 | 70.6 | 52.6 | 23.1 | 6.3 |
| | | | (867) | (3.8) | (4.1) | (3.2) | (1.9) |

SOURCE: National Medical Care Utilization and Expenditure Survey, NCHS, 1980.

NOTES: Standard error appears below each estimate, in parentheses. Excludes very low income families—those reporting annual income less than $1,000 or less than 20 percent of the poverty level. Excludes families with zero total charges.

probably due to the fact that the regression controls for a large number of variables, such as the illnesses experienced by family members, while the frequency table does not. As a result, the tenfold difference in total charges associated in the frequency table with hospitalization probably includes differences due to other factors (such as the presence of major illnesses) that differ between families with hospitalizations and families that did not experience hospitalizations.

Given the large effect of hospitalization on total family charges for health care, it is not surprising that families that experienced one or more hospitalizations of their members in 1980 were much more likely to have high total charges than families with no hospitalizations. Table B shows, for example, that 66 percent of older, multiple-person families with hospitalizations in 1980 had total charges of $3,000 or more compared with 2 percent of families with no hospitalizations. For total charges of $5,000 or more, the corresponding statistics were 45 percent compared with less than 1 percent. (Note that, in these comparisons of frequencies, the two family categories are not comparable with respect to characteristics other than hospitalization. In contrast, in the regression analyses families are comparable in other characteristics.)

Older families with one or more members having major illnesses experienced higher total family charges for health care than otherwise similar families that did not have such illnesses. Families with a member having cancer or other neoplasms in 1980 had total charges averaging 46 percent higher than similar families that did not have any members with these diseases (Table A). It should, moreover, be noted that this is the difference that was found after controlling for the effects of other factors included in the regression—in particular, after controlling for general health status and hospitalization. Cancer often leads to hospitalization and impaired general health status, and to the extent it does so, the 46 percent increase underestimates its full effect on total charges.

A similar effect was found for the presence of heart or circulatory disease among members of older families. Families with a member having these diseases in 1980 had total charges averaging 62 percent higher than similar families with no members having any of these diseases (Table A). Again it should be noted that this is the difference after controlling for the effects of other factors included in the regression, such as general health status and hospitalization.

Finally, accidents, injuries, and poisonings were found to have a similar, but smaller, effect. On average, families with a member who suffered from an accident, injury, or poisoning in 1980 had total charges approximately 31 percent higher than those of like families in which no members suffered from such incidents (Table A).

General family health status had an additional effect on total family charges of older multiple-person families in 1980. Family illness days in bed was the general

health status variable that was found to be statistically significant (Table A). However, if this variable had been omitted from the regression, it is quite possible that some other variable measuring general family health status would have been significant in its place. In any case, the estimated effect of family illness days was such that each 1 percent increase in the quantity (family illness days in bed plus 1) increased total family charges by approximately 0.19 percent. This works out, for example, to a family with 10 illness days in bed in 1980 having total family charges about 12 percent higher than a similar family with 5 illness days in bed. Again, this is the estimated effect after controlling for other variables included in the regression, which means it is the effect over and above the effects (described in previous paragraphs) that are associated with specific illnesses or hospitalizations.

Finally, family income had a statistically significant effect on total family charges of older, multiple-person families in 1980. Each 1 percent increase in family income was associated with approximately a 0.15 percent increase in total charges (Table A). This is equivalent to saying that if one of two otherwise similar families had twice the income of the other, total charges of the higher income family would have been about 11 percent higher than those of the lower income family.

## Younger, Lower Income Families

Younger, lower income multiple-person families are families with all members under 65 years of age, with characteristics (income, family size, and so forth) that place them below 200 percent of the poverty level, and with an average family size of 1.5 members or greater. Statistically significant results from the regression analysis for these families are shown in Table C, with more details of the regression analysis shown in Table III. The squared multiple correlation coefficient ($R^2$) for the regression equation for this family population was 0.60, which means the independent variables shown in Table III explained 60 percent of the variance in the dependent variable (which was the natural logarithm of annualized total family charges for health care). Eight of the 23 variables in the preferred model were found to be statistically significant determinants of total family charges for younger, lower income multiple-person families.

A particularly strong association again was found between total family charges and hospitalization of a family member. The partial $F$-statistic, which measures the statistical significance of this association, is again far larger for hospitalization than for any other independent variable. Moreover, the numerical values of the regression coefficients involved are such that hospitalization again had a large effect on total family charges. (Again, there are two variables involved, one recording whether a family experienced any hospitalization of its members and the other recording how many hospitalizations. Thus

**Table C**

**Significant regression findings for total family charges for health care for younger, lower income multiple-person families**

| Significant factor | Effect (all other factors assumed constant) |
|---|---|
| **Special health event:** | |
| Hospitalization . . . . . . . . . . . . . . . . . . . | The regression coefficient for families with one or more hospitalizations (variable H13) was 1.099 and the regression coefficient for number of discharges (variable H14) was 0.280. Together, these imply a multiplication of total family charges by 3.97 for families with one discharge. In addition, the latter coefficient implies a multiplication of total family charges by 1.32 for each additional discharge. Thus, families with one discharge had total family charges about 3.97 times as high as families with no hospitalization and each additional discharge was associated with a further increase of approximately 32 percent in total family charges. |
| **Type of illness:** | |
| Cancer . . . . . . . . . . . . . . . . . . . . . | The regression coefficient for families with member(s) having cancer or other neoplasms was 0.377. This implies a multiplication by 1.46. Thus, total family charges for these families were approximately 46 percent higher than for families with no members having these diseases. |
| **General health status:** | |
| Poor perceived health . . . . . . . . . . . . | The regression coefficient for families with member(s) reported in poor health was 0.334. This implies a multiplication by 1.40. Thus, total family charges for these families were approximately 40 percent higher than for families with all members reported to be in excellent health or in good or excellent health. |
| Major limitation in activity . . . . . . . . . . | The regression coefficient for families with member(s) reported unable to perform their usual major activity (work, housekeeping, school, and so on) was 0.262. This implies a multiplication by 1.30. Thus, total family charges for these families were approximately 30 percent higher than for families with no members reported unable to perform their usual major activity. |
| Family illness days in bed . . . . . . . . . . | The regression coefficient for family illness days in bed was 0.126. This means that each 1 percent increase in the quantity (family illness days in bed + 1) was associated with an increase of approximately 0.13 percent in total family charges. |
| Completeness of health care coverage . . . . | The regression coefficient for families with no members having any health care coverage was − 0.565. This implies a multiplication by 0.57. Thus, total family charges for these families were approximately 43 percent lower than for families with all members having full-year coverage. |
| Region . . . . . . . . . . . . . . . . . . . . . | The regression coefficient for families residing in the South was − 0.241. This implies multiplication by 0.79. Thus, families residing in the South had total family charges approximately 21 percent lower than families residing elsewhere in the United States. |

NOTES: For further details of the regression, see Appendix Table III. The probability for the 0.05 level of significance for the preferred model for younger, lower income families using the multiple $F$-test discussed in Appendix I is 0.0022.

For an explanation of the above interpretations of the regression coefficients, see Appendix I.

two regression coefficients are involved.) A family with one hospitalization of a family member in 1980 typically had total charges approximately four times as large as those of an otherwise identical family with no hospitalizations. "Otherwise identical" here means (among other things) that both families had the same broad pattern of illnesses (such, for example, as having or not having a member with heart or circulatory disease) and had the same general family health status.

Additional hospitalizations were associated with a further increase in total family charges of approximately 32 percent for each hospitalization. Thus a family with two hospitalizations typically had total charges about 32 percent higher than an equivalent family with only one discharge. The former family typically had total charges more than five times as high as those of a comparable family with no hospitalizations.

These estimated effects of hospitalization are broadly similar to those which appear in a frequency table for the same population of younger, lower income multiple-

person families (Table D). This table shows families with hospitalizations in 1980 experiencing mean total charges about six times as large as those of families with no hospitalizations. (Statistics in Table III indicate that there were an average of almost two discharges per family that experienced a hospitalization, so the appropriate comparison is between a ratio of approximately 5 to 1 in total charges shown by the regression as distinguishing families with two hospitalizations from families with no hospitalizations and a ratio of approximately 6 to 1 in total charges shown by the frequency table.)

Given the large effect of hospitalization on total family charges for health care, it is not surprising that younger, lower income families that experienced one or more hospitalizations of their members in 1980 were much more likely to have high total charges than families with no hospitalizations. Table D shows, for example, that 43 percent of younger, lower income families with hospitalizations in 1980 had total charges of $3,000

**Table D**

**Total family charges for health care for younger, lower income multiple-person families, by hospitalization status and income: United States, 1980**

| Characteristic | Sample size | Number of families in thousands | Mean total charges | Percent of families with total charges of— | | | |
|---|---|---|---|---|---|---|---|
| | | | | $3,000 or more | $5,000 or more | $10,000 or more | $20,000 or more |
| All . . . . . . . . . . . . . . . . . . . . . . . . . . | 1,013 | 13,128 | $1,845 (91) | 15.6 (1.2) | 8.0 (0.9) | 2.8 (0.5) | *0.8 (0.3) |
| **Hospitalization** | | | | | | | |
| No discharges . . . . . . . . . . . . . . . . . . | 675 | 8,704 | 698 (46) | *1.6 (0.5) | *0.4 (0.3) | *0.3 (0.3) | – |
| One or more discharges . . . . . . . . . . . . . | 338 | 4,423 | 4,104 (210) | 43.0 (2.8) | 23.0 (2.4) | 7.7 (1.5) | *2.5 (0.8) |
| **Annualized family income in 1980** | | | | | | | |
| Less than $10,000 . . . . . . . . . . . . . . . . | 526 | 7,089 | 1,702 (121) | 13.6 (1.6) | 7.0 (1.3) | *1.8 (0.6) | *0.7 (0.3) |
| $10,000 or more . . . . . . . . . . . . . . . . . | 487 | 6,039 | 2,014 (152) | 17.9 (1.8) | 9.2 (1.2) | 3.9 (1.0) | *0.9 (0.4) |
| **Hospitalization and income** | | | | | | | |
| No discharges: | | | | | | | |
|   Income less than $10,000 . . . . . . . . . . . . | 350 | 4,683 | 622 (34) | *1.1 (0.5) | – | – | – |
|   Income $10,000 or more . . . . . . . . . . . . . | 325 | 4,021 | 786 (87) | *2.3 (0.9) | *0.9 (0.7) | *0.6 (0.6) | – |
| One or more discharges: | | | | | | | |
|   Income less than $10,000 . . . . . . . . . . . . | 176 | 2,405 | 3,805 (304) | 38.0 (3.9) | 20.7 (3.5) | *5.3 (1.9) | *2.2 (1.0) |
|   Income $10,000 or more . . . . . . . . . . . . . | 162 | 2,018 | 4,462 (331) | 49.0 (3.9) | 25.8 (3.2) | 10.7 (2.4) | *2.8 (1.2) |

SOURCE: National Medical Care Utilization and Expenditure Survey, NCHS, 1980.

NOTES: Standard error appears below each estimate, in parentheses. Excludes very low income families—those reporting annual income less than $1,000 or less than 20 percent of the poverty level. Excludes families with zero total charges.

or more compared with approximately 2 percent of families with no hospitalizations. For total charges of $5,000 or more, the corresponding statistics were 23 percent compared with less than 1 percent. (Note that, in these comparisons of frequencies, the two family categories are not comparable with respect to characteristics other than hospitalization.)

Younger, lower income multiple-person families with a member having cancer or other neoplasms in 1980 were found to have total charges averaging 46 percent higher than similar families that did not have any members with these diseases (Table C). It should again be noted that thi· is the difference that was found after controlling for the effects of other factors included in the regression—in particular, after controlling for general health status and hospitalization. Thus, again, to the extent cancer leads to impaired general health status and hospitalization, the 46 percent increase is an underestimate of its full effect on total charges. For heart and circulatory disease and for accidents, injuries, and poisonings, the regression (Table III) shows apparently similar effects which, however, are not statistically significant by the significance test used here ($p < 0.0022$).

In contrast, three measures of general family health status had statistically significant effects on total family charges for health care for younger, lower income multiple-person families in 1980. These were poor reported health, major limitation in activity, and family illness days in bed (Table C). As regards the first of these, families with one or more members reported to be in poor health had, on average, total charges 40 percent higher than those of otherwise similar families with all members reported in excellent health or with all members reported in good or excellent health. Second, families with member(s) unable to perform their usual major activity (usual major activity is defined as work, housekeeping, school, and so on) had total charges averaging 30 percent higher than those of otherwise similar families with no members limited or with members having less severe limitations. Third, family illness days in bed also had a statistically significant effect. The estimated effect was such that each 1 percent increase in the quantity (family illness days in bed plus 1) increased total family charges by approximately 0.13 percent. This works out, for example, to a family with 10 illness days in bed in 1980 having total family charges

17

about 8 percent higher than a similar family with 5 illness days in bed. It should again be noted that these estimated effects of variables measuring general family health status are estimates of effects over and above the effects of associated specific illnesses or hospitalizations.

Moreover, the estimate of the effect of each of the three general health status variables controls for the effects of the others. Thus, if one is interested in the total effect of general family health status (apart from the effects of specific illnesses and hospitalizations), it is necessary to combine these effects. For example, a family member in bad health might be unable to perform his usual major activity, be rated in poor health, and be confined to bed a relatively large number of days. A family with such a member would be expected to have total charges about twice as great as an otherwise similar family with all members rated in excellent or good health, with less severe or no limitations in activity among its members, and with only half as many bed days.

Low total charges were found for younger, lower income families with no members having any health care coverage. (Health care coverage includes both private health insurance and coverage by public programs such as Medicare or Medicaid.) Total charges for families with no coverage were about 43 percent lower than for families with all members having full-year coverage (Table C).

Finally, the regression shows younger, lower income families that lived in the South in 1980 had lower total family charges for health care than otherwise similar families living elsewhere in the United States. Total charges were approximately 21 percent lower in the South (Table C).

## Younger, Better Off Families

Younger, better off multiple-person families are families with all members under 65 years of age, with characteristics (income, family size, and so forth) that place them at 200 percent of the poverty level or higher, and with an average family size of 1.5 members or greater. Statistically significant results from the regression analysis for these families are shown in Table E, with more details of the regression analysis shown in Table IV. The multiple correlation coefficient ($R^2$) for the regression equation for this family population was 0.57, which means the independent variables shown in Table IV explained 57 percent of the variance in the dependent variable (which was the natural logarithm of annualized total family charges for health care). Sixteen of the 31 variables in the preferred model were found to be statistically significant determinants of total family charges for younger, better off multiple-person families. (However, 1 of the 16 significant variables—namely, reported health status of all family members is unknown—should be regarded as a limitation

of the data on the less than 1 in each 1,000 families in the sample with this characteristic, not as a substantive finding. It does not appear in Table E and is not discussed further.)

A particularly strong association again was found between total family charges and hospitalization of a family member. The partial $F$-statistic, which measures the statistical significance of each independent variable in the regression, again is far larger for hospitalization than for any other independent variable. Moreover, the numerical values of the regression coefficients involved are such that hospitalization had a large effect on total family charges. (As previously pointed out, two variables are involved, one recording whether a family experienced any hospitalization of its members and the other recording how many hospitalizations. Thus two regression coefficients are pertinent.) A family with one hospitalization of a family member in 1980 typically had total charges approximately 3.6 times as large as those of an otherwise identical family with no hospitalization. "Otherwise identical" here again means (among other things) that both families had the same broad pattern of illnesses (such as having or not having a member with heart or circulatory disease) and had the same general family health status.

Additional hospitalizations were associated with a further increase in total family charges of approximately 31 percent for each hospitalization. Thus a family with two hospitalizations typically had total charges about 31 percent higher than an equivalent family with only one discharge. That is, its total charges were close to five times as high as those of a family with no hospitalizations.

These estimated effects of hospitalization are similar to those which appear in a frequency table for the same population of younger, better off multiple-person families (Table F). That table shows families with hospitalizations experiencing total charges somewhat over five times as large as those of families with no hospitalizations.

Given the large effect of hospitalization on total family charges for health care, it is not surprising that, like other family categories, younger, better off families that experienced one or more hospitalizations of their members in 1980 were much more likely to have high total charges than families with no hospitalizations. Table F shows, for example, that 51 percent of younger, better off families with hospitalizations in 1980 had total charges of $3,000 or more compared with 2.5 percent of families with no hospitalizations. For total charges of $5,000 or more, the corresponding statistics were 25 percent, compared with 0.4 percent. (Note that, in these comparisons of frequencies, the families with and without hospitalizations are not comparable with respect to characteristics other than hospitalization.)

Four sociodemographic factors had a statistically significant effect on total family charges for younger, better off multiple-person families in 1980. These were

## Table E
### Significant regression findings for total family charges for health care for younger, better off multiple-person families

| Significant factor | Effect (all other factors assumed constant) |
| --- | --- |
| Family structure . . . . . . . . . . . . . . . . . | The regression coefficient for families with children was 0.199. This implies a multiplication by 1.22. Thus, total family charges for families with children were 22 percent higher than for families without children, and this was apart from effects due to family size. |
| Age . . . . . . . . . . . . . . . . . . . . . . . | The regression coefficient for age of head was 0.008. This means that each additional year of age of the family head was associated with an increase of approximately 0.8 percent in family total charges. |
| Race . . . . . . . . . . . . . . . . . . . . . . | The regression coefficient for families with a black head of family was − 0.366. This implies a multiplication by 0.69. Thus, total family charges for families with a black head of family were approximately 31 percent lower than for families with heads of white race. |
| Education . . . . . . . . . . . . . . . . . . . . | The regression coefficient for education of family head was 0.026. This means that each additional year of education of the family head was associated with an increase of approximately 2.6 percent in total family charges. |
| Special health event: | |
| Hospitalization . . . . . . . . . . . . . . . . . | The regression coefficient for families with one or more hospitalizations (variable H13) was 1.018 and the regression coefficient for number of discharges (variable H14) was 0.272. Together, these imply a multiplication of total family charges by 3.63 for families with one discharge. In addition, the latter coefficient implies a multiplication of total family charges by 1.31 for each additional discharge. Thus, families with one discharge had total family charges about 3.63 times as high as families with no hospitalization and each additional discharge was associated with a further increase of approximately 31 percent in total family charges. |
| Type of illness: | |
| Heart and circulatory diseases . . . . . . . . | The regression coefficient for families with member(s) having heart or circulatory disease was 0.179. This implies a multiplication by 1.20. Thus, total family charges for these families were approximately 20 percent higher than for families with no members having these diseases. |
| Accidents, injuries, and poisonings . . . . . . | The regression coefficient for families with member(s) suffering accidents, injuries, or poisonings was 0.227. This implies a multiplication by 1.25. Thus, total family charges for these families were approximately 25 percent higher than for families with no members affected by such incidents. |
| General health status: | |
| Fair perceived health . . . . . . . . . . . . . | The regression coefficient for families with member(s) reported in fair health was 0.224. This implies a multiplication by 1.25. Thus, total family charges for these families were approximately 25 percent higher than for families with all members reported to be in excellent health. |
| Poor perceived health . . . . . . . . . . . . . | The regression coefficient for families with member(s) reported in poor health was 0.419. This implies a multiplication by 1.52. Thus, total family charges for these families were approximately 52 percent higher than for families with all members reported to be in excellent health. |
| Family illness days in bed . . . . . . . . . . | The regression coefficient for family illness days in bed was 0.112. This means that each 1 percent increase in the quantity (family illness days in bed + 1) was associated with an increase of approximately 0.11 percent in total family charges. |
| Family work-loss days due to illness . . . . . | The regression coefficient for families work-loss days due to illness was 0.072. This means that each 1 percent increase in the quantity (family work-loss days + 1) was associated with an increase of approximately 0.07 percent in total family charges. |
| Family income . . . . . . . . . . . . . . . . . | The regression coefficient for family income was 0.264. This means that each 1 percent increase in family income was associated with approximately a 0.26 percent increase in total family charges. |
| Completeness of health care coverage . . . . | The regression coefficient for families with some (but not all) member(s) having no health care coverage was − 0.285. This implies a multiplication by 0.75. Thus, total family charges for these families were approximately 25 percent lower than for families with all members having full-year coverage. |
| Region . . . . . . . . . . . . . . . . . . . . . . | The regression coefficient for families residing in the West was 0.135. This implies multiplication by 1.14. Thus, families residing in the West had total family charges approximately 14 percent higher than families residing elsewhere in the United States. |

NOTES: For further details of the regression, see Appendix Table IV. In addition to effects described above, Table IV also shows a statistically significant finding for the approximately 1 in 1000 families with all members having an unknown perceived health status; this should be regarded as a limitation of the data rather than a substantive finding regarding an analytically meaningful category of families. The probability for the 0.05 level of significance for the preferred model for younger, better off families using the multiple $F$-test discussed in Appendix I is 0.0016.

For an explanation of the above interpretations of the regression coefficients, see Appendix I.

Total family charges for health care for younger, better off multiple-person families, by hospitalization status and income:
United States, 1980

| Characteristic | Sample size | Number of families in thousands | Mean total charges | Percent of families with total charges of— | | | |
|---|---|---|---|---|---|---|---|
| | | | | $3,000 or more | $5,000 or more | $10,000 or more | $20,000 or more |
| All | 2,883 | 33,372 | $1,840 (73) | 15.5 (0.7) | 7.1 (0.6) | 2.2 (0.3) | *0.5 (0.2) |
| **Hospitalization** | | | | | | | |
| No discharges | 2,131 | 24,412 | 845 (20) | 2.5 (0.3) | 0.4 (0.1) | *0.0 (0.0) | — |
| One or more discharges | 752 | 8,960 | 4,548 (245) | 51.1 (1.9) | 25.4 (1.9) | 7.9 (1.2) | 1.9 (0.6) |
| **Annualized family income in 1980** | | | | | | | |
| Less than $15,000 | 1,181 | 13,554 | 1,650 (110) | 13.8 (1.0) | 5.6 (0.8) | 1.8 (0.4) | *0.6 (0.3) |
| $15,000 or more | 1,702 | 19,818 | 1,969 (83) | 16.7 (0.9) | 8.1 (0.7) | 2.4 (0.4) | *0.4 (0.2) |
| **Hospitalization and income** | | | | | | | |
| No discharges: | | | | | | | |
| Income less than $15,000 | 885 | 10,035 | 689 (25) | *1.6 (0.5) | — | — | — |
| Income $15,000 or more | 1,246 | 14,377 | 955 (27) | 3.1 (0.4) | *0.7 (0.2) | *0.1 (0.1) | — |
| One or more discharges: | | | | | | | |
| Income less than $15,000 | 296 | 3,519 | 4,391 (346) | 48.8 (2.9) | 21.5 (2.7) | 7.0 (1.6) | *2.3 (0.9) |
| Income $15,000 or more | 456 | 5,442 | 4,649 (261) | 52.5 (2.3) | 27.9 (2.5) | 8.5 (1.5) | *1.6 (0.6) |

SOURCE: National Medical Care Utilization and Expenditure Survey, NCHS, 1980.

NOTES: Standard error appears below each estimate, in parentheses. Excludes very low income families—those reporting annual income less than $1,000 or less than 20 percent of the poverty level. Excludes families with zero total charges.

presence of children, age, education, and race of family head (Table E).

Families with children age 16 or younger had total family charges about 22 percent higher than otherwise comparable families that did not have children. As with all regression coefficients, this statistic gives the estimated effect of this variable after controlling for the influence of all other variables in the regression. Thus, because family size was included in the regression, larger family size is not the explanation for the higher total charges found for families with children.

Total family charges were higher in 1980 for families with older heads of family. Each additional year of age of the family head increased total charges by approximately 0.8 percent. Although this may sound like a small effect, it becomes quite substantial with substantial age differences. For example, a family with a head 30 years older than that of an otherwise similar family would have had total charges about 25 percent higher than the family with the younger head.

Among younger, better off families in 1980, more education was associated with higher total charges. Each additional year of education completed by the head of the family increased total charges by about 2.6 percent. Thus a family with a head who completed college would

have had total charges about 10 percent higher than those of an otherwise similar family (in particular, a family with a similar income) in which the head had completed high school but had had no further formal education.

Black families had relatively low total charges for health care. Their total charges were about 31 percent lower than those for comparable white families. Note that this difference is not attributable to differences among races in economic, educational, or insurance status, for these factors are included as variables in the regression and their effects thereby controlled for.

Some types of major illnesses led to increased total family charges for younger, better off multiple-person families (Table E). Specifically, families with a member having heart or circulatory disease had total charges approximately 20 percent higher than similar families with no members having these diseases; and families with a member who suffered from an accident, injury, or poisoning in 1980 on average had total charges approximately 25 percent higher than did similar families with no members affected by such incidents. Again, it should be noted that these estimated effects of illnesses on total charges are measured after controlling for the effects of general health status, hospitalization, and other factors

included in the regression. To the extent that major illnesses affect general health status and give rise to hospitalization (as they do), their effect on total charges is greater than the statistics in this paragraph suggest.

As Table E shows, multiple measures of general family health status also had statistically significant effects on total family charges for health care of younger, better off multiple-person families in 1980. For one, families with one or more members reported to be in fair health (but with no member reported in poor health) had, on average, total charges 25 percent higher than those of otherwise similar families with all members reported in excellent health. And families with a member reported in poor health had total charges averaging more than 50 percent above those of otherwise similar families with all members reported in excellent health. Second, families with more illness days in bed experienced higher charges than those with fewer illness days in bed. The estimated effect was such that each 1 percent increase in the quantity (family illness days in bed plus 1) increased total family charges by approximately 0.11 percent. This works out, for example, to a family with 10 illness days in bed in 1980 having total family charges about 7 percent higher than a similar family with 5 illness days in bed. Finally, families with many work-loss days due to illness had high total charges. The estimated effect was such that each 1 percent increase in the quantity (family work-loss days due to illness plus 1) increased total family charges by approximately 0.07 percent. This works out, for example, to a family that experienced 10 work-loss days due to illness in 1980 having total charges about 4 percent higher than a similar family with 5 work-loss days.

It should again be noted that the estimate of the effect of each of these general health status variables controls for the effects of the others. Thus, if one is interested in the total effect of general family health status (apart from the effects of specific illnesses and hospitalization), it is necessary to combine these effects. For example, a family member in bad health might be rated as being in poor health, missing many work days, and being confined to bed a relatively large number of days. A family with such a member would be expected to have total charges about 70 percent higher than an otherwise similar family with all members rated in excellent health and with only half as many bed days and half as many days lost from work due to illness.

Higher family income was associated with higher total charges among younger, better off multiple-person families in 1980. Each 1 percent increase in family income was associated with approximately a 0.26 percent increase in total charges (Table E). This is equivalent to saying that if one of two otherwise similar families had twice the income of the other, total charges of the higher income family would have been about 20 percent higher than those of the lower income family.

Completeness of health care coverage also had a statistically significant effect on total charges for younger, better off families. Families with some but not all members having no health care coverage in 1980 had total charges approximately 25 percent lower than similar families with all members having full-year health care coverage (Table E). (As Table IV shows, an effect similar in magnitude, but not meeting tests of statistical significance, was also found for families with no members having any health care coverage.)

Finally, the regression shows younger, better off families residing in the West in 1980 had higher total family charges for health care than otherwise similar families living elsewhere in the United States. Total charges were approximately 14 percent higher in the West (Table E).

# Discussion

This section focuses on three aspects of the findings presented in the previous section: (1) the power of the regression analyses to explain variation in total family charges for health care, (2) the contribution of selected individual variables to the explanation, and (3) the contribution of overall patterns of variables to the explanation.

The Introduction to this report points out that fully understanding financially burdensome health expenses requires comparing the determinants of the potential financial burden families could face (consisting of total charges as measured in this report) and the actual burden that families do face (their total out-of-pocket expenses). Total out-of-pocket expenses were examined in a previous report (Dicker and Sunshine, 1988) which pointed out that they are the sum of (1) out-of-pocket expenses for health care and (2) family-paid premiums for health care coverage. That report also pointed out that the best measure of family financial burden is not total out-of-pocket expenses *per se*, but the ratio of these expenses to family income. This ratio measure was called the financial burden index; it is the portion of total family income consumed by total family out-of-pocket expenses for health.

The following discussion frequently compares the regression findings previously presented for the financial burden index (Dicker and Sunshine, 1988) and the regression findings for total family charges for health care developed in this report. The regressions in this report use the same data base and the same independent variables as the previous report did. Because the methodology and data base used in both reports are the same, the comparison highlights differences between the determinants of actual financial burden, as measured by the financial burden index, and the determinants of potential financial burden, as measured by total charges.

Another central topic of this discussion section is the comparison of regression results among the three family populations of interest. This comparison is important because it indicates whether particular findings are common to all family populations or are particular to only one family population and it shows whether variables have similar-sized or dissimilar effects across the three populations. The three family populations, as previously indicated, were older multiple-person families (those with a member 65 years of age or over); younger,

lower income multiple-person families (those with all members under 65 years of age and with incomes below 200 percent of the poverty level); and younger, better off multiple-person families (those with all members under 65 years of age and with incomes of 200 percent of the poverty level or more). These populations, combined, constitute the entire civilian, multiple-person family population of the United States.

The dependent variable in this report, total family charges for health care, is the annualized total amount billed to a family for all health care services and supplies regardless of whether the charges were paid out-of-pocket, paid by health care coverage, or remained unpaid. (Premiums for health care coverage are not included.) In principle, total family charges for health care should include charges for all types of health care services and supplies. However, because of limitations in NMCUES data coverage, total family charges, as used in this report, omit some categories of charges. Most importantly, they do not include charges for nonprescription medicines, nursing homes, or other types of long-term care.

## Explanatory Power

$R^2$, the multiple correlation coefficient squared, is a measure of the overall explanatory power of an entire regression. $R^2$ is equal to the proportion of the variance in the dependent variable that is explained by all the independent variables in combination. In order to provide a better understanding of the relative explanatory power of the regression equations reported in Tables II, III, and IV, $R^2$ for these equations is compared here with the $R^2$ statistics reported in other, similar regression studies. As shown in the first column of Table G, $R^2$ statistics in the regressions for total family charges for health care ranged from 0.57 to 0.72, depending on the socioeconomic family category involved.

This is a relatively high $R^2$ compared with that reported in most studies. For example, three recent papers that use NMCUES data or data from the similar 1977 National Medical Care Expenditure Survey (NMCES) in regression equations similar to those presented here report an $R^2$ of 0.04 to 0.27 (Farley, 1986), 0.31 (Taube, Kessler, and Burns, 1986), and 0.18 to 0.20 (Buczko,

**Comparison of multiple correlation coefficients squared, by dependent variable and age and family status relative to the poverty level**

| Age and status of family | Dependent variable and model | | |
|---|---|---|---|
| | Total family charges for health care (preferred model)[1] | Index of financially burdensome family health expenses (preferred model)[2] | Total family charges for health care (full model)[1] |
| Older | 0.72 | 0.53 | 0.72 |
| Younger: | | | |
|   Lower income | 0.60 | 0.27 | 0.60 |
|   Better off | 0.57 | 0.23 | 0.57 |

[1]Dependent variable is natural logarithm of stated statistic.
[2]Regression results not shown in this report.

NOTES: Older families are families with member(s) 65 years of age or over. Younger families are families with no member 65 years of age or over. Lower income families are families with income below 200 percent of the poverty level. Better off families are families with income of 200 percent of the poverty level or more.

1986). These studies, however, differ from this report in that they deal with individuals, not families, with physician visits, not total health care, and—except for the Buczko study—with number of visits, not spending.

Similarly, an analysis exactly paralleling that described in this report, but using the financial burden index instead of total family charges for health care as the dependent variable, also obtained lower $R^2$ statistics (Dicker and Sunshine, 1988). As the second column in Table G shows, $R^2$ statistics in the regression analyses of the financial burden index were 0.23 to 0.53, depending on the family population involved. These comparisons suggest that while some variance still remains unexplained, the preferred sets of independent variables selected for use in the three regression analyses described in this report are relatively powerful in explaining differences among families in total charges for health care. This is particularly true for the older, multiple-person family population for which 72 percent of the variance in total family charges is explained.

It is also interesting to compare the $R^2$ statistics of two alternate versions of the regressions for total family charges. Preferred models with 23 to 31 independent variables are the source of the findings reported and discussed in this chapter and are shown in Appendix Tables II, III, and IV. However, full models with 43 to 45 independent variables were also run using SURREGR, and are shown in Appendix Tables V, VI, and VII (see Appendix I for more information on the two types of models). The third column of Table G shows the $R^2$ statistics for the full models. For all three family populations, it differs by less than 0.01 from the $R^2$ statistics for the preferred model. This is a typical finding for preferred models developed with stepwise regression (as the models presented here were), and shows that very little explanatory power is lost by using the preferred models rather than the full models.

Interestingly, for both total family charges for health care and the financial burden index, the independent variable sets used in this report account for more of the variance among older families than among younger

family populations. This suggests that factors not included in the regressions (or poorly measured by the included variables) are less important for determining levels of these two health cost measures among older families than among younger families. Why this should be true is not obvious.

## Individual Variables

Two variables were statistically significant as determinants of total family charges for health care in all three populations. These were hospitalization (whether a family had members who experienced one or more hospital inpatient episodes) and family illness days in bed (the annualized total number of illness days in bed experienced by all family members).

The regressions indicate a very strong role for *hospitalization*, and they do so in two ways. First, not only was hospitalization one of only two variables found statistically significant for all three family populations, but in addition, for all three populations its statistical significance, as measured by the partial $F$-statistic, was far stronger than that of any other independent variable. Second, quantitatively, it had a very large effect on the total of family charges. In all three family populations, a family with one hospitalization in 1980 had total family charges roughly 4 times as large as those of an otherwise identical family with no hospitalizations (Table H). In addition, each further hospitalization increased total family charges by approximately 30 percent in each of the three family populations. As Table H shows, the estimated effect of one hospitalization and of each further hospitalization on the sum total of family charges was fairly similar in size across all three family populations.

In all these comparisons, it is important to note that the analysis controls for the effects of all other variables included in the regressions and in particular that "otherwise identical family" here means, among other things, that the two families did not differ in broad

**Effects of selected factors on family health cost measures for the three family socioeconomic populations**

| Description | Effect, other factors assumed equal, on— | | |
| --- | --- | --- | --- |
| | Older families | Younger lower income families | Younger better off families |
| Effects of hospitalization: | | Ratio | |
| Ratio of total charges for family with one hospitalization to total charges for family with no hospitalizations | 4.15 | 3.97 | 3.63 |
| | | Percent | |
| Increase in total family charges for each additional hospitalization of a family member | 28 | 32 | 31 |
| Effects of income: | | | |
| Increase in total family charges for each 1 percent increase in income | 0.15 | (1) | 0.26 |
| Decrease in financial burden index for each 1 percent increase in income | 0.85 | 0.56 | 0.87 |
| Increase in total out-of-pocket expenses for each 1 percent increase in income | 0.15 | 0.44 | 0.13 |
| Effects of illness days in bed: | | | |
| Increase in total family charges for each 1 percent increase in the quantity (family illness days in bed + 1) | 0.19 | 0.13 | 0.11 |
| Effects of specific illnesses: | | | |
| Increase in total charges for a family with member(s) having cancer relative to total charges for a family with no member having cancer | 46 | 46 | (1) |
| Increase in total charges for a family with member(s) having heart or circulatory disease relative to total charges for a family with no member having these diseases | 62 | (1) | 20 |
| Increase in total charges for a family with member(s) suffering accidents, injuries, or poisonings relative to total charges for a family with no member suffering these incidents | 31 | (1) | 25 |

[1]Data not statistically significant.

SOURCES: Tables A, C, and E; Dicker and Sunshine, 1988.

patterns of illness, such as whether they had a member with cancer, and did not differ in general health status. One descriptive study that did not control for other factors showed elderly individuals with hospitalizations having, on average, total charges approximately 10 times as great as those of elderly individuals with no hospitalizations (Kovar, 1986), and the findings of this report are similar if there is no controlling for factors other than hospitalization (Table B).

Hospitalization was far less important as a determinant of the financial burden index than it was as a determinant of total family charges for health care. In the study of determinants of the financial burden index mentioned above (Dicker and Sunshine, 1988), hospitalization of a family member was found to be significant for only one family population—younger, better off families—and its $F$-statistic in that regression was not particularly large. Also, the magnitude of its effect on the level of the financial burden index was much smaller than the size of its effect on the total of a family's charges for health care. A younger, better off family with one hospitalization typically had a financial burden index only about 1.4 times as high as that of an otherwise identical family with no hospitalizations. In contrast, the same family typically had total charges for health care about 4 times as high.

The difference between hospitalization's great importance as a determinant of total family charges and its limited importance as a determinant of the financial burden index is not hard to explain. Inpatient care (both that provided by physicians and that provided by hospi-

tals) accounts for a very large part of total family charges for health care. Although only 30 percent of multiple-person families had a member hospitalized in 1980, inpatient care accounted for 56 percent of total charges for *all* multiple-person families (Sunshine and Dicker, 1987b). However, inpatient care is particularly well insured. Thus, costs for it are much less important in out-of-pocket expenses for health care—and hence much less important in the financial burden index—than in total charges for health care. (Inpatient care directly affects the index only through out-of-pocket expenses for health care, which are part of the index's numerator.) To be exact, comparison of NMCUES data on total family charges (Sunshine and Dicker, 1987b) and on family out-of-pocket expenses (Sunshine and Dicker, 1987a) shows that for multiple-person families, only 10 percent of total charges for inpatient care (including both hospital and physician inpatient care) were paid out of pocket in 1980, whereas 49 percent of charges for all other types of health care included in this study were paid out of pocket. As a result, inpatient care accounted for only one-fifth of all dollars spent out of pocket for health care; yet it was, as noted, responsible for over half of total charges. Put another way, each dollar of charges for inpatient care resulted in only one-fifth as much in out-of-pocket spending as did a dollar of charges for other kinds of care. Thus, hospitalization has a major effect on total charges without having nearly so great an effect on the financial burden index.

*Family illness days in bed* is the other variable that was statistically significant as a determinant of the level

of total family charges for health care in all three family populations in 1980. As Table H shows, the magnitude of its effect on total charges was positive in all three populations but relatively small. Depending on the family population examined, a 1 percent increase in the quantity, family illness days in bed plus 1, increased total family charges by between 0.11 percent and 0.19 percent. Thus, it would take a relatively large increase in family illness days in bed to produce a large increase in total charges, with the size of the increase needed varying with the population. In particular, the difference in Table H between the one-tenth of 1 percent increase for younger, better off families and the two-tenths of 1 percent increase for older families is statistically significant at the 0.05 level using a multiple $t$-test. Thus, the effect on total charges for health care of changes in family illness days in bed for older families was approximately twice that for younger, better off families. For example, for a younger, better off family, it typically took an increase of about 9 percent in bed days to increase total charges by 1 percent. For an older family, a similar increase in bed days typically would produce an increase of about 2 percent in total family charges for health care. This suggests that illness days in bed for an older family typically involve more severe illnesses requiring more expensive medical care than is true for younger families.

Family illness days in bed did not play as important a role as a determinant of a family's financial burden for health care, as measured by the financial burden index (Dicker and Sunshine, 1988), as it did as a determinant of a family's total charges for health care. As a determinant of the financial burden index, this health status variable was statistically significant in only one family population—younger, better off families—in 1980.

Seven variables were significant determinants of 1980 total family charges for health care in two of the three populations. These were family income, three major illness category variables (cancer; heart and circulatory disease; and accidents, poisonings, and injuries), a general health status variable (worst perceived health status of any family member), a health insurance variable (completeness of health care coverage), and region of the country.

The findings for family *income* are very different from what was found for this variable in the study of financially burdensome health costs (Dicker and Sunshine, 1988). Although income was one of the two most important factors determining the financial burden index (the other was type and completeness of health care coverage), it was only of moderate importance as a determinant of total family charges for health care. In regressions analyzing determinants of the financial burden index, income was the only independent variable statistically significant in all three family populations, and it had a large effect on the level of the index. For example, if one of two otherwise identical families

had half the income of the other, the former was found to have an index 1.5 to 1.8 times that of the latter.

In contrast, income was much less important as a determinant of total family charges for health care. It was statistically significant in regressions for only two family populations—older families and younger, better off families—and the size of its effect on the total of a family's charges for health care was modest. Its effect was such that if the first of two otherwise identical families had twice the income of the second, the first family was typically found to have total charges only 1.1 to 1.2 times that of the second. Thus, the size of income's effect on total charges was less than a third the size of its effect on the financial burden index.

Some of the dynamics of the creation of family financial burden are revealed by comparing the relationship between family income and total family charges for health care on the one hand, and between family income and total family out-of-pocket expenses for health on the other hand. Total family out-of-pocket expenses for health is the numerator of the financial burden index, and income's effect on total out-of-pocket expenses can be calculated from its effect on levels of the index. The relationship between income and total out-of-pocket expenses is presented in Table H in the third line under the heading "Effects of income." When this relationship is compared with the relationship of income to total charges for health care (also found in Table H, this time in the first line under the heading "Effects of income"), a different pattern is found for each of the three family populations. For older families, each 1 percent increase in income was associated with approximately a 0.15 percent increase in total charges and a 0.15 percent increase in total out-of-pocket expenses. (Recall, again, that like all regression results, these effects are the measured effects after controlling for all other variables in the regressions. In particular, differences in insurance coverage, which have a major effect on out-of-pocket expenses, are controlled for.) In other words, for older families there was a modest increase in total charges with increasing income, and total out-of-pocket expenses increased at an equal rate. This means that the percent of total charges that a family itself typically paid in 1980 (either through family-paid premiums or through its out-of-pocket payments for care) was constant across the income spectrum for this family population.

In contrast, for younger, lower income families, the finding was that for each 1 percent increase in income there was no statistically significant increase in total charges for health care, although total out-of-pocket expenses increased by 0.44 percent. Apparently as younger, lower income families increased their income modestly in 1980 (that is, by not enough to lift them out of the less-than-200-percent-of-poverty-level category) their out-of-pocket health expenses increased very substantially, while total charges for the health care they received increased only negligibly. Data presented here

do not indicate the exact sources of the large increase in out-of-pocket expenses. It may stem from the loss of charity care (for which charges are forgiven) or from decreased coverage by public programs, such as Medicaid, that have very low or no out-of-pocket costs.

In any case, such a large increase in out-of-pocket costs, if it still exists, may pose a significant disincentive to increasing earnings among the younger, lower income family population. Moreover, if this phenomenon still persists, it seems to define an economic and social inequity. Only younger, lower income families face rapid increases in out-of-pocket costs for health when their income increases. As well, they alone apparently gain no accompanying increase in care (as shown by unchanged total charges for health care).

The extent of this apparent inequity becomes clearer when the data for younger, better off families are examined, for these data show a reverse pattern. For younger, better off families, care received in 1980 (if measured by total charges) increased more rapidly with increasing income than did out-of-pocket spending for health. Among this family population, each 1 percent increase in income was accompanied by a 0.26 percent increase in total family charges for health care but only a 0.13 percent increase in total out-of-pocket expenses (Table H). This means that the proportion of total charges typically paid by a family in this category decreased as income increased. Quite possibly this phenomenon stems from increases in income among younger, better off families being associated with better employer-provided health insurance coverage, leaving these families with a smaller portion of total charges to pay out of their own resources.

Three *categories of major illnesses* were statistically significant determinants of total family charges for health care in 1980 in two of the three family populations. These categories are cancer; heart and circulatory disease; and accidents, injuries, and poisonings. A presentation of the comparative effects of these illness categories is found in Table H. This table shows that all three illness categories were statistically significant determinants of total charges for health care for older families, but this was not true for the two populations of younger families. Among younger, lower income families, only cancer was statistically significant. Among younger, better off families, both heart and circulatory disease and accidents, poisonings, and injuries were statistically significant, but cancer was not.

It is not clear from the data on hand why the types of illness significantly affecting total charges for health care should differ among the family populations examined. Family composition may be a contributing factor to differences in the determinants of total charges between the two younger family populations. Lower income families in the contemporary United States are more likely to be headed by females than better off families, and it may be the relative distribution of gender-related illnesses in each category—specifically such

illnesses as breast cancer or heart disease—that makes cancer a determinant among lower income families and heart disease a determinant among better off families. (That is, holding family size constant, the presence or absence of adult males in the family would affect the proportion of all family charges attributable to specific illnesses.) More generally, differences among the populations in the prevalence of specific illnesses may partially explain differences in findings, for each category of major illness includes several specific illnesses. Severity of the illnesses may also be an explanation, with severity varying among the different family populations. The data also suggest the possibility that the lack of statistical significance among younger, lower income families for accidents, poisonings, and injuries is due to the limited size of the sample for this population. The regression coefficient for this illness category in this family population is not different (in a statistically significant sense) from those found for the other two family populations, although the coefficient itself is not statistically significant.

For the family populations where major illness categories are statistically significant determinants of total charges, the size of the effect of a particular category of major illnesses on total charges for health care was similar across family populations except for one illness category, heart and circulatory disease. For example, families with a member suffering from accidents, poisonings, and injuries had total charges, depending on the population examined, between 25 and 31 percent higher (a statistically nonsignificant difference) than otherwise identical families with no members suffering from these illnesses. For heart and circulatory disease, however, older families had total charges 62 percent higher while younger, better off families had total charges only 20 percent higher than otherwise identical families. While the above data indicate some consistency across family populations in statistically significant findings on the effect of illness categories on the total of a family's charges for health care in 1980, the overall finding (including the statistically nonsignificant relationships) is that different combinations of illness categories and family populations showed different effects. However, whatever the explanation of differences among populations in which illnesses were statistically significant and whatever the size of their effects were, the significant findings all make sense in that the presence of an illness led, as would be expected, to an increase in total family charges rather than a decrease.

By comparison with the importance of categories of major illnesses as determinants of total family charges for health care, only one such category was a statistically significant determinant of the financial burden index in 1980, and it was significant in only one population (Dicker and Sunshine, 1988). Specifically, heart and circulatory disease was a significant determinant of the index in older families. In short, major illness categories were more important as determinants of total family

charges for health care than they were as determinants of financially burdensome family health expenses.

The general health status variable *worst perceived health status of any family member* was found to be statistically significant as a determinant of total family charges for health care in 1980 in two of the three family populations when it was reported to be "poor." These populations were younger, lower income families and younger, better off families. For younger, better off families, the variable *worst perceived health status of any family member* was also statistically significant as a determinant of total family charges for health care when it was reported to be "fair." In this population, then, the regression allows for the assessment of the relative effects of differences in the severity of bad health as a determinant of total family charges for health care. For younger, better off families with a member reported to be in poor health, total charges averaged 52 percent higher than in otherwise identical families with all members reported to be in excellent health. For families in which the worst perceived health status of any family member was reported to be only fair, not poor, the elevation of total charges averaged only 25 percent, not 52 percent. This supports the hypothesis presented above that differences in the severity of an illness may account for some of the differences found among the three family populations by type of illness category. In contrast to findings regarding total family charges for health care, none of the variables measuring perceived health status that were in the regressions were found to be statistically significant determinants of financially burdensome family health costs, as measured by the financial burden index (Dicker and Sunshine, 1988).

As with the findings for hospitalization, illness days in bed, categories of major illnesses, and perceived health status, statistically significant findings for *other health-related variables* are in the expected direction. That is, the presence of such health problems as more work-loss days due to illness or a family member with a limitation in a major activity all led to *increases* in total family charges for health care in those populations for which a statistically significant effect was found.

The effects on total family charges for health care of the *completeness of health care coverage* are the subject of extensive literature (see Manning, Newhouse, Duan, et al., 1987, for a recent summary), and findings of this study are consistent with that literature. Completeness of coverage had a statistically significant effect in both of the younger family populations, with incomplete coverage leading, as expected, to lower total charges.

The parallel study of the financial burden index (Dicker and Sunshine, 1988) showed that incomplete coverage lowered the index. This means that families with incomplete coverage used so much less health care than comparable families with complete coverage that the former families' total out-of-pocket expenses fell even though the amount they had to pay out of pocket

per service increased. (In economists' terms, the price elasticity of demand was greater than one.) This represents more of a change in the use of health care than the literature usually has reported.

The finding that *race* affected total charges, with black families having lower total charges in 1980 than comparable white families among the younger, better off family population, raises concerns. If this pattern remains true, there may be substantial racial discrimination in access to health care in the United States. The regression analysis shows that black families in the younger, better off population had total charges for health care 31 percent lower than those of otherwise identical white families, and the fact that the regression controls for the effects of income, health status, family structure, location of residence, and education (among other variables) makes it more likely that the difference was genuinely a racial one. A similar difference was found in the financial burden index in 1980 (Dicker and Sunshine, 1988). Concern about possible discrimination is reinforced by the finding that total charges for older black families were also about 30 percent below those for older families of other races, although this difference was not statistically significant.

Other *sociodemographic variables* also showed statistically significant effects in 1980 only in the younger, better off family population. This may reflect the fact that the number of families in this population in NMCUES was about three times that in either of the other two multiple-person family populations, for differences of a given magnitude generally have greater statistical significance in larger samples. In any case, the increase in total charges with *age* among the younger, better off family population was in the expected direction and possibly reflects health status effects not captured by other variables.

The increase in total charges accompanying more *education* of the family head (found among younger, better off families) is a familiar finding but, because the regression controls for the effect of income, in this study it most likely is a genuine effect of education. Education and income are correlated and apparent effects of education in analyses that do not control for income— for example, in most frequency tables—may be primarily income-related effects. The effects of education found in this study may be cultural, reflecting differences in family valuation of health and health care that accompany differences in education. And, indeed, cultural differences may also be partially responsible for the race-linked differences in total charges described just above.

Analyses of determinants of the financial burden index (Dicker and Sunshine, 1988), like the analyses of determinants of total charges presented here, showed increases in the index with increasing family age and education. Indeed, the effects of these two variables on the financial burden index were statistically significant in both populations of younger families.

The higher total charges experienced by younger,

better off families with *children* than by comparable families with no children are more puzzling. This pattern of total charges does not occur in other family populations, nor does an effect of children appear in analyses of the financial burden index (Dicker and Sunshine, 1988). The finding is not related to family size, for the regressions include, and therefore control for, family size. This means that the observed effect is a difference found between families with equal numbers of members but with one having children and the other not having children. In such a comparison, the family without children must have more adults (in order to have an equal total number of members), and might be expected to have higher, not lower, total charges because children use less health care than adults.

The findings with respect to *region* are complex. Younger, better off families had higher total charges in the West than elsewhere, while younger, lower income families had lower total charges in the South than elsewhere. The latter finding may reflect regional differences in the availability of public or charity care to lower income families, or in its price. The finding for younger, better off families presumably reflects differences in practice patterns or in the prices for each health care service (for example, the price of a physician visit). It is, as might be expected, paralleled by a finding for older families which, however, is not statistically significant.

## Patterns Among Variables

One important aspect of the findings of this report is the restricted range of factors that are found to be statistically significant determinants of total family charges for health care. The study analyzed the effects of family variables that can be classified into seven categories: demographic, sociocultural, major illnesses and special health events (such as hospitalization), general health status, economic, health insurance, and geographic. For older families, variables in only three of these categories were found to be significant; for younger, lower income families, variables in only four categories were significant; and for both these populations, two of the significant categories were the closely related categories of (1) general health status and (2) major illnesses/special health events. In contrast, statistically significant determinants of the financial burden index included variables from at least five categories for all three family populations (Dicker and Sunshine, 1988).

However, for the younger, better off family population, variables in all seven categories were significant determinants of both total charges and the financial burden index. Given the points made a few paragraphs above about sample sizes and statistical significance, as well as the relatively demanding tests of significance used in the regressions in this study ($p$ of 0.0022 or less for individual variables), the range of variables found significant for the other two family populations might well have been more extensive if the NMCUES samples of these populations had been larger.

Another way to assess the importance of the findings from the regression analyses for total family charges is to examine how the statistically significant independent variables in the regressions were distributed among the three family populations. Table J shows that 15 variables were significant in one or more of the three family socioeconomic populations. (This counts hospitalization

## Table J

Statistically significant variables from a set of regressions for total family charges for health care arranged by the number of family socioeconomic populations in which each variable was statistically significant

| | Statistically significant in— | | |
|---|---|---|---|
| Variable | 3 populations | 2 populations | 1 population |
| Hospitalization | O, L, B | | |
| Family illness days in bed | O, L, B | | |
| Cancer | | O, L | |
| Heart and circulatory disease | | O, B | |
| Accidents, poisonings, and injuries | | O, B | |
| Family income | | O, B | |
| Perceived health status | | L, B | |
| Region | | L, B | |
| Completeness of health care coverage | | L, B | |
| Limitation in major activity | | | L |
| Family work-loss days due to illness | | | B |
| Presence of child | | | B |
| Age of head of family | | | B |
| Race of head of family | | | B |
| Education of head of family | | | B |

NOTES: O represents older families. L represents younger, lower income families. B represents younger, better off families.

as one variable.) Two variables were statistically significant in all three populations, seven others were statistically significant in two of the populations, and another six were statistically significant in only one family population. Table J thus gives one measure of the importance of variables as determinants of total family charges for health care in the U.S. family population in 1980. This measure is the universality of the effects of the variables in major subpopulations of the multiple-person family population.

From this perspective, health variables are clearly most important. Of the 15 variables in Table J, 8 are health variables—that is, variables involving general health status, major illnesses, or special health events. Moreover, as Table J shows, the health variables include all the variables found statistically significant in three populations and four of the seven variables found significant in two populations, while only two of the health variables are among the six variables statistically significant solely in one family population.

How large a role health variables play as determinants of total charges becomes even more apparent when Table J is compared with the findings from the companion regression analysis of the statistically significant determinants of the financial burden index (Dicker and Sunshine, 1988). Table K shows the variables found significant in that analysis and the populations for which each was significant. Only 4 of the 16 variables found significant as determinants of the financial burden index were health variables, and all 4 were significant for only 1 family population. An examination of Table K shows that income and health insurance variables were the most important determinants of the financial burden index, in marked contrast to the predominant role of health variables as determinants of total family charges for health care.

The particularly prominent role of hospitalization among the several health variables that are statistically significant determinants of total family charges is puzzling. A plausible expectation would be that variables involving general health status and major illnesses would be the main health variables affecting total charges, for they give rise to the need for health care. Thus it might be expected that hospitalization would not be prominent as a determinant of total charges.

Moreover, one likely explanation of hospitalization's actual prominence appears untrue. This explanation is that hospitalization's importance as a determinant of total charges reflects differences in severity of illness not measured by general health status variables or major illness variables but measured by hospitalization. (For example, a family with a member with heart or circulatory disease who is hospitalized may have a far more serious illness on its hands than a family with a member with these diseases but no hospitalizations.) However, if an important determinant of total charges is, in fact, severity of illness not measured by general health status or major illness variables, then the variable reflecting the death of a family member would be expected to show statistical significance. To note the obvious, death clearly reflects an extreme severity of illness. Surprisingly, however, the regressions did not find the death of a family member to be a significant determinant of total family charges for any of the three family populations. This negative finding is unexpected both because other health status variables were significant and because an extensive study of older individuals found that those who died had medical expenses in the last year of their life several times as high as the annual expenses of those who did not die (Lubitz and Prihoda, 1984). (Note that, unlike the regressions, this study did

## Table K

**Statistically significant variables from a set of regressions for the index of financially burdensome family health care expenses arranged by the number of family socioeconomic populations in which each variable was statistically significant**

| Variable | Statistically significant in— | | |
|---|---|---|---|
| | 3 populations | 2 populations | 1 population |
| Family income | O, L, B | | |
| Head-spouse structure | | O, L | |
| Coverage by Medicaid | | O, L | |
| Source of health care coverage unknown | | O, L | |
| Other public coverage | | L, B | |
| Other public and private coverage | | L, B | |
| Age of head of family | | L, B | |
| Education of head of family | | L, B | |
| Region | | O, B | |
| Medicare and other public coverage | | | O |
| Heart and circulatory disease | | | O |
| Family work-loss days | | | L |
| Family illness days in bed | | | B |
| One or more hospitalizations | | | B |
| Race of head of family | | | B |
| Completeness of health care coverage | | | B |

NOTES: O represents older families. L represents younger, lower income families. B represents younger, better off families.

not simultaneously control for the effect of multiple other variables.) However surprising, the finding that death is not statistically significant in the regressions is nonetheless clear, and this finding makes it unlikely that hospitalization's great significance as a determinant of total charges reflects a role for it as a measure of severity of illness.

Two other explanations of hospitalization's major significance seem more likely. One is that hospitalization is a particularly costly way to care for illnesses and thus families with the same health problems (including equal severity of illness) have total charges that differ greatly depending on whether or not these problems are treated in the inpatient hospital setting. This explanation is consistent with much of the conventional wisdom in the fields of health care cost control and utilization management. (See, for example, Ginsburg and Sunshine, 1987.) The second likely explanation is that it is the actual use of health care services that directly and *proximately* gives rise to total charges, even though bad health is the dominant factor in the use of these services. This fact is likely to lead mathematically to a particularly strong statistical relationship between hospitalization— that is, the use of a particular type of health care service— and total charges, especially since hospitalization (as noted above) accounts for more than half of all total charges for family health care. Both of these explanations, incidentally, could partially explain the absence of a statistically significant effect of death on total charges.

## Concluding Remarks

Whatever the reasons for the relative importance of different health variables, it is reassuring to find that health variables, collectively, were the predominant determinants of total family charges for health care in 1980, as Table J shows. This finding suggests that overall in the United States, the need for health care has been the most important factor in determining how much care families receive.

True, the regressions do not absolutely establish this conclusion. For one thing, this conclusion requires an assumption that the amount of health care a family receives generally parallels its total charges, and there clearly is some difference between the amount of care and total charges due to differences in the price of health care. Second, it should be noted that there were a number of variables in the preferred models that were not statistically significant in any of the three socioeconomic family populations covered in this study. Whether these variables would be statistically significant in other populations or with a larger sample cannot be definitively assessed at this time, although we have presented suggestive evidence that larger samples would lead to more variables being found statistically significant. Moreover, it should also be noted that a single regression analysis

that encompassed the entire U.S. multiple-person family population might produce different findings about the relative importance of variables than do the three separate regressions reported here. So might an analysis that did not involve the exclusions made in the regressions. (The most important of these exclusions are one-person families, families with zero or very low incomes, families with zero total charges, institutionalized persons, and charges for long-term care.) Nonetheless, it seems reasonable to interpret the findings of this study as showing that the need for health care was the principal factor in determining how much care U.S. families received in 1980.

It seems appropriate that the need for health care should be the principal determinant of how much care U.S. families receive. This is especially true as compared, for example, with a possible alternative health care system in which income and insurance were the primary determinants of care received. Health care coverage and income *are* the primary determinants of the burden on families of health care expenses, as measured by the financial burden index. (See Table K and Dicker and Sunshine, 1988.)

However, although the pattern of receipt of care and of total charges does seem basically appropriate in that need-related factors are its most important determinants, troubling findings should not be ignored. First, there is considerable evidence that in 1980 black families received less care than similar white families, suggesting the existence of race-related barriers in access to care. Second, younger, lower-income families experienced rapid increases in out-of-pocket costs for health care as their income increased and benefited from no accompanying increase in care. This raises not only an equity issue, but also suggests these families may have faced important work disincentives in the form of out-of-pocket health care costs. The work disincentives would arise because increasing income, which would be achieved by more work, led to no more health care and to increased out-of-pocket expenses for health care. The increased out-of-pocket expenses cut into the increase in discretionary income available from more work, making the financial return from working relatively small.

Third, the analysis suggests that inpatient care, which accounts for more than half of all health care costs recorded in NMCUES, may frequently have represented a particularly expensive care modality rather than a necessary response to severe health problems.

These troubling findings from 1980 may no longer characterize the U.S. health care system. Indeed, the very substantial decline in inpatient care seen during the 1980's in response to cost containment measures indicates some amelioration of the inpatient care cost problem. However, to the extent that these findings have persisted, they point to the possible continued existence of serious problems despite the appropriateness, in broad outline, of the pattern of receipt of health care in the United States.

# References

Aday, L. A., Fleming, G. V., and Andersen, R.: Access to medical care in the U.S.: Who has it, who doesn't. *Research Series*. No. 32. Center for Health Administration Studies, University of Chicago. Chicago. Pluribus Press, 1984.

Andersen, R., and Benham, L.: Factors Affecting the Relationship Between Family Income and Medical Care Consumption, in H. E. Klarman, ed., *Empirical Studies in Health Economics*. Baltimore. The Johns Hopkins Press, 1970.

Anderson, J. M., and Thorne, E.: Estimates of Aggregate Personal Health Care Expenditures in 1980. Comparison of the National Health Accounts and the National Medical Care Utilization and Expenditure Survey Data. Working Paper No. 30, Division of Health Interview Statistics, National Center for Health Statistics, Hyattsville, Md., 1985.

Arnold, S. F.: *Theory of Linear Models*. New York. John Wiley and Sons, 1981.

Berki, S. E.: A look at catastrophic medical expenses and the poor. *Health Affairs* 5(4):138–145, Winter 1986.

Berki, S. E., Wyszewianski, L., Magilavy, L. J., et al.: Families with high out-of-pocket health service expenditures relative to their income. Final Report on Phase II of Contract No. 233–81–3022. Rockville, Md. National Center for Health Services Research and Health Care Technology Assessment, Public Health Service, Jan. 30, 1985.

Binder, D. A.: On the variances of asymptotically normal estimators from complex surveys. *Int. Statistical Rev.* 51:279–292, 1983.

Bonham, G. S.: Procedures and questionnaires of the National Medical Care Utilization and Expenditure Survey. *National Medical Care Utilization and Expenditure Survey*. Series A, Methodological Report No. 1. DHHS Pub. No. 83–20001. Public Health Service. Washington. U.S. Government Printing Office, Mar. 1983.

Buczko, W.: Physician utilization and expenditures in a medicaid population. *Health Care Financ. Rev.* 8(2):17–26, Winter 1986.

Campbell, D. T., and Julian, C. S.: *Experimental and Quasi-Experimental Designs for Research*. Chicago. Rand McNally College Publishing Co., 1980.

Catastrophic Illness Expenses. Department of Health and Human Services report to the President. U.S. Department of Health and Human Services, Nov. 1986.

Cox, G., Parker, A. E., Sweetland, S. S., et al.: *Imputation of Missing Item Data for the National Medical Care Utilization and Expenditure Survey*. Prepared for the National Center for Health Statistics under Contract No. HRA–233–79–2032. Research Triangle Park, N.C. Research Triangle Institute, 1982.

Cox, G., and Sweetland, S. S.: *Imputation of Attribution-Related Missing Data for the National Medical Care Utilization and Expenditure Survey*. Prepared for the National Center for Health Statistics

under Contract No. HRA–233–79–2032. Research Triangle Park, N.C. Research Triangle Institute, 1982.

Dicker, M.: Health care coverage and insurance premiums of families. United States, 1980. *National Medical Care Utilization and Expenditure Survey*. Preliminary Data Report No. 3. DHHS Pub. No. 83–20000. Public Health Service. Washington. U.S. Government Printing Office, May 1983a.

Dicker, M.: Panel Surveys and Family Level Measures, Problems and Solutions, in *1983 Proceedings of the American Statistical Association*. Social Statistics Volume. Washington. American Statistical Association, 1983b.

Dicker, M.: Demographic Analysis and a General Theory for Measuring Social Change in Organizations. The Case of the Family. Paper presented at the annual meeting of the Southern Regional Demographic Group. Orlando, Fla., Oct. 19, 1984.

Dicker, M., and Casady, R. J.: A Reciprocal Rule Model for Defining Longitudinal Families for the Analysis of Panel Survey Data, in *1982 Proceedings of the American Statistical Association*. Social Statistics Volume. Washington. American Statistical Association, 1982.

Dicker, M., and Casady, R. J.: Empirical Findings on the Distribution and Construction of Longitudinal Families, Part I, Modeling the Universe, Inscope Changes, and Static and Dynamic Families, in *1984 Proceedings of the American Statistical Association*. Survey Research Methods Volume. Washington. American Statistical Association, 1984.

Dicker, M., and Casady, R. J.: An Introductory Discussion of Issues in the Methodology of Longitudinal Family Surveys. Paper presented at the All Day Measurement Workshop of the Family Health Section at the annual meeting of the National Council on Family Relations. Dallas, Texas, 1985.

Dicker, M., and Sunshine, J. H.: Family Use of Health Care, United States, 1980. *National Medical Care Utilization and Expenditure Survey*. Series B, Descriptive Report No. 10. DHHS Pub. No. 87–20210. Hyattsville, Md. National Center for Health Statistics, Public Health Service, 1987.

Dicker, M., and Sunshine, J. H.: Determinants of financially burdensome family health expenses, United States, 1980. *National Medical Care Utilization and Expenditure Survey*. Series C, Analytical Report No. 6. DHHS Pub. No. 88–20406. Public Health Service. Washington. U.S. Government Printing Office, Apr. 1988.

Draper, N. R., and Smith, H.: *Applied Regression Analysis*. New York. John Wiley and Sons, 1982.

Duan, N., Manning, W. G., Jr., Morris, C. N., et al.: *A Comparison of Alternative Models for the Demand for Medical Care*. Contract No. R–2754–HHS. Santa Monica, Calif. Rand Corporation, Jan. 1982.

Farley, P. J.: Private insurance and public programs, Coverage of health services. *National Health Care Expenditures Study*. Data Preview 20. DHHS Pub. No. (PHS) 85–3374. Rockville, Md. National Center for Health Services Research and Health Care Technology Assessment, Mar. 1985.

Farley, P. J.: Hospital and ambulatory services for selected illness. *Health Serv. Res.* 21(5):587–634, Dec. 1986.

Feldstein, P. J.: *Health Care Economics,* 2d ed. New York. John Wiley and Sons, 1983.

Ginsburg, P. B., and Sunshine, J. H.: *Cost Management in Employee Health Plans.* Contract No. R–3543–RWJ. Santa Monica, Calif. Rand Corporation, Oct. 1987.

Hershey, J. C., Luft, H. S., and Gianaris, J. M.: Making sense out of utilization data. *Med. Care* 13(10):838–852, Oct. 1975.

Holt, M. M., and Shah, B. V.: *SURREGR, Standard Errors of Regression Coefficients From Sample Survey Data.* Research Triangle Park, N.C. Research Triangle Institute, Apr. 1982.

Kaspar, J. A., Walden, D. C., and Wilensky, G. C.: Who are the uninsured? *National Health Care Expenditures Study*. Data Preview 1. Hyattsville, Md. National Center for Health Services Research, Public Health Service, 1980.

Koretz, D.: *Catastrophic Medical Expenses: Patterns in the Non-Elderly, Non-Poor Population.* Washington, D.C. Congressional Budget Office, Dec. 1982.

Kovar, M. G.: Expenditures for the medical care of elderly people living in the community in 1980. *Milbank Mem. Fund Q.* 64(1):100–132, 1986.

Landis, J. R., Lepkowski, J. M., Eklund, S. A., and Stehouwer, S. A.: A statistical methodology for analyzing data from a complex survey, the first National Health and Nutrition Examination Survey. *Vital and Health Statistics.* Series 2, No. 92. DHHS Pub. No (PHS) 82–1366. National Center for Health Statistics, Public Health Service. Washington. U.S. Government Printing Office, Sept. 1982.

Levy, P. S., and Lemeshow, S.: *Sampling for Health Professionals.* Belmont, Calif. Lifetime Learning Publications, 1980.

Lubitz, J., and Prihoda, R.: Use and costs of medicare services in the last two years of life. *Health Care Financ. Rev.* 5(3):117–131, Spring 1984.

Mallows, C. L.: Some comments on C(p). *Technometrics* 15:661–675, 1973.

Manning, W. G., Newhouse, J. P., Duan, N., et al.: Health insurance and the demand for medical care: Evidence from a randomized experiment. *Amer. Econ. Rev.* 77(3):251–277, June 1987.

McCarthy, P. J.: Replication, an approach to the analysis of data from complex surveys. *Vital and Health Statistics.* Series 2, No. 14. DHEW Pub. No. (PHS) 79–1269. National Center for Health Statistics, Public Health Service. Washington. U.S. Government Printing Office, Apr. 1966.

McMillen, D. B., and Herriot, R.: Towards a Longitudinal Definition of Households. SIPP Working Paper Series No. 8402. U.S. Bureau of the Census. Washington, D.C., 1984.

Moser, B., Whitmore, R., Frick, G. G., et al.: *National Medical Care Utilization and Expenditure Survey: Analytic Family File Construction Methodology Report.* RTI Project No. 251U–1898–14. Research Triangle Park, N.C. Research Triangle Institute, Oct. 1983.

Neter, J., and Wasserman, W.: *Applied Linear Statistical Models.* Homewood, Ill. Richard D. Irwin, Inc., 1974.

Newacheck, P. W.: *Utilization and Expenditures for Medical Care Services Provided to Children with Activity Limitations.* Grant MCJ–063468–01–0. Prepared for National Maternal and Child Health Resource Center. Institute for Health Policy Studies, University of California, San Francisco, 1985a.

Newacheck, P. W.: *Prevalence and Severity of Chronic Conditions Among Children.* Prepared for National Maternal and Child Health Resource Center. Institute for Health Policy Studies, University of California, San Francisco, 1985b.

Newhouse, J. P., Manning, W. G., Morris, C. N., et al.: Some interim results from a controlled trial of cost sharing in health insurance. *N. Engl. J. Med.* 305:1501–1507, 1981.

Public Use Data Tape Documentation: Family Data. *National Medical Care Utilization and Expenditure Survey, 1980.* Hyattsville, Md. National Center for Health Statistics, Public Health Service. Nov. 1986.

SAS Institute, Inc.: *SAS Users Guide, Basics,* 1982 ed. Cary, N.C. SAS Institute, Inc., 1982.

SAS Institute, Inc.: *SAS/STAT Guide for Personal Computers, Version 6 Edition.* Cary, N.C. SAS Institute, Inc., 1985.

Shah, B. V.: *SESUDAAN, Standard Errors Program for Computing Standardized Rates From Sample Survey Data.* Research Triangle Park, N.C. Research Triangle Institute, Apr. 1981.

Sunshine, J. H.: Medicare, past, present, and future. *Natl. J.* 14(23):1030–1033, June 5, 1982.

Sunshine, J. H.: How many Americans lack outside sources of payment for major health care costs? *J. Health Hum. Resour. Adm.* 6(3):341–360, Winter 1984.

Sunshine, J. H., and Dicker, M.: Family out-of-pocket expenditures for health care: United States, 1980. *National Medical Care Utilization and Expenditure Survey.* Series B, Descriptive Report No. 11. DHHS Pub. No. 87–20211. National Center for Health Statistics, Public Health Service. Washington. U.S. Government Printing Office, Aug. 1987a.

Sunshine, J. H., and Dicker, M.: Total family expenditures for health care: United States, 1980. *National Medical Care Utilization and Expenditure Survey.* Series B, Descriptive Report No. 15. DHHS Pub. No. 87–20215. National Center for Health Statistics, Public Health Service. Washington. U.S. Government Printing Office, Sept. 1987b.

Taube, C. A., Kessler, L. G., and Burns, B. J.: Estimating the probability and level of ambulatory mental health services use. *Health Serv. Res.* 21(2, Pt II):321–340, June 1986.

Whitmore, R. W., Cox, B. G., and Folsom, R. E.: *Family Unit Weighting Methodology for the National Household Survey Component of the National Medical Care Utilization and Expenditure Survey.* Contract No. HRA–233–79–2032. Prepared for the National Center for Health Statistics. Research Triangle Park, N.C. Research Triangle Institute, 1982.

Wilensky, G. R., and Walden, D. C.: Minorities, Poverty, and the Uninsured. Paper presented at the 109th Annual Meeting of the American Public Health Association. Los Angeles, Nov. 1981.

Wyszewianski, L.: Financially catastrophic and high-cost cases: Definitions, distinctions, and their implications for policy formation. *Inquiry* 23(4):382–394, Winter 1986a.

Wyszewianski, L.: Families with catastrophic health care expenditures *Health Serv. Res.* 21(5):617–634, Dec. 1986b.

# Appendixes

## Contents

**List of Appendix Tables**

34

# Appendix I
# Technical Notes on Regression Methods

## Introduction

Multiple regression analysis is a statistical method for examining the effect of a set of independent (or causal) variables on a continuous dependent variable. It permits the analysis of the effects of a large number of independent variables, both continuous and categorical, and provides separate estimates for the effects of each. The set of variables used in the regression analysis, together with the functional form of the equation that relates the independent variables to the dependent variable, is called the regression model. By analyzing the coefficients in the regression model, one can explore relationships between the dependent and independent variables and make predictions about the future behavior of the independent variable.

This appendix first presents a technical description of multiple regression analysis with special attention given to the application of this technique to complex surveys (such as the National Medical Care Utilization and Expenditure Survey, NMCUES, which is the source of data used in this report). The appendix then presents a series of somewhat less technical sections. These cover, in turn, the regression model used in this report, the interpretation of regression coefficients and of means of variables, and the analytic procedures followed in this report.

## Technical Description of Multiple Regression Analysis

### Introduction

This section discusses the use of multiple regression analysis with special emphasis on its application to complex survey data and, in particular, on its use in analyzing the NMCUES data of this study. The following topics are covered:

- The basic structure of the regression model and how it is usually applied in finite population sampling,

- Estimation of regression parameters using complex survey data, and

- Estimation of variances of parameter estimates.

## The Regression Model in Finite Population Sampling

In most statistical literature, the regression model is stated as

$$y_i = \underline{x_i}' \underline{b} + e_i \qquad (1)$$

where $y_i$ is an observable random variable, $\underline{x_i}$ is a $p \times 1$ vector of independent variables, $\underline{b}$ is a $p \times 1$ vector of unobservable regression coefficients, and $e_i$ is a random variable with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. The sequences $y_1, y_2,\ldots$ and $e_1, e_2,\ldots$ are generally assumed to be independent and identically distributed. In the normal regression model, it is further specified that $e_i$ has a normal distribution.

In finite population sampling, the regression model looks similar but is formulated in a fundamentally different way. When sampling from a finite population, it is assumed that there is a population of N pairs $(y_i, \underline{x_i})$:

$$P = \{(y_1, \underline{x_1}),(y_2, \underline{x_2}),\ldots,(y_N, \underline{x_N})\} \qquad (2)$$

from which a sample of n pairs is selected. Note that $y_i$ is considered to be fixed in finite population sampling and the randomness is introduced by the selection process, whereas each $y_i$ in the infinite population regression model is considered to be a random variable. The same is true of the $e_i$. In the finite population setting, the expression "independent and identically distributed" has no real meaning.

The regression coefficients are determined in the finite population setting by the least squares equation:

$$\underline{b} = (X'X)^{-1}X'\underline{y}, \qquad (3)$$

where X is the matrix whose rows are made up of the $\underline{x_i}'$. In the infinite population situation neither the $X'X$ nor $X'\underline{y}$ exist in any meaningful sense, because each would consist of divergent infinite sums. (Equation (3) would, however, be used in a sample from an infinite population to estimate the regression coefficients.)

Finally, the error terms $e_i$ in the finite population setting are defined as the residuals from the least squares equation in (3):

$$e_i = y_i - \underline{x_i}' \underline{b}. \qquad (4)$$

Thus, in sampling from a finite population, the $e_i$ can be seen as a population of fixed values, from which a sample of size $n$ is drawn. Again, this contrasts with the usual regression model, where each $e_i$ is considered to be random.[1]

## Estimation of Regression Parameters

In equation (3), it was indicated that $\underline{b}$ is given by

$$\underline{b} = (X'X)^{-1} X' \underline{y}.$$

The elements of the matrix $X'X$ are given by

$$z_{ij} = \sum_k x_{ki} x_{kj} \qquad (5)$$

and the elements of the vector $X'\underline{y}$ are given by:

$$t_i = \sum_k x_{ki} y_k. \qquad (6)$$

To estimate $\underline{b}$, then, it is necessary to estimate $z_{ij}$ and $t_i$. Let

$$w_i = \text{sampling weight of } i\text{th sample unit.} \qquad (7)$$

Unbiased estimates of $z_{ij}$ and $t_i$ are given by

$$\hat{z}_{ij} = \sum_k x_{ki} x_{kj} w_k \qquad (8)$$

and

$$\hat{t}_i = \sum_k x_{ki} y_k w_k. \qquad (9)$$

It must be noted that although these estimates are unbiased, the estimate of $\underline{b}$ obtained by forming

$$\underline{\hat{b}} = \hat{Z}^{-1} \hat{\underline{t}} \qquad (10)$$

is not. Nor is the variance of this estimate easy to calculate, as the next section shows.

## Estimating Variances of Regression Coefficients

When sampling from an infinite population, the covariance matrix of $\underline{\hat{b}}$ is given by

$$\text{Cov}_\infty(\underline{\hat{b}}) = (X'X)^{-1}\sigma^2 \qquad (11)$$

where $\sigma^2$ is the variance of the independent variable $y$. When sampling from a finite population, this formula is somewhat more complicated.

Define the vector $\underline{u}$ by

$$u_i = \sum_k (y_k - x_k'\underline{b}) x_{ki} w_k. \qquad (12)$$

Notice that $u_i$ is a Horvitz-Thompson type sum whose variance can be calculated using the familiar rules of stratified and/or cluster sampling. We can now write a formula for the approximate covariance matrix of $\underline{\hat{b}}$ as

$$\text{Cov}(\underline{\hat{b}}) = (X'X)^{-1}\hat{\Sigma}(\underline{u})(X'X)^{-1} \qquad (13)$$

where $\hat{\Sigma}(\underline{u})$ is an estimate of $\text{Cov}(\underline{u})$. (See Binder (1983), pp. 279–292.)

This method of variance estimation is often called "Taylorizing" or "linearizing," because the Taylor expansion is used to develop the linear approximation on which equation (13) is based. SURREGR (Holt and Shah, 1982), the computer software package used in the final steps of the regression analysis of this report, uses this "Taylorizing" or linearizing technique.

## The Model

### Introduction

In this section, the model used in the regression analysis reported in this study is discussed. Three topics are covered in turn: first, the variables used in the regression analysis; second, the functional form of the hypothesized relationship between the dependent variable and the independent variables; and third, the rationale for using this functional form.

### Variables Used in the Regression Analysis

The variables used in the regression analysis are listed and described in Table I, which appears at the end of this appendix. The 48 variables in this table were either taken directly from or constructed from the variables on the NMCUES family data tape. Total family charges for health care was selected as the dependent variable of interest for this report. The 47 independent variables were chosen from the larger set of family variables available from the NMCUES on the basis of previous research and the desire to include the variables believed likely to have the greatest explanatory power. In developing a regression model, the number of variables is limited by the amount of data available; screening out unnecessary variables greatly facilitates the analysis. The 48 variables are discussed further in the text of this report, in the chapters titled "Introduction" and "Data and Methods."

---

[1] It should be noted that this is the "classical" point of view in survey sampling. Increasingly, superpopulation models are used in survey inference. They assert that the finite population under study is simply a large "sample" from an infinite "superpopulation."

## Functional Form of the Regression Equation

The functional form of the relationship underlying the regression equation used in this report is

$$y_i = \prod_{j=1}^{s} (x_{ij}^{b_j}) \, \mathrm{EXP}\{b_o + \sum_{j>s} b_j x_{ij} + e_i\}, \qquad (14)$$

where $s$ is the number of independent variables transformed into their natural logarithm in the estimating equation (see below). As in equation (1), $y_i$ is the dependent variable (in this report, it is total family charges for health care), the $x_{ij}$ are the independent variables, the $b_i$ are the regression coefficients, and $e_i$ is the error term with $\mathrm{E}(e_i) = 0$. (The notation EXP means that "$e$," the base of natural logarithms $(=2.71828...)$, is to be raised to the power indicated by the expression in braces that follows EXP.) Expanding the products in equation (14) yields

$$y_i = (x_{i1}^{b1})(x_{i2}^{b2})(x_{i3}^{b3}) \ldots \qquad (14a)$$

$$(\mathrm{EXP}\{b_o\})(\mathrm{EXP}\{b_{s+1} x_{i,s+1}\})$$

$$(\mathrm{EXP}\{b_{s+2} \, x_{i,s+2}\}) \ldots (\mathrm{EXP}\{e_i\}).$$

Equation (14) and equation (14a), which are mathematically equivalent, are not linear in the regression coefficients (the $b_i$) and so cannot be estimated in the fashion described earlier in this appendix. However, these equations have a number of desirable features, as described below, and were selected because of these features.

One desirable feature is that they are easily transformed into equations that *are* linear in the regression coefficients. (For more on transformation of variables in regression analysis, see Neter and Wasserman, 1974, pp. 123–127.) The transformation necessary to achieve linearity consists of taking the natural logarithm of both sides of the equations. Taking the natural logarithm of both sides of equation (14) yields

$$\ln(y_i) = b_o + \sum_{j=1}^{s} (b_j \ln(x_{ij})) \qquad (15)$$

$$+ \sum_{j>s} (b_j \, x_{ij}) + e_i.$$

Similarly, taking the natural logarithm of both sides of equation (14a) yields

$$\ln(y_i) = b_o + b_1 \ln(x_{i1}) \qquad (15a)$$

$$+ b_2 \ln(x_{i2}) + \ldots$$

$$+ b_{s+1} x_{i,s+1} + b_{s+2} \, x_{i,s+2}$$

$$+ \ldots + e_i.$$

Equation (15) and equation (15a), which are algebraically the same, *are* linear in the regression coefficients (the $b_j$) and are the regression equation used in the analyses in this report. Their parameters and the variances of these parameters were estimated by the techniques described in the first section of this appendix.

A point to note here with respect to equation (15) is that, in this equation, some of the original NMCUES variables are transformed into their natural logarithms. That is, these variables are replaced by their natural logarithms. Table I indicates the variables for which the natural logarithm, rather than the untransformed variable, was used.

Also, equations (14) and (15) require any categorical independent variables to be expressed in numerical form. This is readily accomplished for variables that take on only two values. For these variables, it is accomplished by assigning "1" to one of the values and "0" to the other. For example, sex of the head of the family was coded as female $= 1$ and male $= 0$. For categorical variables that take $k \geq 3$ values, an extension of this procedure was used. Such categorical variables were represented in the regression equation by a series of $k-1$ "dummy" variables, each of which can take the value of "1" or "0." Each $k-1$ value of the original categorical variable was associated with one dummy variable, which was assigned the value of 0 or 1 for each family in the sample depending on whether the value (of the original categorical variable) was true (dummy $= 1$) or not (dummy $= 0$). For example, the head-and-spouse structure of a family takes on three values: (1) head and spouse always present, (2) family always has only a head, and (3) changing head-and-spouse structure. Dummy variables, D1 and D2, were created from the second and third of these three values, while the first value was the omitted value. This procedure is sometimes called "blocking" (Draper and Smith, 1982, p. 241)

In creating dummy variables, one of the original $k$ values of the original variable must be omitted, because the $k$th dummy variable would be a linear combination of the first $k-1$ dummy variables. (If one independent variable in a regression is a linear combination of others, the matrix X'X (see equation (3)) cannot be inverted and the regression coefficients are thus undefined.) Typically, the omitted value was the one regarded conceptually as the "base case" or was the most common state. Table I shows the dummy variables that were used and indicates which value was omitted.

### Rationale for the Functional Form

*The dependent variable.* The functional form shown in equation (14) requires that the dependent variable be used in logarithmic form when estimating the regression model using equation (15). This functional form was chosen for three reasons.

First, it is believed that the relationship between

the independent and dependent variables is primarily multiplicative. For example, it was expected that the reduction in total family charges found for families lacking health care coverage would be multiplicative rather than additive—that is, would reduce total charges by a given percentage rather than by a given dollar amount. To simplify greatly, in an additive model, if the absence of coverage reduced total charges by $600—from $1,500 to $900—for some families, it should reduce total charges from $500 to minus $100—an obvious absurdity—for other families. In contrast, a multiplicative relationship would, following this example, generally reduce total charges to 60 percent of their with-insurance level—that is to $300, a plausible figure, for the second group of families. A multiplicative effect of this type is what the authors believed likely. When underlying relationships are multiplicative, a functional form with a logarithmic dependent variable in the regression model (that is, in equation (15)) is appropriate because such a functional form is multiplicative in its untransformed version (that is, equation (14)).

Second, a logarithmic dependent variable was used because prior research in evaluating the appropriateness of different functional forms for regression analysis of medical expenditure data indicates that a logarithmic dependent variable should be used in the estimating equation. Duan et al. (1982) carried out an extensive analysis of residuals from various models and found that they approximated a normal distribution (as assumed by the linear regression model) more closely when a logarithmic dependent variable was used than when an untransformed dependent variable was used in the regression model.

Third, the literature on health expenditures, perhaps because of the preceding two reasons, almost always uses a logarithmic dependent variable in regression equations. Some examples, which are similar to this report in the data bases they use or the subjects they investigate, are Taube et al. (1986) and Farley (1986). By following the literature, comparability of results is enhanced.

*Independent variables.* Given a logarithmic dependent variable in the estimating equation (equation (15)), there remains the question of whether or not to transform the (continuous) independent variables in the equation into their logarithms. The alternatives of carrying out such a transformation or not doing so imply different relationships. The following paragraphs first describe in nontechnical terms the relationships implied by each alternative and then describe the choices made.

When the dependent variable is logarithmic, a logarithmic independent variable in the estimating equation implies constant elasticity. In nontechnical terms, this means that a 1 percent increase in the untransformed independent variable produces a fixed percent increase (or decrease) in the untransformed dependent variable, with this increase (or decrease) called the elasticity. (Technically, elasticity is defined as $(\partial y_i / \partial x_{ij})$ $(x_{ij} / y_i)$, and the nontechnical description in the preceding sentence is usually a close approximation to elasticity as

measured by this formula.) A logarithmic independent variable also implies that a fixed percent change in the untransformed independent variable—for example, increasing it by 100 percent (that is, doubling it)—produces a uniform percent change in the untransformed dependent variable. In contrast, with a logarithmic dependent variable, using a continuous independent variable in nontransformed form in the estimating equation implies that it is a unit increase in the nontransformed independent variable (not a percent increase) that produces a uniform percent change in the nontransformed dependent variable.

Manipulation of equation (14) will show that these relationships hold. [Independent variables used in logarithmic form in the estimating equation (equation (15)) are the first $s$ of the $x_{ij}$, and these are found in the product term of equation (14). Independent variables used without transformation in the estimating equation (equation (15)) are the remaining $x_{ij}$, and these are found in the exponential term of equation (14).]

In light of the different relationships that hold true for logarithmic and untransformed continuous independent variables in the estimating equation (equation (15)), a choice was made between these two forms for continuous independent variables. The choice was based on beliefs about the nature of the underlying relationships. For example, it was believed that if an increase from 10 to 20 annualized family illness days spent in bed ("bed days") resulted in, say, a 20 percent increase in total charges, then a further increase from 20 to about 40 bed days would be required to produce another 20 percent increase in total charges. Hence bed days was used in logarithmic form in the estimating equation (equation (15)). Using bed days untransformed would imply that an increase in bed days from 20 to about 30 would produce the second 20 percent increase. In contrast, to take a second example, it was believed that if an increase in the age of the family head from 30 to 40 years produced a given percent increase in total charges (say, increasing it by 30 percent—that is, multiplying it by 1.3), then an increase in age of the family head from 54 to approximately 64 years would produce an equally large percent change (that is, a 30 percent increase) in total charges. Hence, age of the family head was used untransformed. If it were used in logarithmic form, an increase in the head's age from 54 to 72 years would be required to generate the same effect (in percent terms) as the increase from age 30 to 40.

Based on previous research about the underlying relationships, the following four continuous variables were used in logarithmic form in the estimating equation (equation (15)):

- Average family size.
- Annualized family bed days.
- Annualized family work days lost because of illness ("work-loss days").
- Family annual income.

Because family bed days and family work-loss days can take on the value zero, and the logarithm of zero is undefined, the value of these variables was increased by one before performing the logarithmic transformation.

The following three continuous independent variables were used in untransformed form:

- Age of head of family.
- Years of education of head.
- Annualized number of hospital discharges for family members.

Using the PC SAS computer program (SAS Institute, 1985), informal, exploratory tests were conducted on the effect of the choice between no transformation of the above independent variables and a logarithmic transformation of them. There was little difference in $R^2$ and, in general, relatively little difference in tests of statistical significance for each of these independent variables when a logarithmic transformation was substituted for no transformation and vice versa. This would indicate that significant factors can be detected using either the logarithmic or untransformed dependent variable. This can be explained partially by the fact that strong logarithmic effects will also have roughly linear patterns. It should also be noted that the results of significance tests will generally be valid in regressions when using large data sets because of the asymptotic normality properties of least squares estimators. That is, even though the residuals may not be normally distributed (one of the key assumptions on which the regression $F$-tests are based), the $F$-tests used in regression will be valid for large data sets. (See Arnold, 1981.) Thus the results of the significance tests may be similar even if the logarithmic transformation improves the normality of the regression residuals.

## Interpreting the Regression Coefficients

Because the estimating equation (equation (15)) involves variables in logarithmic form, interpretation of the regression coefficients is somewhat complex. The reader may find the following explanation helpful in interpreting the regression coefficients, which appear in Tables II, III, and IV.

When the dependent variable in a regression model is in logarithmic form in the estimating equation, as is total family charges for health care in all the regressions in this report, three different types of independent variables can be distinguished.

1. First are dummy variables. (Again, "dummy variable" designates a categorical variable that takes on only the values 0 and 1.) The regression coefficient, $b$, for a dummy variable has the following interpretation. The presence of the characteristic indicated by the dummy variable is associated with multiplication of the underlying, nonlogarithmic value of the dependent variable by approximately antilog($b$),

where antilog($b$) is the number whose logarithm is $b$. For example, in Table IV, the regression coefficient of D9, a dummy variable denoting families with a head of black race, is $-0.366$. The antilog of $-0.366$ means that black families have, other things equal, total charges about 0.69 times as large as white families or, equivalently, total charges about 31 percent less than those of comparable white families. Table E, which interprets Table IV, presents these findings.

2. Second are continuous independent variables used in logarithmic form in the estimating equation. For such variables, the regression coefficient, $b$, is the elasticity. This means that each 1 percent increase in the underlying, nonlogarithmic independent variable is associated with approximately a $b$ percent increase in the underlying, nonlogarithmic dependent variable. For example, I30, the natural logarithm of a family's annual income, is a continuous independent variable used in logarithmic form in Table IV. Its regression coefficient there is 0.264. This means that each 1 percent increase in annual family income (the underlying, nonlogarithmic independent variable) is, other things equal, associated with approximately a 0.26 percent increase in total family charges (the underlying, nonlogarithmic dependent variable). Again, Table E, which interprets Table IV, presents this finding.

3. Finally, there are continuous independent variables used in nontransformed (that is, nonlogarithmic) form. If the regression coefficient for such a variable is $b$, then each increase of one unit in the independent variable is associated with a multiplication of the underlying, nonlogarithmic form of the dependent variable by approximately antilog($b$). Again, Tables IV and E can serve to illustrate this point. The age of the family head in years, D6, is a continuous independent variable used in nontransformed form in the regression equation whose results are presented in Table IV. Its regression coefficient in that equation is found to be 0.008. The antilog of 0.008 is 1.008. Hence, the interpretation of the regression result is that each increase of one year (the unit in which age is measured) in the age of a family's head is, other things equal, associated with a multiplication of total family charges by approximately 1.008, which is an increase of 0.8 percent. Again, Table E, which interprets the results reported in Table IV, presents this finding.

These interpretations of regression coefficients can be demonstrated by suitable manipulation of equation (14) (or of equation (14a), which is mathematically equivalent). However, in using these interpretations of regression coefficients, it should be noted that antilog($\hat{b}$), the antilog of the estimated regression coefficient, is a biased estimator of the antilog of the regression coefficient, although statistical significance tests associated with the

39

coefficients are sound. Thus, the regressions correctly indicate which variables are statistically significant. If extensive estimation using antilogs is to be carried out, corrections for the bias are available.

## Interpreting Means of Variables

Because of the mixture of dummy variables, untransformed variables, and logarithmic variables in the regressions, readers may find helpful the following information on interpreting the means of variables. (Means of the variables used in the regressions are shown in Tables II–IV.)

The mean value of a dummy variable is the proportion of the population that has the characteristic denoted by the variable. For example, Table IV shows that the mean of D9, the dummy variable denoting black families, is 0.064 for the U.S. population of younger, better off multiple-person families included in this table. This means that about 6 percent of the families were black in 1980.

The mean value of an untransformed continuous variable is simply the (familiar) arithmetic mean of the variable for the population in question. For example, the mean of D6 in Table IV is 42. This variable measures the age of the family head in years and thus shows that the (arithmetic) mean age of the family head for the U.S. population of younger, better off multiple-person families included in the table was 42 in 1980.

The mean value of a logarithmic continuous variable is the logarithm of the geometric mean of the variable. Taking the antilogarithm of the mean does not give the arithmetic mean of the untransformed variable. The geometric mean often differs very substantially from the arithmetic mean and should not be confused with it.

## Analytic Procedures Used

### Introduction

This section describes the analytic procedures used in the regression analyses in this report. Several steps were involved. These were weighting and standardizing the data; selecting the initial variable set; finding a core set of variables through stepwise regression; choosing criteria for evaluating the statistical significance of variables; and estimating the statistical significance of the core variables with SURREGR (Holt and Shah, 1982), a computer program that takes account of the complex sample design of the NMCUES. These steps are described in turn in this section.

### Weighting and Standardizing the Data

Before regression analysis (or other analysis) of the data could begin, certain weighting and standardizing procedures had to be carried out. These procedures are

described in more detail in Appendix II, but a summary of them is included here in order to present in sequence the procedures followed in the regression analysis.

Weighting of each case (family) in the data set began with a weight that previous reports on NMCUES family data have called FWEIGHT. Described simply, FWEIGHT is the reciprocal of the sampling probability adjusted for undercoverage and nonresponse and smoothed to agree with population totals from the March 1980 Current Population Survey. For each case (family), FWEIGHT was multiplied by the proportion of the survey year (calendar year 1980) that the family was eligible for the survey. This time-adjusted weight, called AWEIGHT, is the weight used in the regression analyses and in other analyses in this report. AWEIGHT differs from FWEIGHT only for families not in the sample for a full year.

For these families, standardization of data on income, health spending, health care use, and other variables that measure rates was carried out. Data on these variables covering the period the family was in the sample were divided by the proportion of the year the family was in the sample in order to derive an annualized rate. For example, a family in the sample for half the year with $10,000 of income and $150 of total charges recorded during this half year had its annualized income recorded as $20,000 and its annualized total charges recorded as $300. Annualized statistics, like these, were used in the regression analyses as the measure of all variables involving rates.

### Selecting the Initial Variable Set

Regressions were run separately for three categories of multiple-person families:

- Older families, those with one or more members age 65 or older.

- Younger, lower income families, those with no member 65 or older and with income below 200 percent of the poverty level.

- Younger, better off families, those with no member 65 or older and with income of 200 percent of the poverty level or more.

For each family category, a small number of the 47 independent variables shown in Table I were omitted from the initial regressions because they were not applicable or relevant. Thus, the initial regressions involved total family charges as the dependent variable and slightly fewer than 47 independent variables. The omitted independent variables and the reasons for their omission are as follows. First, one dummy variable for source of health care coverage was omitted for each of the three populations, for the reason described above in the section "Functional Form of the Regression Equation." The omitted variable was that representing the most common source of coverage—Medicare plus private insurance, I38, for older families; and private insurance

only, I34, for both younger family populations. The dummy variable identifying families with all members age 65 or older, D7, was omitted from the regressions for both younger family categories because such families, by definition, have no members age 65 or older. The variable measuring work days lost due to illness, H29, was omitted from the regression for older families because many such families have no working members, which makes this variable meaningless for them. The dummy variable for families with health care coverage entirely from "other public" sources, I40, was omitted from the regression for older families because none of the older families in the sample had such coverage. The dummy variable for families with all members having an unknown perceived health status rating, H25, was omitted from the regressions for older families and for younger, lower income families because no families in the sample in these two categories had this rating. Thus there were 43 independent variables in the initial regressions for older families, 44 independent variables in the initial regressions for younger, lower income families, and 45 independent variables in the initial regressions for younger, better off families.

## Identifying a Core Set of Variables Through Stepwise Regression

For each of the three family categories, stepwise regression was used to select a preferred subset of independent variables from among the original 43 to 45 independent variables. The stepwise regression was carried out using PC SAS (SAS Institute, 1985).

Stepwise regression was used for two reasons. For one, a number of factors were operationalized by multiple variables. For example, family health status was operationalized by four sets of variables: (1) family bed days due to illness, (2) family work-loss days due to illness, (3) the excellent-good-fair-poor scale of reported health status, and (4) the limitations in main activity scale. Because of multicollinearity, the use of multiple variables that operationalize the same concept in a regression equation often yields distorted regression coefficients and large standard errors indicating that none of the variables is significant. Stepwise regression generally selects out a subset of the variables operationalizing a given concept, thus avoiding severe multicollinearity problems.

Second, as more variables are entered into a regression equation, there is a reduction in the precision with which the effects of any one variable can be identified. Standard errors increase, which tends to reduce the number of variables identified as significant. Stepwise regression permits a tradeoff between the additional explanatory power obtained by adding more variables to a regression and the loss of precision in identifying the effects of any one of them.

The preferred independent variable set was defined by the step of the stepwise regression that had the lowest

value of $C(p)$. (For $C(p)$, see Mallows, 1973.) The variable entered in the step at which $C(p)$ reached a minimum and all variables entered in preceding steps were included in the preferred variable set. All other variables were excluded. However, if $C(p)$ was still decreasing when all independent variables that entered the stepwise regression with probabilities less than 0.20 had been added, then the last step before the entry probability for a variable exceeded 0.20 was chosen as defining the preferred variable set. This step was used to define the preferred variable set in the same fashion that $C(p)$ was otherwise used to define it. Finally, all preferred variable sets were required to contain the variables I30 (natural logarithm of annual family income), H13 (indicating whether or not any family member was hospitalized), and D5 (natural logarithm of average family size). If any of these three variables was missing from the preferred variable set developed by the stepwise regression, the missing variable(s) was added and the variable set that included all three of these variables was considered the preferred variable set. Income and hospitalization were included because the literature has found them very important. Family size was included in order to try to assure that effects of family size were distinguished from effects of family structure variables. (The family structure variables used were whether a family had children, whether it had a head and a spouse or only a head, and whether its composition was stable.)

The PC SAS stepwise regression program weights each case (family) in the regression, but does not take account of the complex sample design of NMCUES. That is, it estimates variances according to equation (11), the formula appropriate for noncomplex samples, while equation (13), which is more involved, is the appropriate formula to use in estimating variances of NMCUES data. There appears to be no stepwise regression program available that takes complex sample design into consideration in estimating variances.

## Estimating Statistical Significance Using SURREGR

Because the stepwise regression procedure in PC SAS does not take account of complex sample design in estimating variances, its estimates of statistical significance can involve large errors when it is used in analysis of data from surveys, such as the NMCUES, which have complex sample designs. Therefore, the next steps in the regression analysis procedure involved the use of a regression software program that does estimate variances of complex samples appropriately, using equation (13). The program used was SURREGR (Holt and Shah, 1982), which runs within the SAS system. Nonstepwise SURREGR regressions were run on the preferred variable sets, and it is the results of these regressions that are shown in Tables II–IV.

It should be noted that identical estimates of regression coefficients and of $R^2$ (the proportion of the variance in the dependent variable explained by the independent

variables) are produced by the PC SAS stepwise regression procedure and the SURREGR regression procedure. The differences of concern between the two procedures are in the statistical significance levels that they report for independent variables.

In the regressions shown in Tables II–IV, the probability that the regression coefficient associated with any variable in the regressions was different from zero was computed by SURREGR using the $F$-statistic. This probability is shown in the last column in each table. A regression coefficient was considered significant if its probability of occurring by chance was less than 0.05.

However, because there are 23 to 31 regression coefficients (one for each independent variable) in the preferred variable sets, a simple use of a 0.05 probability test would not be appropriate. Approximately one coefficient meeting a simple 0.05 probability test would be expected for every 20 regression coefficients, and thus approximately one such coefficient would be expected in each of Tables II, III, and IV purely by chance.

The significance test actually used was that the probability associated with any one variable had to be less than $0.05 \div n$, where $n$ is the number of independent variables in the preferred variable set. This test was used in a companion report (Dicker and Sunshine, 1988) and is analogous to the multiple $t$-test used in frequency tables in this report (see Levy and Lemeshow, 1980, p. 296) and in previous reports on NMCUES family data (Dicker and Sunshine, 1987; Sunshine and Dicker, 1987a; Sunshine and Dicker, 1987b). The actual proba-

bility corresponding to $0.05 \div n$ was as follows for the three populations of multiple-person families studied:

| Population | Probability |
|---|---|
| Older families (Table II) . . . . . . . . . . . . . | 0.0020 |
| Younger, lower income families (Table III) . . . | 0.0022 |
| Younger, better off families (Table IV) . . . . . | 0.0016 |

Variables with probabilities below these levels were considered significant and appear in Tables A, C, and E, which report significant findings and accompany the text discussions of findings for each of the three family populations.

### A Further Check Using SURREGR

The preferred variable sets should include all statistically significant independent variables, for stepwise regression selects the independent variables with the greatest statistical significance first, and then moves to progressively less significant variables. However, in light of possible problems arising because the PC SAS stepwise procedure does not compute variances (and hence significance levels) based on the NMCUES complex sample design, a check for omitted significant variables was performed. For each of the three family populations, the full regression model, with all 43 to 45 independent variables, was run using SURREGR. The results are shown in Tables V–VII. These results were generally as expected, and no statistically significant variables were found that were not also statistically significant in the (smaller) preferred variable sets.

### Table I

### Initial set of variables used in the stepwise regression

| Variable type | Variable indicator | Description |
|---|---|---|
| **Dependent variable** | | |
| Total family charges for health care (continuous) | Y1 | Annualized total family charges for health care, transformed into its natural logarithm.[1] |
| **Independent variables** | | |
| Demographic and social | | |
| Head-spouse structure of family (3 categories, 2 dummy variables)[2] | | |
| Head-only family | D1 | 1 = Family had a head only (no spouse) during entire time in survey; 0 = All other head and spouse combinations. |
| Head-spouse change[3] | D2 | 1 = Family had an unstable head-spouse structure during time in survey; 0 = All families with stable head only or stable head-and-spouse. |
| Dynamic-static nature of family (3 categories, 2 dummy variables)[4] | | |
| Head-spouse change[3] | D2 | 1 = Family had an unstable head-spouse structure during time in survey; 0 = All families with stable head only or stable head-and-spouse. |
| Other change | D3 | 1 = Family had a stable head-spouse structure during time in survey but other family member changed, or family did not exist full year; 0 = Other family change status. |

See footnotes at end of table.

42

| Variable type | Variable indicator | Description |
|---|---|---|
| Presence of children in family (2 categories, 1 dummy variable) | D4 | 1 = Family had a member 16 years of age or younger; 0 = All family members 17 years of age or older. |
| Family size (continuous) | D5 | Average family size (in persons) during time in survey, transformed into its natural logarithm. |
| Age of head of the family (continuous) | D6 | Age of the head of the family in years, as of January 1, 1980. |
| Age of family members (2 categories, 1 dummy variable)[5] | D7 | 1 = All family members are 65 years of age or over; 0 = Some or all family members are less than 65 years of age. |
| Sex of the head of the family (2 categories, 1 dummy variable) | D8 | 1 = Female head of family; 0 = Male head of family. |
| Race of head of the family (3 categories, 2 dummy variables)[6] | | |
|   Black | D9 | 1 = Black head of family; 0 = Head of family of other race. |
|   Other | D10 | 1 = Other (neither black nor white) head of family; 0 = Head of family either black or white. |
| Ethnicity of the head of the family (2 categories, 1 dummy variable) | D11 | 1 = Hispanic head of family; 0 = Head of family of other ethnicity. |
| Education of head of the family (continuous) | D12 | Formal education of the head of the family in years of education (18 was the highest value used). |
| Health related | | |
| Hospitalization of a family member (2 categories, 1 dummy variable) | H13 | 1 = Family had one or more members discharged from a hospital during its time in the survey; 0 = No family members discharged from a hospital. |
| Total number of hospital discharges (continuous) | H14 | Annual rate of hospital discharges for all family members. |
| Institutionalization of a family member (2 categories, 1 dummy variable) | H15 | 1 = Family had one or more members institutionalized during its time in the survey or, if it did not continue until the end of 1980, at its termination; 0 = No family members were institutionalized. |
| Death of a family member (2 categories, 1 dummy variable) | H16 | 1 = Family had one or more members die during its time in the survey or, if it did not continue until the end of 1980, at its termination; 0 = No family member died. |
| Birth of a family member (2 categories, 1 dummy variable) | H17 | 1 = Family had one or more members who gave birth to a child during its time in the survey; 0 = No family member gave birth to a child. |
| Illness in a family member (4 dummy variables)[7] | | |
|   Cancer and other neoplasms[7] | H18 | 1 = Family had one or more members with some type of neoplasm during its time in the survey; 0 = No family member had a neoplasm. |
|   Circulatory and heart disease[7] | H19 | 1 = Family had one or more members with some type of circulatory or heart disease during its time in the survey; 0 = No family member had circulatory or heart disease. |
|   Accidents, injuries, and poisonings[7] | H20 | 1 = Family had one or more members with some type of accident, injury, or poisoning during its time in the survey; 0 = No family member had an accident, injury, or poisoning. |
|   Other illnesses only[7] | H21 | 1 = Family members had none of the above illnesses, but one or more members had some other illness during his or her time in the survey; 0 = One or more family members had one of the above illnesses or all family members had no illness. |

See footnotes at end of table.

| Variable type | Variable indicator | Description |
|---|---|---|
| Perceived health status rating of family (5 categories, 4 dummy variables)[8] | | |
| Good | H22 | 1 = Worst perceived health status of any family member was reported as "good"; 0 = All family members were reported in excellent health or some family members were reported in fair or poor health. |
| Fair | H23 | 1 = Worst perceived health status of any family member was reported to be "fair"; 0 = All family members were reported in excellent or good health or some member was reported in poor health. |
| Poor | H24 | 1 = Worst perceived health status of any family member was reported to be "poor"; 0 = No family member had a "poor" rating. |
| Unknown | H25 | 1 = Reported health status of all family members is "unknown"; 0 = Reported health status of at least some family members is known. |
| Limitation in usual activity rating of family (3 categories, 2 dummy variables)[9] | | |
| Secondary limitation | H26 | 1 = Most severe limitation reported for any family member was either a limitation in secondary activity or a limitation in amount or kind of main activity (work, housekeeping, school, and so on); 0 = No family member was reported to have a limitation or a major limitation was reported for one or more family members. |
| Major limitation | H27 | 1 = Most severe limitation reported for any family member was inability to perform a usual major activity (work, housekeeping, school, and so on); 0 = No family member was reported as unable to perform his or her usual major activity. |
| Family illness days in bed (continuous) | H28 | Annual rate of total illness days spent in bed for all family members. One day was added to the annual rate and the resulting statistic then transformed into its natural logarithm. |
| Family work-loss days (continuous)[10] | H29 | Annual rate of total work-loss days due to illness. One day was added to the annual rate and the resulting statistic then transformed into its natural logarithm. |
| Income and insurance | | |
| Family income (continuous) | I30 | Annualized family income in dollars transformed into its natural logarithm. |
| Completeness of health care coverage (4 categories, 3 dummy variables)[11] | | |
| Partial coverage 1 | I31 | 1 = All family members covered but some or all only part year; 0 = All other types of coverage. |
| Partial coverage 2 | I32 | 1 = Some family members had coverage, but some had no coverage; 0 = All other types of coverage. |
| No coverage | I33 | 1 = No family member covered; 0 = Partial or full coverage. |
| Source of health care coverage (8 categories, 7 dummy variables used for each population group)[12] | | |
| Private insurance[12] | I34 | 1 = Family members only had coverage from private health insurance; 0 = All other sources of coverage. |
| Medicaid | I35 | 1 = Family members only had coverage from Medicaid; 0 = All other sources of coverage. |
| Medicare | I36 | 1 = Family members only had coverage from Medicare; 0 = All other sources of coverage. |
| Medicare and other public | I37 | 1 = Family members only had coverage from Medicare and other public programs; 0 = All other sources of coverage. |
| Medicare and private[12] | I38 | 1 = Family members only had coverage from Medicare and private health insurance; 0 = All other sources of coverage. |
| Other public and private | I39 | 1 = Family members had coverage from both (1) public sources other than Medicare and (2) private insurance; 0 = All other sources of coverage. |

See footnotes at end of table.

| Variable type | Variable indicator | Description |
|---|---|---|
| Other public | I40 | 1 = Family members only had coverage from public source(s) other than those listed above; 0 = All other sources of coverage. |
| Unknown | I41 | 1 = Every family member either had no coverage or had coverage from sources not identified; 0 = One or more family members had coverage from identified sources. |
| Geographic | | |
| Region of United States (4 categories, 3 dummy variables)[13] | | |
| North Central | G42 | 1 = Head of family resided in North Central census region; 0 = Head resided in other region of U.S. |
| South | G43 | 1 = Head of family resided in South census region; 0 = Head resided in other region of U.S. |
| West | G44 | 1 = Head of family resided in West census region; 0 = Head resided in other region of U.S. |
| Urban-rural location (4 categories, 3 dummy variables)[14] | | |
| Metropolitan suburb | G45 | 1 = Head of family resided in a suburb of a metropolitan statistical area; 0 = Head resided in another location. |
| Nonmetropolitan urban area | G46 | 1 = Head of family resided in an urban area that was not a part of a metropolitan statistical area; 0 = Head resided in another location. |
| Nonurban area | G47 | 1 = Head of family resided in non-urban area; 0 = Head resided in another location. |

[1]For the types of health care charges included in total family charges for health care, see the Introduction section in the text.
[2]Omitted category is families with both a head and a spouse during time in survey.
[3]The variable is entered only once in the regression, but functions both as a measure of head-spouse structure and as a measure of dynamic-static nature of the family.
[4]Omitted category is static families—that is, families that had no change in membership and were in the survey the full survey year.
[5]This variable was only used in regressions for older families (families with a member 65 years of age or over). When used with this population, the "0" category designates families with members both over and under 65 years of age.
[6]Omitted category is families with a white head of family.
[7]Omitted category is families in which no family member reported an illness. The dummy variables for cancer; heart and circulatory disease; and accidents, illnesses, and poisonings are not mutually exclusive.
[8]Omitted category is all family members reported to be in excellent health.
[9]Omitted category is families in which no family member reported a limitation. Family members for whom limitation status was unknown were coded as having no limitation.
[10]This variable was not used in regressions for older families (families with a member 65 years of age or over).
[11]Omitted category is families in which all family members had full-year coverage by private health insurance and/or public health care coverage program(s).
[12]For regressions for older families, the omitted category is coverage by both Medicare and private insurance (I38). For regressions for younger families, the omitted category is coverage from private health insurance only (I34).
[13]Omitted category is residence in the Northeast census region.
[14]Omitted category is residence in the central city of a metropolitan statistical area.

NOTE: Further information on the variables in this table may be found either in Appendix III, "Definition of Terms," or in the text.

**Table II**

**SURREGR regression for total family charges for health care for older multiple-person families: preferred model**

| Independent variable | Mean | Regression coefficient | Standard error of regression coefficient | F-value | Probability |
|---|---|---|---|---|---|
| D1  Head only | 0.254 | −0.122 | 0.083 | 2.18 | 0.1447 |
| D2  Head-spouse change | 0.038 | 0.291 | 0.148 | 3.84 | 0.0542 |
| D3  Other change | 0.106 | 0.253 | 0.082 | 9.51 | 0.0030 |
| D4  Children | 0.110 | 0.143 | 0.130 | 1.21 | 0.2745 |
| D5  Family size | 0.866 | 0.027 | 0.135 | 0.04 | 0.8436 |
| D9  Black race | 0.090 | −0.356 | 0.172 | 4.28 | 0.0424 |
| D12 Education of head | 9.850 | 0.017 | 0.008 | 4.45 | 0.0385 |
| H13 Hospitalization | 0.398 | 1.179 | 0.088 | 180.83 | 0.0000 |
| H14 Discharges | 0.727 | 0.244 | 0.045 | 29.79 | 0.0000 |
| H18 Cancer | 0.167 | 0.377 | 0.065 | 33.45 | 0.0000 |
| H19 Heart disease, etc. | 0.730 | 0.480 | 0.060 | 64.87 | 0.0000 |
| H20 Accidents, etc. | 0.281 | 0.268 | 0.055 | 24.11 | 0.0000 |
| H22 Good health | 0.337 | 0.220 | 0.125 | 3.12 | 0.0819 |
| H23 Fair health | 0.285 | 0.317 | 0.121 | 6.84 | 0.0110 |
| H24 Poor health | 0.245 | 0.322 | 0.149 | 4.65 | 0.0345 |
| H28 Bed days | 1.911 | 0.190 | 0.025 | 59.00 | 0.0000 |
| I30 Income | 9.588 | 0.148 | 0.045 | 10.53 | 0.0018 |
| I31 Partial coverage 1 | 0.061 | −0.182 | 0.112 | 2.66 | 0.1077 |
| I32 Partial coverage 2 | 0.105 | −0.308 | 0.116 | 7.12 | 0.0095 |
| I36 Medicare coverage only | 0.090 | −0.170 | 0.105 | 2.64 | 0.1089 |
| I39 Other public and private coverage | 0.012 | −0.567 | 0.286 | 3.94 | 0.0512 |
| G43 South Region | 0.348 | 0.099 | 0.070 | 2.01 | 0.1610 |
| G44 West Region | 0.210 | 0.144 | 0.088 | 2.70 | 0.1047 |
| G46 Nonmetropolitan urban area | 0.147 | −0.232 | 0.097 | 5.75 | 0.0193 |
| G47 Nonurban area | 0.191 | −0.196 | 0.089 | 4.89 | 0.0304 |

Intercept = 3.810
Number of observations = 839
Multiple correlation coefficient squared ($R^2$) = 0.720
$F$ = 167.77 with 25 degrees of freedom

Mean of dependent variable = 7.072
Probability = 0.0000
68 denominator degrees of freedom

NOTES: Older families are families with member(s) 65 years of age or over. A probability of 0.0020 was needed for statistical significance at the 0.05 level using an $F$-test analogous to a multiple $t$-test (see Appendix I).

**Table III**

**SURREGR regression for total family charges for health care for younger, lower income multiple-person families: preferred model**

| Independent variable | Mean | Regression coefficient | Standard error of regression coefficient | F-value | Probability |
|---|---|---|---|---|---|
| D2 Head-spouse change . . . . . . . . . . . . . . | 0.041 | 0.355 | 0.194 | 3.34 | 0.0721 |
| D5 Family size . . . . . . . . . . . . . . . . . . . | 1.252 | 0.151 | 0.081 | 3.44 | 0.0680 |
| D9 Black race . . . . . . . . . . . . . . . . . . . | 0.214 | −0.130 | 0.078 | 2.78 | 0.0998 |
| D12 Education of head . . . . . . . . . . . . . . | 10.762 | 0.018 | 0.011 | 2.74 | 0.1022 |
| H13 Hospitalization . . . . . . . . . . . . . . . . | 0.337 | 1.099 | 0.100 | 120.96 | 0.0000 |
| H14 Discharges . . . . . . . . . . . . . . . . . . | 0.604 | 0.280 | 0.041 | 47.04 | 0.0000 |
| H17 Birth . . . . . . . . . . . . . . . . . . . . . . | 0.063 | −0.256 | 0.088 | 8.42 | 0.0050 |
| H18 Cancer . . . . . . . . . . . . . . . . . . . . | 0.076 | 0.377 | 0.109 | 12.07 | 0.0009 |
| H19 Heart disease, etc. . . . . . . . . . . . . . | 0.248 | 0.152 | 0.054 | 7.98 | 0.0062 |
| H20 Accidents, etc. . . . . . . . . . . . . . . . | 0.470 | 0.171 | 0.058 | 8.76 | 0.0042 |
| H23 Fair health . . . . . . . . . . . . . . . . . . | 0.262 | 0.225 | 0.080 | 7.91 | 0.0064 |
| H24 Poor health . . . . . . . . . . . . . . . . . | 0.134 | 0.334 | 0.104 | 10.30 | 0.0020 |
| H27 Major limitation . . . . . . . . . . . . . . . | 0.151 | 0.262 | 0.070 | 14.18 | 0.0003 |
| H28 Bed days . . . . . . . . . . . . . . . . . . . | 2.282 | 0.126 | 0.025 | 26.47 | 0.0000 |
| H29 Work-loss days . . . . . . . . . . . . . . . | 1.100 | 0.061 | 0.023 | 7.05 | 0.0099 |
| I30 Income . . . . . . . . . . . . . . . . . . . . | 9.049 | 0.042 | 0.058 | 0.52 | 0.4720 |
| I31 Partial coverage 1 . . . . . . . . . . . . . . | 0.237 | −0.176 | 0.074 | 5.66 | 0.0201 |
| I32 Partial coverage 2 . . . . . . . . . . . . . . | 0.139 | −0.183 | 0.098 | 3.51 | 0.0651 |
| I33 No coverage . . . . . . . . . . . . . . . . . | 0.085 | −0.565 | 0.123 | 21.14 | 0.0000 |
| I38 Medicare and private coverage . . . . . . . . | 0.003 | −1.101 | 0.786 | 1.96 | 0.1656 |
| I40 Other public coverage only . . . . . . . . . . | 0.007 | −0.665 | 0.773 | 0.74 | 0.3940 |
| I41 Coverage source unknown . . . . . . . . . . | 0.269 | −0.189 | 0.073 | 6.81 | 0.0111 |
| G43 South Region . . . . . . . . . . . . . . . . . | 0.327 | −0.241 | 0.055 | 18.93 | 0.0000 |

Intercept = 5.023
Number of observations = 1,012
Multiple correlation coefficient squared ($R^2$) = 0.595
$F$ = 87.68 with 23 degrees of freedom

Mean of dependent variable = 6.684
Probability = 0.0000
69 denominator degrees of freedom

NOTES: Younger, lower income families are families with no member 65 years of age or over and with incomes below 200 percent of the poverty level. A probability of 0.0022 was needed for statistical significance at the 0.05 level using an $F$-test analogous to a multiple $t$-test (see Appendix I).

**Table IV**

**SURREGR regression for total family charges for health care for younger, better off multiple-person families: preferred model**

| Independent variable | Mean | Regression coefficient | Standard error of regression coefficient | F-value | Probability |
|---|---|---|---|---|---|
| D1 Head only | 0.143 | −0.210 | 0.081 | 6.70 | 0.0117 |
| D4 Children | 0.594 | 0.199 | 0.057 | 12.33 | 0.0008 |
| D5 Family size | 1.136 | 0.120 | 0.066 | 3.30 | 0.0737 |
| D6 Age of head | 42.045 | 0.008 | 0.002 | 27.60 | 0.0000 |
| D8 Female family head | 0.132 | 0.133 | 0.079 | 2.79 | 0.0992 |
| D9 Black race | 0.064 | −0.366 | 0.083 | 19.52 | 0.0000 |
| D10 "Other" race | 0.153 | −0.208 | 0.140 | 2.20 | 0.1429 |
| D12 Education of head | 12.656 | 0.026 | 0.007 | 15.55 | 0.0002 |
| H13 Hospitalization | 0.268 | 1.018 | 0.055 | 338.77 | 0.0000 |
| H14 Discharges | 0.423 | 0.272 | 0.025 | 119.90 | 0.0000 |
| H15 Institutionalization | 0.001 | −0.698 | 0.405 | 2.97 | 0.0893 |
| H18 Cancer | 0.082 | 0.157 | 0.055 | 8.04 | 0.0060 |
| H19 Heart disease, etc. | 0.267 | 0.179 | 0.036 | 24.96 | 0.0000 |
| H20 Accidents, etc. | 0.433 | 0.227 | 0.031 | 52.21 | 0.0000 |
| H22 Good health | 0.467 | 0.094 | 0.035 | 7.23 | 0.0090 |
| H23 Fair health | 0.136 | 0.224 | 0.041 | 29.53 | 0.0000 |
| H24 Poor health | 0.054 | 0.419 | 0.080 | 27.55 | 0.0000 |
| H25 Unknown health status | 0.000 | 1.512 | 0.111 | 186.40 | 0.0000 |
| H27 Major limitation | 0.079 | 0.116 | 0.072 | 2.59 | 0.1120 |
| H28 Bed days | 1.959 | 0.112 | 0.018 | 37.06 | 0.0000 |
| H29 Work-loss days | 1.389 | 0.072 | 0.015 | 23.40 | 0.0000 |
| I30 Income | 10.261 | 0.264 | 0.038 | 49.12 | 0.0000 |
| I31 Partial coverage 1 | 0.141 | −0.123 | 0.057 | 4.56 | 0.0362 |
| I32 Partial coverage 2 | 0.057 | −0.285 | 0.069 | 17.08 | 0.0001 |
| I33 No coverage | 0.020 | −0.294 | 0.128 | 5.28 | 0.0247 |
| I36 Medicare coverage only | 0.003 | −0.786 | 0.530 | 2.20 | 0.1426 |
| I40 Other public coverage only | 0.004 | −0.403 | 0.301 | 1.79 | 0.1855 |
| I41 Coverage source unknown | 0.075 | −0.184 | 0.088 | 4.40 | 0.0396 |
| G44 West Region | 0.192 | 0.135 | 0.638 | 12.66 | 0.0007 |
| G45 Metropolitan suburb | 0.446 | 0.092 | 0.031 | 8.77 | 0.0042 |
| G47 Nonurban area | 0.163 | −0.079 | 0.036 | 4.87 | 0.0306 |

Intercept = 2.209
Number of observations = 2,882
Multiple correlation coefficient squared ($R^2$) = 0.569
$F$ = 301.90 with 31 degrees of freedom

Mean of dependent variable = 6.794
Probability = 0.0000
69 denominator degrees of freedom

NOTES: Younger, better off families are families with no member 65 years of age or over and incomes 200 percent of the poverty level or higher. A probability of 0.0016 was needed for statistical significance at the 0.05 level using an $F$-test analogous to a multiple $t$-test (see Appendix I).

**Table V**

**SURREGR regression for total family charges for health care for older multiple-person families: full model**

| Independent variable | Mean | Regression coefficient | Standard error of regression coefficient | F-value | Probability |
|---|---|---|---|---|---|
| D1 Head only | 0.254 | −0.127 | 0.100 | 1.62 | 0.2073 |
| D2 Head-spouse change | 0.038 | 0.271 | 0.206 | 1.73 | 0.1926 |
| D3 Other change | 0.106 | 0.244 | 0.102 | 5.67 | 0.0200 |
| D4 Children | 0.110 | 0.150 | 0.130 | 1.33 | 0.2534 |
| D5 Family size | 0.866 | 0.077 | 0.138 | 0.31 | 0.5791 |
| D6 Age of head | 68.022 | 0.003 | 0.003 | 0.98 | 0.3250 |
| D7 All members 65+ | 0.386 | −0.023 | 0.080 | 0.08 | 0.7769 |
| D8 Female family head | 0.219 | 0.017 | 0.123 | 0.02 | 0.8819 |
| D9 Black race | 0.090 | −0.336 | 0.175 | 3.68 | 0.0594 |
| D10 Other nonwhite race | 0.020 | −0.044 | 0.179 | 0.06 | 0.8026 |
| D11 Hispanic | 0.035 | 0.175 | 0.122 | 2.08 | 0.1539 |
| D12 Education of head | 9.850 | 0.020 | 0.008 | 5.68 | 0.0200 |
| H13 Hospitalization | 0.398 | 1.166 | 0.090 | 167.97 | 0.0000 |
| H14 Discharges | 0.727 | 0.249 | 0.046 | 29.55 | 0.0000 |
| H15 Institutionalization | 0.019 | 0.156 | 0.253 | 0.38 | 0.5388 |
| H16 Death | 0.049 | 0.002 | † | 0.00 | 0.9895 |
| H17 Birth | 0.002 | −0.120 | 0.220 | 0.30 | 0.5853 |
| H18 Cancer | 0.167 | 0.382 | 0.065 | 35.08 | 0.0000 |
| H19 Heart disease, etc. | 0.730 | 0.444 | 0.087 | 25.86 | 0.0000 |
| H20 Accidents, etc. | 0.281 | 0.266 | 0.060 | 19.70 | 0.0000 |
| H21 Other illnesses only | 0.170 | −0.047 | 0.117 | 0.16 | 0.6949 |
| H22 Good health | 0.337 | 0.233 | 0.133 | 3.08 | 0.0839 |
| H23 Fair health | 0.285 | 0.351 | 0.142 | 6.13 | 0.0158 |
| H24 Poor health | 0.245 | 0.366 | 0.177 | 4.29 | 0.0422 |
| H26 Secondary limitation | 0.081 | 0.035 | 0.097 | 0.13 | 0.7246 |
| H27 Major limitation | 0.538 | −0.068 | 0.085 | 0.64 | 0.4252 |
| H28 Bed days | 1.911 | 0.192 | 0.025 | 57.97 | 0.0000 |
| I30 Income | 9.588 | 0.131 | 0.047 | 7.81 | 0.0068 |
| I31 Partial coverage 1 | 0.061 | −0.195 | 0.115 | 2.89 | 0.0936 |
| I32 Partial coverage 2 | 0.105 | −0.335 | 0.109 | 9.40 | 0.0031 |
| I33 No coverage | 0.007 | −0.367 | 0.387 | 0.90 | 0.3449 |
| I34 Private coverage only | 0.030 | 0.023 | 0.160 | 0.02 | 0.8885 |
| I35 Medicaid coverage only | 0.004 | 0.011 | † | 0.00 | 0.9475 |
| I36 Medicare coverage only | 0.090 | −0.188 | 0.108 | 3.03 | 0.0862 |
| I37 Medicare and other public coverage | 0.061 | −0.046 | 0.115 | 0.16 | 0.6924 |
| I39 Other public and private coverage | 0.012 | −0.584 | 0.300 | 3.78 | 0.0561 |
| I41 Coverage source unknown | 0.031 | −0.118 | 0.235 | 0.25 | 0.6164 |
| G42 North Central Region | 0.234 | 0.075 | 0.096 | 0.61 | 0.4379 |
| G43 South Region | 0.348 | 0.146 | 0.084 | 3.02 | 0.0866 |
| G44 West Region | 0.210 | 0.173 | 0.106 | 2.70 | 0.1048 |
| G45 Metropolitan suburb | 0.362 | 0.051 | 0.069 | 0.54 | 0.4631 |
| G46 Nonmetropolitan urban area | 0.147 | −0.190 | 0.109 | 3.06 | 0.0847 |
| G47 Nonurban area | 0.191 | −0.161 | 0.109 | 2.20 | 0.1422 |

Intercept = 3.642
Number of observations – 839
Multiple correlation coefficient squared $(R^2)$ = 0.723
$F$ = 185.52 with 43 degrees of freedom

Mean of dependent variable = 7.072
Probability = 0.0000
68 denominator degrees of freedom

† Standard error of regression not stated because $F$-statistic had a value of zero after rounding.

## Table VI
## SURREGR regression for total family charges for health care for younger, lower income multiple-person families: full model

| Independent variable | Mean | Regression coefficient | Standard error of regression coefficient | F-value | Probability |
|---|---|---|---|---|---|
| D1 Head only | 0.458 | −0.171 | 0.164 | 1.09 | 0.2999 |
| D2 Head-spouse change | 0.041 | 0.259 | 0.190 | 1.85 | 0.1786 |
| D3 Other change | 0.199 | 0.013 | 0.092 | 0.02 | 0.8937 |
| D4 Children | 0.820 | 0.125 | 0.103 | 1.48 | 0.2275 |
| D5 Family size | 1.252 | 0.087 | 0.089 | 0.95 | 0.3332 |
| D6 Age of head | 37.484 | 0.005 | 0.003 | 2.19 | 0.1439 |
| D8 Female family head | 0.466 | 0.142 | 0.149 | 0.91 | 0.3429 |
| D9 Black race | 0.214 | −0.105 | 0.090 | 1.37 | 0.2455 |
| D10 Other nonwhite race | 0.021 | −0.044 | 0.254 | 0.03 | 0.8544 |
| D11 Hispanic | 0.116 | 0.050 | 0.107 | 0.22 | 0.6414 |
| D12 Education of head | 10.762 | 0.021 | 0.012 | 3.20 | 0.0780 |
| H13 Hospitalization | 0.337 | 1.082 | 0.102 | 113.60 | 0.0000 |
| H14 Discharges | 0.604 | 0.288 | 0.041 | 48.90 | 0.0000 |
| H15 Institutionalization | 0.006 | −0.162 | 0.318 | 0.26 | 0.6092 |
| H16 Death | 0.009 | −0.091 | 0.194 | 0.22 | 0.6414 |
| H17 Birth | 0.063 | −0.243 | 0.123 | 3.88 | 0.0530 |
| H18 Cancer | 0.076 | 0.408 | 0.126 | 10.44 | 0.0019 |
| H19 Heart disease, etc. | 0.248 | 0.211 | 0.081 | 6.76 | 0.0114 |
| H20 Accidents, etc. | 0.470 | 0.284 | 0.088 | 10.40 | 0.0019 |
| H21 Other illnesses only | 0.361 | 0.159 | 0.106 | 2.25 | 0.1386 |
| H22 Good health | 0.377 | 0.058 | 0.065 | 0.80 | 0.3737 |
| H23 Fair health | 0.262 | 0.249 | 0.091 | 7.43 | 0.0081 |
| H24 Poor health | 0.134 | 0.347 | 0.123 | 7.93 | 0.0063 |
| H26 Secondary limitation | 0.092 | 0.045 | 0.106 | 0.18 | 0.6743 |
| H27 Major limitation | 0.151 | 0.243 | 0.083 | 8.53 | 0.0047 |
| H28 Bed days | 2.282 | 0.126 | 0.024 | 26.59 | 0.0000 |
| H29 Work-loss days | 1.100 | 0.066 | 0.023 | 8.11 | 0.0058 |
| I30 Income | 9.049 | 0.048 | 0.067 | 0.51 | 0.4787 |
| I31 Partial coverage 1 | 0.237 | −0.192 | 0.079 | 5.90 | 0.0177 |
| I32 Partial coverage 2 | 0.139 | −0.232 | 0.098 | 5.56 | 0.0212 |
| I33 No coverage | 0.085 | −0.590 | 0.121 | 23.59 | 0.0000 |
| I35 Medicaid coverage only | 0.172 | 0.090 | 0.111 | 0.66 | 0.4193 |
| I36 Medicare coverage only | 0.013 | 0.415 | 0.214 | 3.77 | 0.0564 |
| I37 Medicare and other public coverage | 0.001 | −0.183 | 0.203 | 0.81 | 0.3699 |
| I38 Medicare and private coverage | 0.003 | −1.071 | 0.796 | 1.81 | 0.1828 |
| I39 Other public and private coverage | 0.226 | 0.087 | 0.069 | 1.57 | 0.2143 |
| I40 Other public coverage only | 0.007 | −0.592 | 0.758 | 0.61 | 0.4390 |
| I41 Coverage source unknown | 0.269 | −0.110 | 0.085 | 1.66 | 0.2025 |
| G42 North Central Region | 0.230 | −0.059 | 0.074 | 0.63 | 0.4313 |
| G43 South Region | 0.327 | −0.244 | 0.064 | 14.53 | 0.0003 |
| G44 West Region | 0.227 | 0.003 | † | 0.00 | 0.9738 |
| G45 Metropolitan suburb | 0.318 | 0.059 | 0.072 | 0.67 | 0.4171 |
| G46 Nonmetropolitan urban area | 0.137 | 0.025 | 0.083 | 0.09 | 0.7621 |
| G47 Nonurban area | 0.182 | 0.052 | 0.102 | 0.26 | 0.6119 |

Intercept = 4.504  
Number of observations = 1,012  
Multiple correlation coefficient squared ($R^2$) = 0.600  
$F$ = 205.29 with 44 degrees of freedom

Mean of dependent variable = 6.684  
Probability = 0.0000  
69 denominator degrees of freedom

† Standard error of regression not stated because F-statistic had a value of zero after rounding.

50

**Table VII**

**SURREGR regression for total family charges for health care for younger, better off multiple-person families: full model**

| Independent variable | Mean | Regression coefficient | Standard error of regression coefficient | F-value | Probability |
|---|---|---|---|---|---|
| D1 Head only . . . . . . . . . . . . . . . . . . . . . . | 0.143 | −0.224 | 0.085 | 6.92 | 0.0105 |
| D2 Head-spouse change . . . . . . . . . . . . . . . | 0.020 | −0.090 | 0.114 | 0.62 | 0.4335 |
| D3 Other change . . . . . . . . . . . . . . . . . . . | 0.168 | −0.048 | 0.056 | 0.74 | 0.3916 |
| D4 Children . . . . . . . . . . . . . . . . . . . . . . . | 0.594 | 0.201 | 0.057 | 12.42 | 0.0008 |
| D5 Family size . . . . . . . . . . . . . . . . . . . . | 1.136 | 0.123 | 0.069 | 3.19 | 0.0784 |
| D6 Age of head . . . . . . . . . . . . . . . . . . . . | 42.045 | 0.008 | 0.002 | 24.35 | 0.0000 |
| D8 Female family head . . . . . . . . . . . . . . . | 0.132 | 0.143 | 0.083 | 2.96 | 0.0899 |
| D9 Black race . . . . . . . . . . . . . . . . . . . . . | 0.064 | −0.379 | 0.082 | 21.30 | 0.0000 |
| D10 Other nonwhite race . . . . . . . . . . . . . . | 0.015 | −0.215 | 0.142 | 2.29 | 0.1348 |
| D11 Hispanic . . . . . . . . . . . . . . . . . . . . . . | 0.048 | −0.087 | 0.079 | 1.20 | 0.2762 |
| D12 Education of head . . . . . . . . . . . . . . . . | 12.657 | 0.025 | 0.007 | 13.48 | 0.0005 |
| H13 Hospitalization . . . . . . . . . . . . . . . . . . | 0.268 | 1.010 | 0.056 | 325.98 | 0.0000 |
| H14 Discharges . . . . . . . . . . . . . . . . . . . . | 0.423 | 0.287 | 0.029 | 97.01 | 0.0000 |
| H15 Institutionalization . . . . . . . . . . . . . . . . | 0.001 | −0.677 | 0.375 | 3.26 | 0.0752 |
| H16 Death . . . . . . . . . . . . . . . . . . . . . . . . | 0.006 | 0.001 | † | 0.00 | 1.0000 |
| H17 Birth . . . . . . . . . . . . . . . . . . . . . . . . . | 0.037 | −0.045 | 0.096 | 0.22 | 0.6374 |
| H18 Cancer . . . . . . . . . . . . . . . . . . . . . . . | 0.082 | 0.151 | 0.056 | 7.30 | 0.0087 |
| H19 Heart disease, etc. . . . . . . . . . . . . . . . | 0.267 | 0.170 | 0.045 | 14.10 | 0.0004 |
| H20 Accidents, etc. . . . . . . . . . . . . . . . . . | 0.433 | 0.219 | 0.047 | 21.80 | 0.0000 |
| H21 Other illnesses only . . . . . . . . . . . . . . . | 0.380 | −0.011 | 0.064 | 0.03 | 0.8567 |
| H22 Good health . . . . . . . . . . . . . . . . . . . . | 0.467 | 0.096 | 0.035 | 7.60 | 0.0075 |
| H23 Fair health . . . . . . . . . . . . . . . . . . . . . | 0.136 | 0.228 | 0.042 | 29.84 | 0.0000 |
| H24 Poor health . . . . . . . . . . . . . . . . . . . . | 0.054 | 0.421 | 0.081 | 27.08 | 0.0000 |
| H25 Unknown health status . . . . . . . . . . . . . | 0.001 | 1.486 | 0.123 | 145.68 | 0.0000 |
| H26 Secondary limitation . . . . . . . . . . . . . . | 0.047 | −0.029 | 0.073 | 0.16 | 0.6909 |
| H27 Major limitation . . . . . . . . . . . . . . . . . | 0.079 | 0.108 | 0.075 | 2.08 | 0.1535 |
| H28 Bed days . . . . . . . . . . . . . . . . . . . . . | 1.959 | 0.111 | 0.018 | 37.31 | 0.0000 |
| H29 Work-loss days . . . . . . . . . . . . . . . . . | 1.389 | 0.071 | 0.015 | 22.13 | 0.0000 |
| I30 Income . . . . . . . . . . . . . . . . . . . . . . | 10.260 | 0.264 | 0.038 | 49.17 | 0.0000 |
| I31 Partial coverage 1 . . . . . . . . . . . . . . . . | 0.141 | −0.111 | 0.058 | 3.61 | 0.0615 |
| I32 Partial coverage 2 . . . . . . . . . . . . . . . . | 0.057 | −0.268 | 0.075 | 12.66 | 0.0007 |
| I33 No coverage . . . . . . . . . . . . . . . . . . | 0.020 | −0.279 | 0.132 | 4.45 | 0.0384 |
| I35 Medicaid coverage only . . . . . . . . . . . . . | 0.005 | 0.058 | 0.183 | 0.10 | 0.7508 |
| I36 Medicare coverage only . . . . . . . . . . . . | 0.003 | −0.781 | 0.543 | 2.07 | 0.1546 |
| I37 Medicare and other public coverage . . . . . . | <0.001 | −0.022 | 0.064 | 0.12 | 0.7304 |
| I38 Medicare and private coverage . . . . . . . . | 0.002 | −0.332 | 0.250 | 1.76 | 0.1893 |
| I39 Other public and private coverage . . . . . . . | 0.135 | 0.052 | 0.051 | 1.02 | 0.3156 |
| I40 Other public coverage only . . . . . . . . . . . | 0.004 | −0.389 | 0.307 | 1.61 | 0.2089 |
| I41 Coverage source unknown . . . . . . . . . . . | 0.075 | −0.184 | 0.089 | 4.28 | 0.0424 |
| G42 North Central Region . . . . . . . . . . . . . . | 0.283 | 0.038 | 0.042 | 0.80 | 0.3750 |
| G43 South Region . . . . . . . . . . . . . . . . . . | 0.298 | 0.000 | † | 0.00 | 1.0000 |
| G44 West Region . . . . . . . . . . . . . . . . . . | 0.192 | 0.150 | 0.043 | 12.14 | 0.0009 |
| G45 Metropolitan suburb . . . . . . . . . . . . . . . | 0.446 | 0.066 | 0.038 | 3.09 | 0.0832 |
| G46 Nonmetropolitan urban area . . . . . . . . . . | 0.129 | −0.067 | 0.052 | 1.65 | 0.2032 |
| G47 Nonurban area . . . . . . . . . . . . . . . . . . | 0.163 | −0.115 | 0.044 | 6.84 | 0.0110 |

Intercept =2.263
Number of observations =2,882
Multiple correlation coefficient squared ($R^2$) =0.570
$F$ =579.95 with 45 degrees of freedom

Mean of dependent variable =6.794
Probability =0.0000
69 denominator degrees of freedom

† Standard error of regression not stated because $F$-statistic had a value of zero after rounding.

51

# Appendix II
# Technical Notes on Survey and Nonregression Methods

## Survey Background

The National Medical Care Utilization and Expenditure Survey (NMCUES) was a panel survey designed to collect data about the U.S. civilian noninstitutionalized population in 1980. During the course of the survey, information was obtained on health, access to and use of medical services, associated charges and sources of payment, and health insurance coverage. Information was collected in such a way that data can be provided at the family level as well as for individuals. The survey contained both a household sample and a Medicaid case sample. This report is based on the household sample. NMCUES was cosponsored by the National Center for Health Statistics and the Health Care Financing Administration. Data collection was provided under contract by the Research Triangle Institute and its subcontractors, National Opinion Research Center and SysteMetrics, Inc.

The basic survey plan for NMCUES drew heavily on two surveys, the National Health Interview Survey (NHIS), conducted annually by the National Center for Health Statistics, and the National Medical Care Expenditure Survey (NMCES), cosponsored by the National Center for Health Services Research and the National Center for Health Statistics.

NHIS is a continuing, multipurpose, cross-sectional survey first conducted in 1957. The main purpose of NHIS is to collect information on illness, disability, and the use of medical care. Although some information on medical expenditures and insurance payments has been collected in NHIS, the cross-sectional nature of the survey design is not well suited for providing annual data on expenditures and payments.

NMCES was a panel survey in which a sample of households was interviewed six times over an 18-month period in 1977 and 1978. NMCES was specifically designed to provide comprehensive data on how health services were used and paid for in the United States in 1977.

NMCUES is similar to NMCES in survey design and questionnaire wording, so analysis of some of the changes during the period 1977–80 is possible. Both NMCUES and NMCES used question wording that was similar to NHIS in areas common to the three surveys. Together, NMCES and NMCUES provide extensive information on illness, disability, use of medical care, costs of medical care, sources of payment for medical care, and health insurance coverage at two points in time.

## Sample Design

The NMCUES sample of housing units and group quarters, hereafter jointly referred to as dwelling units, is a concatenation of two independently selected national samples, one provided by the Research Triangle Institute and the other by the National Opinion Research Center. The sample designs used by these two organizations are similar with respect to principal design features; both can be characterized as stratified, four-stage area probability designs. The principal differences between the two designs are the type of stratification variables and the specific definitions of sampling units at each stage. The salient design features of the two sample surveys are summarized in the following sections.

The target population for NMCUES consisted of all persons who were members of the U.S. civilian noninstitutionalized population at any time from January 1, 1980, through December 31, 1980. All persons living in a sample dwelling unit at the time of the first interview contact became part of the national sample. Unmarried students 17–22 years of age who lived away from home were included in the sample when a parent or guardian was included in the sample. In addition, persons who died or were institutionalized between January 1 and the date of the first interview were included in the sample if they were related to persons living in the sampled dwelling units. All of these persons were considered "key" persons, and data were collected for them for the full 12 months of 1980 or for the proportion of time that they were part of the U.S. civilian noninstitutionalized population. In addition, babies born to key persons were considered key persons, and data were collected for them from the time of birth. Relatives from outside the original population (that is, institutionalized, in the Armed Forces, or outside the United States between January 1 and the first interview) who moved in with key persons after the first interview were also considered key persons, and data were collected for them from the time they joined the key person. Relatives who moved in with key persons after the first

interview but were part of the civilian noninstitutionalized population on January 1, 1980, were classified as "nonkey" persons. Data were collected for nonkey persons for the time that they lived with a key person but, because they had a chance of selection in the initial sample, their data are not used for general person-level analysis. However, data for nonkey persons are used in family analysis because nonkey persons contributed to the family's utilization of and expenditures for health care during the time they were part of the family.

Persons included in the sample were grouped into "reporting units" for data collection purposes. Reporting units were defined as all persons related to each other by blood, marriage, adoption, or foster care status and living in the same dwelling unit. The combined NMCUES sample consisted of 7,244 eligible reporting units, of which 6,599 agreed to participate in the survey. In total, data were obtained on 17,123 key persons. The Research Triangle Institute sample yielded 8,326 key persons, and the National Opinion Research Center sample yielded 8,797.

### Research Triangle Institute Sample Design

A primary sampling unit (PSU) is defined as a county, a group of contiguous counties, or parts of counties with a combined minimum 1970 population size of 20,000. A total of 1,686 disjoint PSU's exhaust the land area of the 50 States and Washington, D.C. The PSU's are classified as one of two types. The 16 largest standard metropolitan statistical areas (SMSA's) are designated as self-representing PSU's, and the remaining 1,670 PSU's in the primary sampling frame are designated as non-self-representing PSU's.

PSU's are grouped into strata whose members tend to be relatively alike within strata and relatively unlike between strata. PSU's derived from the 16 largest SMSA's had sufficient population in 1970 to be treated as primary strata. The 1,659 non-self-representing PSU's from the continental United States were stratified into 59 primary strata with approximately equal populations. Each of these primary strata had a 1970 population of about 3⅓ million. One supplementary primary stratum of 11 PSU's, with a 1970 population of about 1 million, was added to the Research Triangle Institute primary frame to include Alaska and Hawaii.

The total first-stage sample for Research Triangle Institute consisted of 59 PSU's, of which 16 were self-representing PSU's. The non-self-representing PSU's were obtained by selecting one PSU from each of the 43 non-self-representing primary strata. These PSU's were selected with probability proportional to 1970 population size.

In each of the 59 sample PSU's, the entire PSU was divided into smaller disjoint area units called secondary sampling units (SSU's). Each SSU consisted of one or more enumeration districts or block groups defined by the 1970 census. Within each PSU, SSU's were ordered and then partitioned to form secondary strata of approximately equal size. Two secondary strata were formed in the non-self-representing PSU drawn from Alaska and Hawaii, and four secondary strata were formed in each of the remaining 42 non-self-representing PSU's. Thus, the non-self-representing PSU's were partitioned into a total of 170 secondary strata. In a similar manner, the 16 self-representing PSU's were partitioned into 144 secondary strata.

In the second stage of selection, one SSU was selected from each of the 144 secondary strata covering the self-representing PSU's, and two SSU's were selected from each of the remaining secondary strata. All second-stage sampling was with replacement and with probability proportional to the SSU's total noninstitutionalized population. The total number of sample SSU's was 2 × 170 + 144 = 484.

For the third stage of selection, each SSU was first divided into smaller disjoint geographic areas, and one area within the SSU was selected with probability proportional to the total number of housing units in 1970. Next, one or more disjoint segments of at least 60 housing units were formed in the selected area. One segment was selected from each SSU with probability proportional to the segment housing unit count. In response to the sponsoring agencies' request that the expected household sample size be reduced, a systematic sample of one-sixth of the segments was deleted from the sample. Thus, the total third-stage sample was reduced to 404 segments.

For the fourth stage of selection, all of the dwelling units within the segment were listed, and a systematic sample of dwelling units was selected. The procedures used to determine the sampling rate for segments guaranteed that all dwelling units had an approximately equal overall probability of selection. All of the reporting units within the selected dwelling units were included in the sample.

### National Opinion Research Center Sample Design

The land area of the 50 States and Washington, D.C., was also divided into disjoint PSU's for the National Opinion Research Center sample design. A PSU consisted of SMSA's, parts of SMSA's, counties, parts of counties, or independent cities. Grouping of counties into a single PSU occurred when individual counties had a 1970 population of less than 10,000. The PSU's were classified into two groups according to metropolitan status—SMSA or not SMSA. These two groups were individually ordered and then partitioned into zones with a 1970 census population size of approximately 1 million.

A single PSU was selected within each zone with a probability proportional to its 1970 population. It should be noted that this procedure allowed a PSU to be selected more than one time. For instance, an SMSA primary sampling unit with a population of 3 million could be selected as many as four times. The full general-purpose sample contained 204 PSU's. These 204 PSU's

were systematically allocated to four subsamples of 51 PSU's. The final set of 76 sample PSU's was chosen by randomly selecting two complete subsamples of 51 PSU's. One subsample was included in its entirety, and 25 of the PSU's in the other subsample were selected systematically for inclusion in NMCUES.

For the second stage, each PSU selected in the first stage was partitioned into a disjoint set of SSU's defined by block groups, enumeration districts, or a combination of the two types of census units. Within each sample PSU, the SSU's were ordered and then partitioned into 18 zones such that each zone contained approximately the same number of households. One SSU had the opportunity to be selected more than once, as was the case in the PSU selection. If a PSU had been hit more than once in the first stage, the second-stage selection process was repeated as many times as there were first-stage hits. The 405 SSU's were identified by selecting 5 SSU's from each of the 51 PSU's in the subsample that was included in its entirety and 6 SSU's from each of the 25 PSU's in the group for which only one-half of the PSU's were included.

The SSU's selected in the second stage were then subdivided into area segments with a minimum size of 100 housing units each. One segment was then selected with probability proportional to the estimated number of housing units. The final-stage sample, in which a selection of housing units was made, was essentially the same as that used by the Research Triangle Institute.

## Collection of Data

Field operations for NMCUES were performed by the Research Triangle Institute and the National Opinion Research Center under specifications established by the sponsoring agencies. Persons in the sample dwelling units were interviewed at approximately 3-month intervals beginning in February 1980 and ending in March 1981. The core questionnaire was administered during each of the five rounds of interviews to collect data on health, health care, health care charges, sources of payment, and health insurance coverage. A summary of responses was used to update information reported in previous rounds. Supplements to the core questionnaire were used during the first, third, and fifth rounds of interviews to collect data that were not expected to change during the year or that were needed only once. Approximately 80 percent of the third and fourth rounds of interviews were conducted by telephone; all remaining interviews were conducted in person. The respondent for the interview was required to be a household member 17 years of age or older. A proxy respondent not residing in the household was permitted only if all eligible household members were unable to respond because of health, language, or mental condition.

## Imputation

Nonresponse in panel surveys such as NMCUES occurs when sample individuals refuse to participate in the survey (total nonresponse), when initially participating individuals drop out of the survey (attrition nonresponse), or when data for specific items on the questionnaire are not collected (item nonresponse). In general, response rates for NMCUES were excellent. Approximately 90 percent of the sample reporting units agreed to participate in the survey, and approximately 94 percent of the individuals in the participating reporting units supplied complete annual information. Even though the overall response rates are quite high for NMCUES, the estimates of means and proportions may be biased if nonrespondents have different health care experiences than respondents or if there is a substantial response rate differential across subgroups of the target population. Furthermore, totals will tend to be underestimated unless allowance is made for the loss of data because of nonresponse.

Two methods commonly used to compensate for survey nonresponse are data imputation and the adjustment of sampling weights. For NMCUES, imputation was used to compensate for attrition and item nonresponse, and weight adjustment was used to compensate for total nonresponse. The calculation of the weight adjustment factors is discussed in the section on sampling weights.

A specialized form of the sequential hot-deck imputation method was used for attrition imputation. First, each sample person with incomplete annual data (recipient) was linked to a sample person with similar demographic and socioeconomic characteristics who had complete annual data (donor). Second, the time periods for which the recipient had missing data were divided into two categories, imputed eligible days and imputed ineligible days. Imputed eligible days were those days for which the donor was eligible (that is, in scope), and imputed ineligible days were those days for which the donor was ineligible (that is, out of scope). For the recipient's imputed eligible days, the donor's medical care experiences (such as medical provider visits, dental visits, or hospital stays) were imputed into the recipient's record. Finally, the results of the attrition imputation were used to make the final determination of a person's respondent status. If more than two-thirds of the person's total eligible days (both reported and imputed) were imputed, then the person was considered to be a total nonrespondent, and all data for the person were removed from the analytic data file.

The data collection methodology and field quality control procedures for NMCUES were designed so that the data would be as accurate and complete as possible subject to budget considerations. However, individuals

cannot report data that are unknown to them, or they may choose not to report the data even if known. This latter situation is especially true for data relating to expenditures, income, and other sensitive topics. Because of the size and complexity of the NMCUES data base, it was not feasible, from the standpoint of cost, to replace all missing data for all data items. The 12-month data files, for example, contain approximately 1,400 data items per person. With this in mind, the NMCUES approach was to designate a subset of the total items on the data base for imputation of the missing data. Thus, for 5 percent of the NMCUES data items, the responses were edited and missing data imputed by a combination of logic and hot-deck procedures to produce revised variables for use in analysis. Items for which imputations were made cover the following data areas:

- Visit charges.
- Source of payment codes and amounts.
- Annual disability days.
- Health insurance premium amount.
- Length of hospital stay.
- Total weeks worked in 1980.
- Average hours worked per week.
- Educational level.
- Hispanic ethnicity.
- Income.
- Age and birth date.
- Race.
- Sex.
- Health insurance coverage.
- Visit dates.

These items were selected as the most important variables for statistical analyses.

## Construction of Longitudinal Families

At the time of the initial interview, a group of persons sharing a common housing unit was designated a family if they were related to each other by blood, marriage, adoption, or a formal foster care relationship. An unmarried student 17–22 years of age living away from home was also considered a part of the family, even though his or her residence was in a different location. When, on subsequent interviews, this initial sampled social unit was found to have had changes in membership, it became necessary to find a decision rule (or set of decision rules) for deciding when a family continued, when it ended, and when a new family began.

The decision rule chosen was initially referred to as a principal-predecessor-principal-successor rule (Dicker and Casady, 1982; Whitmore, Cox, and Folsom, 1982; Moser et al., 1983). The term came from the

understanding that, at any given point in time, a family may have several predecessor families from which its members came and several successor families into which its members would go. The decisionmaking problem, therefore, was to objectively select only one predecessor family (the principal predecessor) and only one successor family (the principal successor) as representing the family through successive stages in time. If no principal successor family could be found, the initial family had ended. If no principal predecessor family could be found, the current family (at the time of the interview) was a new family. Later discussions in the literature referred to the above rule under a different name. It came to be called a "reciprocal, majority population rule" (McMillen, 1984; Dicker, 1984) because the principal-predecessor-principal-successor rule came to be understood as a rule that linked families on the basis of cross-family majorities. Thus, if two families (as defined above) exist at different but adjacent points in time, they are the same family if and only if a majority of the eligible members of the first family are found in the second family and a majority of the eligible members of the second family are also found in the first family. The reciprocity of the comparison is crucial. A unidirectional majority—either from the first family to the second family or from the second to the first—is not sufficient for the two families to be defined as the same.

Several aspects of the rule as applied in this survey need further elaboration. First, the rule was applied to all families in the longitudinal universe (not only to those in the initial sample) that had cross-membership connections with initially sampled families. Second, only persons eligible over time to be in both families being compared were counted when calculating cross-family majorities. For example, persons in family 1 who died or otherwise left the universe were not eligible for membership in family 2 and were not counted. Likewise, persons who entered family 2 from outside the universe during the interval between interviews, such as a newborn baby or a soldier returning to civilian status, could not have been in family 1 (that is, were not eligible for inclusion in that family) and also were not counted. Third, the reciprocal majority population rule, as stated above, links only two families adjacent in time. However, transitivity between linkages is implied in the rule. This means that given three families (families A, B, and C) existing at three different points in time, if family A is the same as family B and family B is the same as family C, then family A is also the same as family C. A longitudinal family, therefore, is either one or a series of point-interval families linked by the reciprocal majority population rule. Fourth, the final sample of families was limited to initially sampled families and all other families derived from these families that had at least one initially sampled person (a key individual) in them on their beginning date. Thus, the collection of families examined for family construction purposes was divided into key families (a family with

a key individual), which were in the sample and given a positive sampling weight, and nonkey families (a family without a key individual), which were not in the sample and given a sampling weight of zero. One reason for not including nonkey families in the sample is that very little data for them were available. Moreover, assumptions were often required to construct these families. (For more details on this methodology, see Dicker and Casady, 1982, and Whitmore, Cox, and Folsom, 1982.)

The dynamic sample of longitudinal families derived from this process tended to have characteristics that are generally sociologically believed to define the beginning and ending of families. For example, an even merger of two individuals through marriage always produced a new family. Similarly, an even split in a two-person family as the result of divorce or separation always ended the family. On the other hand, an uneven split in a larger family would not necessarily end such a family. In most cases, the original family continued as the larger part of the split. For example, if an adult child left a family of three persons or more to set up a separate household, in most cases the original family continued as the same but smaller family. Such an outcome appears to be in agreement with the sociological consensus that the loss of a single family member, other than the head or spouse, does not usually end the original family. The majority of uneven splits arise from this type of situation.

By the same reciprocal majority rule, however, a separation of husband and wife in a situation where children remained with one of the spouses in most cases continued the old family, now reconstituted as a single-spouse family with children. This result may not appear to be the sociologically preferred one. However, a more detailed review of the class of events of which this is a special case suggests that this result is in line both with the results based on sampling criteria for other members of the class and with sociological expectations of what the result should be for those class members. For example, given a head-spouse family with children, the loss of a head or spouse because of death or institutionalization is rarely thought of sociologically as an event ending the family. Rather, the social consensus appears to be that the original family continues, although in a recognizably changed state. The same may be said for the situation in which a head or spouse enters the military or goes overseas and is absent from the family for long periods. The family is not defined as ended but as continuing with an absent spouse. In this survey, all of the above events are defined as out-of-scope sampling events that cannot affect the identity of the family over time. Therefore, families would not end because of their occurrence. Only when the separating head or spouse remains within the noninstitutionalized U.S. population (the universe of inference) does the dilemma arise from sampling and sociological considerations as

to whether the original family has ended. This inscope event, however, is similar in its effect on family functioning as the four previously mentioned out-of-scope events. In all of these situations, the family loses a significant role player. As a consequence, important family role obligations go unfulfilled (or only partially fulfilled). It seemed appropriate, therefore, to treat all of these events in the same manner (as a functionally equivalent happening) for the purpose of constucting longitudinal families. Given the lack of a sociological consensus for treating the above class of events, the reciprocal majority population rule produces an appropriate, if not consensual, decision. When the separating head or spouse or adult child remains within the universe, the reciprocal majority population rule must also be applied to find out if he or she has formed a new family. The decision will depend on whether the person joins a previously existing family in the universe and the size of the family joined.

An uneven merger of two preexisting families also presents some decisionmaking problems from a sociological perspective. Such mergers occur when one or more related persons join another set of related persons or when a marriage occurs and one or more of the marriage partners bring children from a previous marriage (or another related person) with them. The first type of situation presents few problems. Most of these cases involve the entering or reentering of continuing families by elderly parents, adult children, or other relatives. Usually these new family members constitute the smaller of the two merging families. The larger of the two families entering the merger generally has reciprocal majority linkages to the newly merged family. (The smaller family never has.) The two reciprocally linked families are considered one continuing family. Occasionally, an uneven merger may produce a totally new family if the merged family cannot be linked to any preexisting family. The above result appears to be in line with the general sociological consensus that a family's identity is not changed by the addition or return of elderly parents, adult children, etc. Of course, if the additional family members come from out of scope (that is, if they are newborn children, come out of an institution, or return from the military or from overseas), they do not affect the identity of the family. These instances probably represent the majority of uneven mergers. However, there is less sociological consensus as to what the merged family represents when an uneven merger results from a marriage. The reciprocal majority population rule treats this situation in the same manner as the preceding one. For situations in which a single spouse enters an already existing larger family, the result appears appropriate. Where both spouses bring large families into the marriage, the result may be questionable. However, these latter situations represent a very small number of cases.

## Construction and Use of Family Weights

### Initial Family Weights

The target population of the household survey (HHS) was civilian noninstitutionalized families existing in the United States at any time during 1980. The universe of families existing on any specific day during 1980 was potentially different from that existing on any other day of the year. Conceptually, one could have conducted a census of the eligible population of the United States on January 1, 1980. By following this initial universe of families throughout the year, every unique longitudinal family unit could be identified and labeled. These longitudinal family units are defined by a beginning date, an ending date, and a set of persons who qualify as eligible (civilian and noninstitutionalized) family members. In addition to all family units that can be linked to the initial January 1 family universe, there are persons and families who were ineligible on January 1, 1980, but subsequently returned to the civilian noninstitutionalized population without merging with families containing individuals who were eligible on January 1. Such individuals and families were eligible for the sample but did not have a chance of entering it. Poststratification weight adjustments partially compensated for this undercoverage.

The family weights for longitudinal families in the household sample were developed from the sampling weights for the initially sampled families, which were called originating base reporting units (OBRU's). For each HHS longitudinal family, the key family members all belonged to the same OBRU. Hence, the initial family weight for the $j^{th}$ key HHS longitudinal family was computed as follows:

$$WF_1(j) = [n(j)/g(j)]\, w_o(j),$$

where $n(j)$ is the number of key individuals in family $j$ on its beginning date, $g(j)$ is the total number of members of family $j$ on its beginning date, and $w_o(j)$ is the OBRU initial sampling weight for the key members of family $j$. Thus, the initial family weight is the OBRU sampling weight adjusted for person-level multiplicity. Essentially, this formula means that the sampling weight of a family beginning on January 1, 1980, is the same as the household sampling weight, regardless of when the family ended or family membership changed in the subsequent 12 months. However, if a family began on some day after January 1, 1980, the household sampling weight was adjusted to take into account the fact that the new family may have had multiple chances of getting into the sample. However, as previously pointed out, positive sampling weights were developed only for key longitudinal families. Further details of the methodology for HHS longitudinal sampling weights are provided by Whitmore, Cox, and Folsom (1982).

## Adjustment for Undercoverage and Nonresponse

Poststratification adjustment of the initial HHS family weights to the family counts based on the March Supplement to the 1980 Current Population Survey (CPS) was used to reduce the variance of estimators and the bias from undercoverage. These counts, however, were from estimates based on an updating of the 1970 census. Therefore, NMCUES family counts and estimates may not agree with family counts and estimates based on the 1980 census. The poststratification adjustments and a weighting class adjustment were also used to reduce the bias from nonresponse of longitudinal families.

A key HHS longitudinal family was classified as responding if it satisfied the following three requirements:

1. At least one key family member was classified as a respondent; that is, at least one key family member responded for at least one-third of his or her eligible days in the survey.

2. The total number of responding (known eligible) days during the family's existence summed over all family members is at least one-third of the total number of eligible days during the family's existence summed over all members of the family.

3. The family contained no students who were listed only on the parents' round 1 secondary reporting unit roster and for whom no other data collection instrument was ever received.

This definition of a responding family was felt to be consistent with the definition of person-level response and was used to create the HHS family response indicator variable. Only about 0.1 percent of all longitudinal families were declared to be nonresponding because of condition 3. Imputation of a full year of data for these students was problematic. Hence, inclusion of condition 3 in the definition of a responding family was felt to be cost effective.

The initial multiplicity-adjusted family weight was computed for all longitudinal families from the initial OBRU weight. A poststratification adjustment was then made for nonresponse of families linked to nonresponding OBRU's, producing an adjusted weight. A weighting class adjustment was performed for nonresponding longitudinal families generated by responding OBRU's. This adjusted weight was then truncated to produce a new family weight. The final adjustment was a poststratification and smoothing to the March Current Population Survey family counts to produce the final HHS longitudinal family weight, FWEIGHT. An alternative family weight, AWEIGHT, which was adjusted for each family's eligible days, was also computed from FWEIGHT to facilitate analytic tabulations. AWEIGHT, a time-adjusted family weight, is equal to FWEIGHT times the proportion of 1980 for which the

family existed. (Computationally, it equals FWEIGHT times the family's survey eligibility days divided by 366, the total number of days in 1980.) The time-adjusted family weights, AWEIGHT, sum to the average daily number of HHS-eligible longitudinal families in the United States in 1980.

## Estimators

This family weighting scheme produces the adjusted family weight, FWEIGHT, which can be used directly for estimation of annual health care utilization and expenditure. For example, if $Y(j)$ represents the total expenditure of the $j^{th}$ HHS longitudinal family for a particular medical service in 1980, then

$$\Sigma FWEIGHT(j)Y(j)$$

estimates the total expenditure of all civilian noninstitutionalized families in the United States for this medical service in 1980, where the summation extends over all longitudinal families in the NMCUES HHS sample.

Rates of utilization and expenditure are, however, of more interest than population totals. The rates of annual utilization and expenditure per family for a given family domain, say domain $d$, are defined at the population level by

$$R(d) = [\sum_{j=1}^{J} X_d(j)Y(j)]/[\sum_{j=1}^{J} X_d(j)PE(j)],$$

where $j = 1,\ldots, J$ indexes the population of all key longitudinal families that ever existed in 1980 (that is, all longitudinal families that had a chance for selection as key NMCUES families);

$X_d(j) = 1$ if family $j$ belongs to domain $d$, 0 otherwise;

$Y(j) = $ total utilization or expenditure for family $j$ during the portion of 1980 that family $j$ was eligible for NMCUES; and

$PE(j) = $ proportion of 1980 that family $j$ was eligible for NMCUES, or (FAMEND − FAMBEG + 1)/366, where FAMEND = family ending date (days of 1980 numbered 1 through 366) and FAMBEG = family beginning date.

The family aggregates, $Y(j)$, can be viewed as sums of associated person-level visit counts or expenditures for key and nonkey individuals belonging to family $j$ during the time period in which they were members of the family. The denominator of $R(d)$ is the average daily number of families of type $d$ that existed during 1980. The bracketed portion of the numerator of $R(d)$ is simply the total number of health care visits or the total expenditures of a specified type experienced by NMCUES eligible persons while they belonged to families of type $d$.

Unbiased estimators for the numerator and denominator of $R(d)$ lead to the ratio estimator $r(d)$, for which the equation is

$$r(d) = [\Sigma FWEIGHT(j)X_d(j)Y(j)] / [\Sigma FWEIGHT(j)X_d(j)PE(j)],$$

where the summation extends over all longitudinal families in the sample. Of course, it is necessary to compute $X_d(j)$ and $PE(j)$ only for responding families because FWEIGHT is zero for all other families. Two alternative formulations of this estimator that may be more convenient for some computations are

$$r(d) = [\Sigma AWEIGHT(j)X_d(j)Y(j)/PE(j)] / [\Sigma AWEIGHT(j)X_d(j)],$$

and

$$r(d) = [\Sigma FWEIGHT(j)X_d(j)Y(j)] / [\Sigma AWEIGHT(j)X_d(j)],$$

where the summations extend over all longitudinal families and AWEIGHT$(j)$, as previously noted, is the final time-adjusted weight for family $j$; that is,

$$AWEIGHT(j) = FWEIGHT(j) PE(j).$$

Throughout this report, all estimates are based on the first of these two alternative formulations. All counts of expenditures for health care employ as the measure of expenditure used

$$\Sigma AWEIGHT(j)X_d(j)Y(j)/PE(j),$$

and all counts of families employ as the number of families in question

$$\Sigma AWEIGHT(j)X_d(j).$$

To be more specific, the statistics presented in the detailed tables of this report are estimated as follows.

The *number of families* with given characteristic(s) is estimated as

$$\Sigma AWEIGHT(j)X_d(j),$$

where $X_d(j) = 1$ if family $j$ has the characteristic(s) in question and 0 otherwise.

Note that this estimator estimates the number of family years experienced by families with the given characteristic(s) or, equivalently, the average number of families with the given characteristic(s) that would have been found at a randomly chosen point in time in 1980. It is, in general, less than the cumulative total of distinct longitudinal families with the given characteristic(s) that ever existed at any time in 1980, some of which existed for only part of the year.

The *mean* for use or expenditure is always the mean rate per family year and is estimated as

$$[\Sigma \text{AWEIGHT}(j)X_d(j)Y(j)/PE(j)] / [\Sigma \text{AWEIGHT}(j)X_d(j)].$$

The *percent* of families with a given characteristic is estimated as

$$[\Sigma \text{AWEIGHT}(j)X_d(j)X_u(j)] / [\Sigma \text{AWEIGHT}(j)X_d(j)]$$

where $X_u(j) = 1$ if family $j$ has the given characteristic and 0 otherwise.

Note that this estimator has as its denominator the estimated number of family years experienced by all families in a domain defined by a set of family characteristics and has as its numerator the estimated number of family years experienced by families in the domain that also have the utilization characteristic in question. In other words, the estimator involves a ratio of family years.

## Special Requirements for Imputation of Family Data

As noted in the previous section, estimation of utilization and expenditure rates requires family aggregate data, say $Y(j)$, where the aggregates can be obtained as sums of associated person-level visit counts or expenditures. To compute the family aggregate $Y(j)$, it is necessary to sum over all members of family $j$, both key and nonkey. Moreover, computation of annual utilization and expenditure statistics requires a full year of data for every member of each responding family. Hence, in the attrition imputation, a weighted sequential hot-deck procedure was used to produce complete data for individuals who did not respond for the full year. In the attrition task (Cox and Sweetland, 1982), each individual was first classified as either having complete data or having incomplete data, based on whether the individual had responded for all 366 days in 1980. The data records for individuals who had not responded for the full year were completed by attrition imputation, including imputation of eligibility status (eligible or ineligible) for each day in 1980. The major importance of the attrition task is that it provided a full year of data for every individual from which family aggregates, $Y(j)$, can be computed. The concept of a key responding family was defined in such a way, however, that minimal use of data from the attrition task is required. Of course, missing item data can also lead to missing values for the family aggregate, $Y(j)$. Hence, item imputation procedures (Cox et al., 1982) were performed in addition to attrition imputation to assure the availability of complete data for important analytical variables for every eligible day for each family member.

## Reliability of Estimates

### Standard Errors

The estimates presented in this report are based on a sample of the target population rather than on the entire population. Thus, the values of the estimates may be different from values that would be obtained from a complete census. The difference between a sample estimate and the population value is referred to as the sampling error, and the expected magnitude of the sampling error is measured by the standard error. Estimated standard errors for the estimates in Tables A–F are generally next to each estimate.

The SESUDAAN (Shah, 1981) standard error estimation software package was used to produce the estimates of standard errors. SESUDAAN is a Taylor Series procedure, developed and released by the Research Triangle Institute. It runs within the Statistical Analysis System (SAS Institute, Inc., 1982).

In addition to sampling errors, the estimates presented in this report are subject to nonsampling errors, such as biased interviewing and reporting, undercoverage, and nonresponse. The standard error does not provide an estimate of nonsampling errors. However, as discussed in preceding sections, every effort was made to minimize these errors.

### Confidence Intervals

The estimates in this report are subject to sampling error. The true values are unknown. But the sampling error can be used to determine a range of values such that the true value will be within that range with a known probability. This range is called a confidence interval.

Suppose that $\hat{\theta}$ is an unbiased estimator for the parameter $\theta$, and $S_{\hat{\theta}}$ is a consistent estimator for the standard error of $\hat{\theta}$. Under appropriate central limit theorem assumptions regarding $\hat{\theta}$, the statistic $Z = (\hat{\theta} - \theta)/S_{\hat{\theta}}$ has an approximate standard normal distribution for large samples. Thus, an approximate $(1 - \alpha) \times 100$ percent confidence interval for $\theta$ is given by

$$(\hat{\theta} + z_{\alpha/2}S_{\hat{\theta}}, \hat{\theta} + z_{1-\alpha/2}S_{\hat{\theta}}),$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the appropriate values from a standard normal table.

As an example, Table B shows that, of all older multiple-person families in the civilian noninstitutionalized population of the United States, an estimated 27.6 percent had total charges for health care of \$3,000 or more in 1980. The estimated standard error of 27.6 percent is 1.7 (Table B). As $Z_{.025} = -1.96$ and $Z_{.975} = 1.96$, a 95-percent confidence interval for the percent of all older multiple-person families with such charges in 1980 is $27.6 \pm (1.96 \times 1.7)$, or the interval 24.3–30.9 percent. Approximately 95 percent of the confidence intervals constructed in this

manner will contain the true percent of families with total charges for health care of $3,000 or more in 1980.

Confidence intervals for the difference of two parameters can be constructed in a similar manner. Suppose $\theta_1$ and $\theta_2$ are the values of the parameter of interest in two mutually exclusive population subgroups. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators of $\theta_1$ and $\theta_2$, respectively, then $\hat{d} = \hat{\theta}_1 - \hat{\theta}_2$ is unbiased for $d = \theta_1 - \theta_2$ and

$$\text{Var}(\hat{d}) - \text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2) - 2\,\text{Cov}(\hat{\theta}_1,\hat{\theta}_2).$$

Unfortunately, the estimation of $\text{Var}(\hat{d})$ presents a problem because it is not possible for the National Center for Health Statistics to provide the reader with covariance estimates for all possible pairs of subdomains of potential interest. However, if it is reasonable to assume that $\text{Cov}(\hat{\theta}_1,\hat{\theta}_2) = 0$, the standard error of $d$ can be estimated by

$$S_{\hat{d}} = \sqrt{S_{\hat{\theta}_1}^2 + S_{\hat{\theta}_2}^2}.$$

Then, under appropriate central limit theorem assumptions regarding $d$, the statistic $Z_d = (\hat{d} - d)/S_{\hat{d}}$ has an approximate standard normal distribution for large samples, and the interval

$$(\hat{d} + z_{\alpha/2}S_{\hat{d}}, \ \hat{d} + z_{1-\alpha/2}S_{\hat{d}})$$

is an approximate $(1 - \alpha) \times 100$ percent confidence interval for the difference $d$.

For example, suppose we wanted to construct a 95 percent confidence interval for the difference between the percent of older families with incomes of $15,000 per year or more and total charges for health care of $3,000 or more ($\theta_1$) and the percent of older families with incomes of less than $15,000 per year and the same amount of total charges ($\theta_2$). It can be seen in Table B that $\hat{\theta}_1 = 29.9$ and $\hat{\theta}_2 = 25.5$, so

$$\hat{d} = \hat{\theta}_1 - \hat{\theta}_2$$

$$= 29.9 - 25.5$$

$$= 4.4.$$

From Table B, it can be seen that $S_{\hat{\theta}_1} = 2.3$ and $S_{\hat{\theta}_2} = 2.1$, so

$$S_{\hat{d}} = \sqrt{S_{\hat{\theta}_1}^2 + S_{\hat{\theta}_2}^2}$$

$$= \sqrt{5.29 + 4.41}$$

$$= \sqrt{9.70}$$

$$= 3.11.$$

Then, as $\alpha = .05$, it follows that $z_{\alpha/2} = -1.96$ and $z_{1-\alpha/2} = 1.96$, so the 95 percent confidence interval for the difference of interest is $(10.48, 0.00)$.

The reader should be aware that the assumption that $\text{Cov}(\hat{\theta}_1,\hat{\theta}_2) = 0$ is frequently not true for complex sample surveys. This warning is especially germane for sample designs, such as the NMCUES design, that rely on cluster sampling at one or more stages of sample selection. If $\text{Cov}(\hat{\theta}_1,\hat{\theta}_2)$ is positive, the confidence interval will tend to be too large, and the confidence level will be understated. More seriously, if $\text{Cov}(\hat{\theta}_1,\hat{\theta}_2)$ is negative, the confidence interval will tend to be too small, and the confidence level will be overstated.

### Hypothesis Testing

The statistics $Z$ and $Z_d$ can be used to test hypotheses. For example, the size $\alpha$ critical region for the composite hypothesis

$$H_0 : d \geq d_0$$

versus

$$H_A : d < d_0$$

is given by

$$Z_{d_0} = \frac{\hat{d} - d_0}{S_{\hat{d}}} \leq z_\alpha.$$

As an example, suppose that before any data were collected one had a reason to believe that the percent of older families with total charges for health care of $5,000 or more that also had incomes of less than $15,000 per year and one or more discharges from a hospital ($\theta_1$) was less than the percent of older families with the same amount of total charges and the same hospital discharge history, but who differed in that they had incomes of $15,000 or more per year ($\theta_2$). Letting $d = \theta_1 - \theta_2$, this can be restated as a formal hypothesis as

$$H_0 : d \geq 0$$

versus

$$H_A : d < 0.$$

Note that what is believed to be the true state of nature is reflected by the one-sided alternative.

It can be seen from Table B that

$$\hat{d} = 39.0 - 52.6 = -13.6$$

and

$$S_{\hat{d}} = \sqrt{18.49 + 16.81}$$

$$= 5.94,$$

so that $Z_{d_0} = -2.29$. A multiple $t$-test based on the Bonferroni inequality (Levy and Lemeshow, 1980) will be used to assess the significance of the comparison. Comparing two categories, two at a time, and not taking sign into account gives one possible comparison. Use of the table in Levy and Lemeshow (1980, p. 296)

gives a one-tail critical value of $-1.65$. Therefore, $H_0$ is rejected in favor of $H_A$ as $Z_{d_0} \leq z_\alpha$.

As discussed earlier, the assumption that $\text{Cov}(\hat{\theta}_1, \hat{\theta}_2) = 0$ must be carefully evaluated. If in fact the covariance is positive, the size of the test will be smaller than $\alpha$, and if the covariance is negative, the size of the test will be larger than $\alpha$. Readers who want to conduct more sophisticated analyses of the NMCUES data are advised to consult with a statistician knowledgeable in the analysis of data from complex sample surveys.

# Appendix III
# Definition of Terms

*Accidents, injuries, and poisonings*—This category includes injuries; wounds; burns; poisonings; toxic effects; complications of medical and surgical care; and early complications, late effects, and impairments due to the previous causes. Conditions were reported by the family respondent and recoded according to the International Classification of Diseases, 1975 revision.

*Age of family head*—Age is as of January 1, 1980.

*Ambulatory physician visit*—A visit by a patient to a physician's office, clinic, or similar place is an ambulatory physician visit. Visits are counted whether a physician or only a member of the physician's staff is seen. House calls and visits to school or workplace clinics are also included. Family visits are the sum of all visits by family members during the time they were in the family.

*Bed days*—Bed days are days spent in bed by a family member because of illness or injury. Family bed days are the sum of all bed days of family members during the time they were in the survey, prorated to the time they were in the family.

*Cancer and other neoplasms*—This category includes malignant neoplasms of all sites and tissues, benign neoplasms, carcinoma in situ, and other and unspecified neoplasms. Conditions were reported by the family respondent and recoded according to the International Classification of Diseases, 1975 revision.

*Circulatory and heart disease*—This category includes rheumatic fever, rheumatic heart disease, hypertensive disease, ischemic heart disease, diseases of the pulmonary circulation, other forms of heart disease, cerebrovascular disease, and other diseases of the circulatory system. Conditions were reported by the family respondent and recoded according to the International Classification of Diseases, 1975 revision.

*Civilian noninstitutionalized family*—This refers to families in which all members are members of the civilian noninstitutionalized population. Families whose heads are members of the military are defined as not being civilian families and are excluded in their entirety from this report, although they were included in the sample and the weighting. In the sample, there were 49 such families (about 0.7 percent). Family members other than the head who were in the military were excluded from the survey even if they resided with the family.

*Dental visit*—A visit to a dentist's office is a dental visit. A dentist or a member of the dentist's office staff may have provided services. Family visits are the sum of all visits by family members during the time they were in the family.

*Education of family head*—The years of school completed by family heads 17 years of age and over constitute the education of family heads. Only years completed in regular schools, where persons are given a formal education, are included. A "regular" school is one that advances a person toward an elementary or high school diploma or a college, university, or professional school degree. Thus, education in vocational, trade, or business schools outside the regular school system was not counted in determining the highest grade of school completed.

*Ethnicity of family head*—The ethnicity of family heads 17 years of age and over is as reported by the family respondent. The ethnicity of family heads under 17 was imputed. Ethnicity is classified as (1) Hispanic, which includes Puerto Rican, Cuban, Mexican, Mexicano, Mexican American, Chicano, other Latin American, and other Spanish or (2) non-Hispanic.

*Family*—A family is a group of people who share a common housing unit and are related to each other by blood, marriage, adoption, or a foster care relationship. An unmarried student 17–22 years of age living away from home is also considered part of a family even though his or her residence was in a different location. The group of people who compose the family may change composition over time, causing the family to take on one or a combination of the following time-related states: existing over time without change in membership; existing over time with change in membership; going out of existence before the end of the survey; coming into existence after the beginning of the survey; or existing for the whole survey. For more detail, see Appendix II.

*Family dynamics*—A family is considered unchanging, or static, if it existed for the whole of 1980 and its membership was unchanged. Families that had changes in membership and/or did not exist for the whole of 1980 are considered changing, or dynamic, families.

*Family illness days in bed*—See bed days.

*Family income in 1980*—For each person in the family, data were collected on 12 categories of income. These included income from employment for persons 14 years of age and over; income from various government programs; income from pensions; alimony or child support; interest income; and net rental income. When

information was missing, income was imputed. The total income of persons who were members of more than one family was allocated to each family they were in, in proportion to the amount of time they were in that family. Person-level incomes in each family were summed to create a family-level total. If a family did not exist for an entire year, the family income was adjusted to an annual basis by dividing actual income by the proportion of the year the family existed.

*Family-paid premiums*—The amount paid for premiums by a family and not reimbursed. Much of the cost of premiums in the United States is paid by employers or by public funds rather than by families.

*Family size*—The time-weighted average number of persons in a family determines the size. Family size was computed by (1) summing the number of days in the family for each person who was ever a family member and (2) dividing this sum by the number of days the family was in existence. For example, if a family existed for 200 days and had two persons who were members throughout its existence and one person who was a member for 80 days, the family size is 2.4.

*Family structure*—Family structure refers to the presence or absence of family head, spouse, and children under 17, and whether these persons were present for the family's entire duration or part of its duration.

*Family work-loss days*—See work-loss day.

*Family years*—Family years refers to the length of time that a family, or a collection of families, existed as a unit of analysis in (were eligible for) the survey, as measured in units of a year or fractions of such units.

For an individual longitudinal family in the NMCUES sample, the number of family years equals the number of days the family was eligible for the NMCUES sample divided by 366, the number of days in 1980 (the NMCUES sample period). For such a family weighted to represent a group of families in the NMCUES universe, the number of family years is AWEIGHT($j$), which is equal to FWEIGHT($j$), the basic adjusted weight, times $PE(j)$, the proportion of the year the family was eligible for the sample. For a group of sample families, the associated number of family years is the sum of the AWEIGHT's. For further details and fuller definitions of variables, see the section on estimators in Appendix II.

*Financial burden index*—See index of financially burdensome health care expenses.

*Group quarters*—This is a structure occupied by five or more unrelated people who lived or ate together, or for whom there was neither direct access from the outside or through a common hall nor complete kitchen facilities. Only noninstitutional group quarters were included in the NMCUES sample frame. Each unrelated person in a group-quarter household was considered a separate one-person family, unless he (or she) was a student away from home. (See definition of family.)

*Head of family*—A person was designated as the family head by the respondent at the time of the first interview. If no head was designated or this information was missing, a family head was imputed. Among families in which the person designated as head changed over time, the characteristics of the person who was designated head the longest were used for all head-of-family variables.

*Health care coverage*—Health care coverage refers to the situation in which a private insurance plan or public health care coverage program (Medicare, Medicaid, and so forth) can be used to pay all or a part of a family's or person's health care costs. For this report, completeness of health care coverage was coded into four categories: (1) "full coverage," meaning all family members had health care coverage during their entire survey eligibility period; (2) "partial coverage 1," meaning all family members had health care coverage at some time, but some or all had coverage for only part of their survey eligibility period; (3) "partial coverage 2," meaning some family members had health care coverage for at least part of their survey eligibility period, but some never had coverage; and (4) "no coverage," meaning no family members had health care coverage at any time during their survey eligibility period.

For this report, a family was coded as having a particular source of health care coverage (such as private insurance, Medicare, Medicaid, or a particular combination of coverages) on the basis of the known coverage of family members. Only when the source of coverage was unknown, or not assignable, for *all* family members was the family coded as having source of coverage unknown. However, the coding categories for individuals upon which the family health care coding in this report was based do not identify the coverage source(s) for individuals with part-year or no coverage. Thus, most families coded as having source of coverage unknown are families with no members having full-year coverage. Also, the coding categories for individuals upon which the family health care coverage coding in this report was based are different from the categories used in a previous family report (Dicker, 1983a) that dealt with only a part of the survey year. As a result, there may be differences in coverage estimates between the reports.

In this report, coverage by CHAMPUS is classified as coverage by private insurance.

*Health care services*—NMCUES includes information on eight types of health care services. These are (1) inpatient hospital care, (2) inpatient physician care, (3) ambulatory physician visits, (4) hospital emergency room and outpatient visits, (5) dental visits, (6) prescription acquisitions, (7) services of other *independent* medical providers such as chiropractors, speech therapists, faith healers, and psychologists (unless such providers are working as part of a physician's staff, in which case their services are counted in physicians' care), and (8) acquisition of other health care supplies such as eyeglasses, orthopedic items, hearing aids, ambulance services, and diabetic items. Excluded from the data in this report are nonprescription medicines, nursing home care, and care in other long-term care institutions.

*Hospital admission*—Hospital admission is the formal acceptance by a hospital of a patient who is provided room, board, and regular nursing care in a unit of the hospital, including patients admitted for childbirth. A patient admitted to the hospital and discharged on the same day is included as a hospital admission. A hospital stay resulting from an emergency department visit is also included. Family hospital admissions are the sum of all admissions of family members during the time they were in the family.

*Hospital discharge*—A hospital discharge is the formal release by a hospital of a patient who was provided room, board, and regular nursing care in a unit of the hospital. A patient admitted to the hospital and discharged on the same day is included as a hospital discharge. A hospital stay resulting from an emergency room visit and subsequent admission of the patient is also included. Family hospital discharges are the sum of all discharges of family members during the time they were in the family.

*Hospital emergency room*—The emergency room is a facility within a hospital organized to provide medical services to people needing immediate medical or surgical intervention. People receiving care in the emergency room may be admitted to a hospital.

*Hospital emergency room visit*—This is a face-to-face encounter between a patient (not necessarily ambulatory) and a medical person in the hospital emergency room. Encounters by patients transported to the emergency room by police or by emergency medical service are included. The visit may result in a hospital admission. Family emergency room visits are the sum of all emergency room visits by family members during the time they were in the family.

*Hospital outpatient department*—This is a hospital-based ambulatory care facility organized to provide non-emergency medical services. Persons receiving services do not receive inpatient nursing care. Examples of outpatient departments or clinics are pediatrics, obstetrics and gynecology, eye, and psychiatric.

*Hospital outpatient department visit*—This is a face-to-face encounter between an ambulatory patient and a medical person in a hospital outpatient department. The patient comes to a hospital-based ambulatory care facility to receive services and departs on the same day. If more than one department or clinic was visited on a single trip, each department or clinic visited was counted as a separate visit. Family outpatient department visits are the sum of all hospital outpatient department visits by family members during the time they were in the family.

*Household*—This refers to occupants of a housing unit or group quarters included in the sample. A household can be one person, a family of related people, a number of unrelated people, or a combination of related and unrelated people. Therefore, a household can contain more than one family. (See definition of family.)

*Housing unit*—A housing unit is a group of rooms or a single room occupied or intended for occupancy as separate living quarters. This means that (1) the occupants do not live and eat with any other persons in the structure, and (2) there is either direct access from the outside or through a common hall or complete kitchen facilities for the use of the occupants only.

*Index of financially burdensome health care expenses*—This is the ratio of total family out-of-pocket expenses for health (including both family-paid premiums and out-of-pocket expenses for health care services) to total family income.

*Inpatient hospital care*—This is health care provided to a patient by a hospital during the period from the patient's admission to the patient's discharge. This includes admissions for deliveries of babies.

*Inpatient physician care*—This care is provided to a patient by a physician (or a physician's staff) during the period from the patient's admission to a hospital to the time of the patient's discharge from the hospital. Such care was only recorded in NMCUES if a physician's bill, separate from the hospital bill, was rendered for such care. Otherwise, a patient was not regarded as having received inpatient physician care.

*Institution*—An institution is a place providing room, board, and certain other services for residents or patients. Correctional institutions, military barracks, and orphanages were always considered institutions in NMCUES. Places that provide long-term health care were also identified as institutions if they provide either nursing or personal care services. Certain other facilities licensed, registered, or certified by a State agency or affiliated with a Federal, State, or local government agency were also defined as institutions. People residing in institutions were not included in the household sample.

*Key person*—See the discussion under "Sample Design" in Appendix II.

*Limitation in major activity*—Four categories were developed for classifying limitation in major activity:

1. Cannot perform usual major activity (such as working, going to school, or keeping house).

2. Can perform usual major activity but limited in kind or amount.

3. Can perform usual major activity but limited in kind or amount of other activity.

4. Not limited.

People 6 years of age and over were classified into any of the categories; children 1–5 years of age were classified into categories 1, 2, and 4; and children under 1 year of age were classified into categories 1 and 4. In this report, categories 2 and 3 are combined into the category "secondary limitation." The NMCUES classified persons with unknown limitations as not limited.

*Longitudinal family*—A longitudinal family is a family identified as the same family over a time period. It may or may not have had changes in family membership during the time period. (See the definition of family.)

*Marital status*—Marital status for each person 17

years of age and over is as indicated by the household respondent.

*Metropolitan area*—See urban-rural location.

*Metropolitan center city*—See urban-rural location.

*Metropolitan suburb*—See urban-rural location.

*Multiple-person family*—A family with an average size of 1.5 members or more is a multiple-person family.

*National household component*—One component of NMCUES, this consists of multiple household interviews with an area probability sample of people in the noninstitutionalized population of the United States in 1980.

*Nonkey person*—See the discussion under "Sample Design" in Appendix II.

*Nonmetropolitan urban area*—See urban-rural location.

*Nonurban area*—See urban-rural location.

*Number of families*—This refers to the average number of families with a given set of characteristics that would have been found at a randomly chosen point in time in 1980. This is equal to the number of family years experienced during 1980 by families with the given characteristics. It is, in general, less than the cumulative total number of distinct longitudinal families with the given characteristics that ever existed at any time in 1980, some of which existed for only part of the year.

*One-person family*—A family with average size less than 1.5 is a one-person family. More than one individual may be involved.

*Out-of-pocket expenses*—This is the amount paid by a family and not reimbursed by either insurance or other health care payment programs.

*Out-of-pocket expenses for health care services*—This is the amount paid out-of-pocket by a family for all eight of the types of health care services covered by NMCUES. (See definition of health care services, above.) This does not include family-paid premiums.

*Perceived health status*—This is the family respondent's rating on a 4-point scale of the health of a family member compared with the health of other persons of the same age, as reported at the time of the first interview. The categories are "excellent," "good," "fair," and "poor." When a family consisted of only one member, this was a self-rating.

*Point-interval family*—A point-interval family is a family with exactly the same family membership over a time period. A change in family membership ends one point-interval family and begins another.

*Poverty status of family*—The poverty status in 1980 was calculated by dividing the family's income in 1980 by the appropriate 1980 poverty level threshold and converting it to a percent. For example, a family with income between two and three times the poverty level threshold that corresponds to its size and other characteristics would be classified in the 200–299 percent category. The poverty level thresholds, as used by the U.S. Bureau of the Census, are determined by the age and sex of the family head and the average number of persons in the family. In 1980, average poverty level thresholds by family size (weighted for the mix of families by sex

and age of head) were: 1-person, $4,190; 2-person, $5,368; 3-person, $6,565; 4-person, $8,414; 5-person, $9,966; 6-person, $11,269; 7-person, $12,761; 8-person, $14,199; 9-person and larger, $16,896.

*Premium*—The premium is the amount paid for private health insurance or other health care coverage.

*Prescription acquisition*—This describes the obtaining of a medication by a family member requiring a prescription from a doctor or dentist. Both initial fillings of prescriptions and refills are counted as acquisitions. Family prescription acquisitions are the sum of all acquisitions by family members during the time they were in the family.

*Principal respondent*—This is the member of the reporting unit who provided most of the information for the people in the reporting unit.

*Proxy respondent*—As used in this survey, a proxy respondent was a person who provided information for people in the reporting unit but who was not a member of the reporting unit. A proxy respondent was used only when no member of the reporting unit could supply the information because of physical or mental incapacity.

*Race of family head*—The race of the family head is as reported by the family respondent or imputed. Race is classified as "white," "black," or "other." The "other" race category includes American Indians, Alaskan Natives, Asians, Pacific Islanders, and people not identified by race. The category "all other" includes the categories "black" and "other."

*Rate per family year*—Amount of care used or dollars expended by a family or group of families is divided by the number of family years experienced by these families while eligible for the NMCUES sample. All data on the use of care or on health expenditure in this report are presented in terms of a rate per family year. For a given family, the rate per family year equals $Y(j)/PE(j)$,

where $Y(j)$ = use of care or expenditure during family's period of eligibility for NMCUES sample, and

$PE(j)$ = proportion of year family was eligible for the NMCUES sample.

The section on estimators in Appendix II presents more details of calculations.

*REF. DATE*—The reference date was the date of the previous interview in most cases. For the first interview, however, it was January 1, 1980. For new persons, it was the date they joined the reporting unit. For the final interview, it spanned the time between the next-to-last interview and December 31, 1980.

*Region*—The 50 States and the District of Columbia are categorized into four regions by the U.S. Bureau of the Census. This classification by region was used in this report. The regions and their constituent parts are as follows. NORTHEAST: Maine, New Hampshire,

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

*Reporting unit (RU)*—A reporting unit is the basic unit for collecting data in the household components of NMCUES at the time of each interview. A reporting unit consisted of all related people residing in the same housing unit or group quarters during the reference period covered by a particular interview. One person could give information for all members of the reporting unit.

*Secondary reporting unit*—Unmarried students 17–22 years of age usually living in a sampled household but away from home as full-time students were considered secondary reporting units. Also, in a household with multiple families, the reporting unit with the largest number of individuals was usually designated the primary reporting unit, and all other families were designated secondary reporting units.

*Sex*—Sex was recorded by the interviewer in the initial NMCUES interview.

*Spouse*—The spouse is the person designated by the respondent as the spouse of the head of the family.

*Total charges*—This is the full amount billed (either actual or imputed) to a family for all eight types of health care services covered by NMCUES, whether these amounts are paid out of pocket by the family, paid by health care coverage, or go unpaid.

*Total family charges for health care*—See total charges.

*Total out-of-pocket expenses*—This is the amount paid by a family for out-of-pocket expenses for health care services plus the amount of family-paid premiums.

*Urban-rural location*—Households were identified as located either within or outside of a Standard Metropolitan Statistical Area (SMSA). The definitions and titles of SMSA's are established by the U.S. Office of Management and Budget with the advice of the Federal Committee on Standard Metropolitan Statistical Areas. Households located inside SMSA's are furthei classified as being located within the SMSA's central city (called "central city") or not ("other"). Households located outside of SMSA's are classified as "urban" if they are located in (1) places of 2,500 inhabitants or more that are incorporated as cities, villages, boroughs (except Alaska), and towns (except in New England, New York, and Wisconsin), but excluding persons living in the rural portions of extended cities; (2) unincorporated places of 2,500 inhabitants or more; or (3) other territory, incorporated or unincorporated, included in urbanized areas. Otherwise, households located outside of SMSA's are classified as "rural."

*Work-loss day*—A work-loss day is a day on which a person did not work at his or her job or business because of a specific illness or injury. The number of days lost from work is determined only for persons 17 years of age and over who reported that at any time during the survey period they either worked at or had a job or business. Family work-loss days are the sum of all work-loss days of family members during the time they were in the survey, prorated to the time they were in the family.