SAFER • HEALTHIER • PEOPLE™

Monitoring the Nation's Health

Vital and Health Statistics

Series 2, Number 144

January 2008

# Statistical Match of the March 1996 Current Population Survey and the 1995 National Health Interview Survey

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

---

**Suggested citation**

Ingram DD, Moriarity CL, O'Hare JF, Turek J. Statistical match of the March 1996 Current Population Survey and the 1995 National Health Interview Survey. National Center for Health Statistics. Vital Health Stat 2(144). 2008.

---

**Library of Congress Cataloging-in-Publication Data**

Statistical match of the March 1996 Current population survey and the 1995 National Health Interview Survey / [by Deborah D. Ingram ... et al.].
p. ; cm.— (DHHS publication ; no. (PHS) 2007–1344) (Vital and health statistics. Series 2 ; no. 144)
"November 2007."
Includes bibliographical references and index.
ISBN 0–8406-0617–6 (alk. paper)
1. National Health Interview Survey (Hyattsville, Md.) 2. Current population survey, March 1996. 3. Health surveys—United States—Methodology. 4. Demographic surveys—United States—Methodology. 5. Statistical matching. 6. Demographic surveys—United States. I. Ingram, Deborah D. II. Series. III. Series: Vital and health statistics. Series 2, Data evaluation and methods research ; no. 144.
[DNLM: 1. National Health Interview Survey (Hyattsville, Md.) 2. Current population survey, March 1996. 3. Data Collection—methods—United States—Statistics. 4. Ethnic Groups—statistics & numerical data—United States. 5. Health Status Indicators—United States—Statistics. 6. Minority Groups—statistics & numerical data—United States. 7. Models, Statistical—United States. 8. Statistics—methods—United States. W2 A N148vb no.144 2007]
RA409.S6839 2007
614.4'273—dc22 2007035075

---

For sale by the U.S. Government Printing Office
Superintendent of Documents
Mail Stop: SSOP
Washington, DC 20402-9328
Printed on acid-free paper.

# Vital and Health Statistics

Series 2, Number 144

# Statistical Match of the March 1996 Current Population Survey and the 1995 National Health Interview Survey

Data Evaluation and Methods Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
January 2008
DHHS Publication No. (PHS) 2008-1344

## National Center for Health Statistics

Edward J. Sondik, Ph.D., *Director*

Jennifer H. Madans, Ph.D., *Acting Co-Deputy Director*

Michael H. Sadagursky, *Acting Co-Deputy Director*

Jennifer H. Madans, Ph.D., *Associate Director for Science*

Jennifer H. Madans, Ph.D., *Acting Associate Director for Planning, Budget, and Legislation*

Michael H. Sadagursky, *Associate Director for Management and Operations*

Lawrence H. Cox, Ph.D., *Associate Director for Research and Methodology*

Linda B. Torian, *Acting Director for Information Technology*

Linda B. Torian, *Acting Director for Information Services*

Linda T. Bilheimer, Ph.D., *Associate Director for Analysis and Epidemiology*

Charles J. Rothwell, M.S., *Director for Vital Statistics*

Jane E. Sisk, Ph.D., *Director for Health Care Statistics*

Jane F. Gentleman, Ph.D., *Director for Health Interview Statistics*

Clifford L. Johnson, M.S.P.H., *Director for Health and Nutrition Examination Surveys*

### Office of Analysis and Epidemiology

Linda Bilheimer, Ph.D., *Associate Director*

Diane Makuc, Dr.P.H., *Associate Director for Science*

Pauline Mendola, Ph.D., *Chief, Infant, Child, and Women's Health Studies Branch*

Christine Cox, M.A., *Chief, Special Projects Branch*

Amy Bernstein, Sc.D., *Chief, Analytic Studies Branch*

James Lubitz, M.A., *Acting Chief, Aging and Chronic Disease Statistics Branch*

Richard Klein, M.A., *Acting Chief, Health Promotion Statistics Branch*

# Acknowledgments

The authors gratefully acknowledge the assistance of the following people: Dr. Jacob Feldman and Dr. Diane Makuc provided thoughtful comments and insights throughout the project and reviewed the report. Dr. Jane Gentleman and Dr. Jennifer Madans reviewed the report. Linda Giannarelli, Urban Institute, provided information on use of statistically matched data in the Transfer Income Model.

The report was edited by Gail V. Johnson, CDC/CCHIS/NCHM/Division of Creative Services, Writer Editor Services Branch; typeset by Zung T. Le, CDC/CCHIS/NCHM/Division of Creative Services; and graphics produced by Odell Eldridge, CDC/CCHIS/NCHM/Division of Creative Services, Nova contractor.

# Contents

## Figures



## Text Tables



## Detailed Tables

## Appendix Tables

## Abstract

### Objectives

Statistical matching is a method used to combine two files when it is unlikely that individuals on one file are also on the other file. The objectives of this report are to document and evaluate statistical matches of the March 1996 Current Population Survey (CPS) and the 1995 National Health Interview Survey (NHIS) and give recommendations for improving future matches. The CPS-NHIS match was motivated by the need for a data set with data on health measures and family resources for use in policy analyses.

### Methods

Three statistical matches between the March 1996 CPS and the 1995 NHIS are described in this report. All three matches used person-level constrained matching with partitioning and a predictive mean matching algorithm to link records on the two files. For two of the matches, the CPS served as the Host file and the NHIS served as the Donor file; for the third match, the NHIS was the Host file and the CPS was the Donor file.

### Results

The results suggest that the constrained predictive mean matches of the March 1996 CPS and the 1995 NHIS successfully combined some of the information on the two files, but that relationships among some Host and Donor variables on the matched file may be distorted. The evaluation of the matches suggested that the variables used to partition the Host and Donor files prior to matching and the variables involved in the predictive mean matching play an important role in determining whether relationships among variables on the matched file correctly represent relationships among those variables in the population. The evaluation also indicated that estimates for small subgroups may be especially subject to error. The results reinforce the need to proceed cautiously when exploring relationships among Host and Donor variables on a statistically matched file.



**Keywords:** *Constrained matching • data fusion • predictive mean matching • statistical matching*


# Statistical Match of the March 1996 Current Population Survey and the 1995 National Health Interview Survey


*by Deborah D. Ingram Ph.D. and Christopher L. Moriarity, Ph.D., National Center for Health Statistics; John F. O'Hare, Ph.D., The Urban Institute; Joan Turek, Ph.D., Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services*


## Introduction

The objectives of this report are to describe the methods used for a statistical match of the March 1996 Current Population Survey (CPS) and the 1995 National Health Interview Survey (NHIS) (1–3), to present an evaluation of the match, and to provide recommendations for improving future matches. Statistical matching is a method used to combine two files when it is very unlikely that an individual is included on both files. This is known to be the case for the CPS and the NHIS. The motivation for the statistical match of the CPS and NHIS was the need to have measures of health status, health care utilization, and family resources (income sources including noncash benefits) on the same large national data set for health policy analyses. The goal of the statistical match of the March 1996 CPS and the 1995 NHIS was to assess the feasibility of using statistical matching to obtain a large national data file with health data and family resource data that could be used to make valid tabulations and inferences that involve both types of variables.

The CPS was selected as the source of information on family resources for two reasons. First, the CPS has detailed information on the income and demographic characteristics of the Nation. Second, as the primary data set used in the Urban Institute's Transfer Income Model (TRIM), the CPS has been a principal data source for examining how major governmental health, tax, and cash and in-kind transfer programs (and changes to them) affect the U.S. population at the individual, family, state, and national level (4). TRIM3, the current version of TRIM, is used to simulate numerous health, tax, and transfer programs. Health programs simulated include Medicare, Medicaid and SCHIP, and employer-sponsored health insurance. Tax programs simulated include payroll taxes, federal income taxes, and state income taxes. Cash and in-kind transfer programs simulated include: Supplemental Security Income (SSI), Temporary Assistance to Needy Families (TANF), Aid to Families with Dependent Children (AFDC), the Food Stamp Program, child care, child support, and public and subsidized housing.

Since the first TRIM model became operational in 1973, the TRIM models have been used to understand the potential outcomes of public policy changes such as welfare reform, tax reform, and national health care reform. Health-related variables are important to TRIM in at least two ways. First, various health-related variables help determine whether an individual or family is eligible for a government program, or the amount of benefit from that program. For example, disability status is particularly relevant to program

eligibility or benefits. Second, some health-related variables are important to include in the TRIM system to provide a more comprehensive picture of the economic well-being of American families. For example, the inclusion of health insurance premiums in the TRIM system allows for creation of a comprehensive measure of income (health costs and taxes subtracted, transfers added) that could be used in looking at how persons spend down into poverty. Other health-related variables may affect health, tax, or transfer program benefits. For instance, tax credits have been proposed to defray some of the costs of purchasing private nongroup health insurance policies. The CPS contains very little information on health-related indicators. Thus, matching the CPS and NHIS would allow important health-related variables on the NHIS, such as functional disability indicators and information on health conditions of interest, to be added to TRIM for use in simulations and tabulations. Additional health-related information benefits the TRIM system by enhancing the model's ability to simulate the various tax and transfer programs as well as possible new policies.

The NHIS was selected as the source of the health data as it contains detailed information on the health characteristics of the U.S. population. While the 1995 NHIS collected fairly detailed data on amounts of family income from various sources (on the Family Resources supplement), beginning with the 1997 NHIS the Family Resource supplement was discontinued and only data on amounts of total annual family income and annual personal earnings were collected. For some types of studies, data on income amounts from specific sources are needed. Therefore, the addition of reliable family resource variables to the NHIS through a statistical match with the CPS could be very useful.

The feasibility of combining income and program participation data from the CPS and health data from the NHIS using statistical matching was assessed in the Statistical Matching Project. The Statistical Matching Project was a collaborative effort among researchers at the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS), the U.S. Department of Health and Human Services' Office of the Assistant Secretary for Planning and Evaluation (ASPE), and the Urban Institute. An Advisory Group, which provided expert advice, consisted of the two project codirectors, Deborah Ingram (NCHS) and Joan Turek (ASPE); John O'Hare (Urban Institute); and the following researchers: Dale Hitchcock (ASPE), John Marcotte (Urban Institute), Chris Moriarity (NCHS), Gene Moyer (ASPE), Jim Scanlon (ASPE), Fritz Scheuren (Urban Institute), Sheila Zedlewski (Urban Institute), the late Pat Doyle (Census Bureau), and Carol Frost (Congressional Budget Office) (note that the affiliations shown are not necessarily current affiliations, but rather are the affiliation of each Advisory Board member when the Advisory Board was active). There have been several iterations of the statistical match of the March 1996 CPS and 1995 NHIS during this project, one of which has been briefly reported previously (5). An early assessment of the final statistical match, the match described in this report, has also been previously published (6).

This report begins with a brief review of statistical matching, followed by a comparison of the CPS and NHIS that assesses the similarity of the two data sets, and a description of the methodology used in this project. The following section presents the results of the match. The final sections contain recommendations that may improve future matches, some thoughts on future research directions, and a summary of the Statistical Matching Project.

# Overview of Statistical Matching

Statistical matching (also called synthetic, stochastic, or attribute matching or merging, data integration, or data fusion) involves combining two or more data files to construct one file. The purpose of statistical matching is to enrich an existing data file by adding more accurate, more detailed, or more comprehensive information to meet research, evaluation, or analytic needs. In a statistical match, it is very unlikely that an entity that is in one file is also in the other file, so the records on one file are matched with records on the other file that they resemble or are in some sense ''close to'' based on the values of a set of common variables. Statistical matching is in contrast to *exact matching* or *record linkage* where the same entities appear in both data files and unique identifiers exist to combine the records from the two data sources. Statistical matching has been widely used because it is faster and cheaper to construct a statistically matched data file than to conduct a new survey. Adding variables to a data file by statistically matching files also has the benefit of reduced respondent burden. Statistical matching may sometimes be used instead of exact matching because of legal restrictions on the use of exact matching due to confidentiality concerns. When a data file will be used by many analysts for many different purposes, producing a statistically matched file may be an efficient way to obtain the needed flexibility. The development of computers has made use of statistical matching procedures feasible. As computational power has increased, the size of data files that can be matched and the complexity of matching algorithms that can be used have also increased.

Statistical matching procedures developed in the United States and Canada, as well as evaluations and applications of them, have been published extensively in the economics and statistics literature (7–70). An early comprehensive review of the theoretical and applied approaches to statistical matching is contained in Sims (13). The Subcommittee on Matching Techniques of the Federal Committee on Statistical Methodology produced a comprehensive report containing an overview of statistical and exact matching techniques, examples of both types of matching, and a limited comparison of statistical and exact matching (36). The National Academy of Science's report

*Combining Information: Statistical Issues and Opportunities for Research* demonstrates how statistical matching relates to the broader topic of combining information across numerous data sources to assist in better decision-making (57). A number of other reviews of statistical matching procedures and evaluations of various approaches have appeared in the literature (17,23,30, 39,41,46,48,49,52,54,56,67). In the United States and Canada, statistical matching has been performed since the late 1960s by government agencies and research institutions—including the Bureau of Economic Analysis, the Brookings Institution, Mathematica Policy Research, the Office of Tax Analysis (U.S. Department of the Treasury), the Social Security Administration, Statistics Canada, and Yale University (7,8,10, 15–18,20–24,27,29,30,32,34,35,37, 38,40,43,45,47,51,55,59). The data files typically used in these statistical matches have been large national administrative record and economic microdata files containing information on a nationally representative sample of (or a major subset of) entities such as individuals, families, or firms. The resulting statistically matched data files have been used to make estimates of the distributions of economic variables (e.g., income, wealth, and taxes) and as input to microsimulation models that examine the impact of policy changes on population subgroups and projections of program needs. At least one early statistical match involved combining a national health data file with an economic data file to produce a data file for use in health services research (47). In recent years, statistical matching (usually referred to as data fusion in these applications) has been widely used in market research to produce cross tabulations of product usage data and media exposure data (61). Statistical matching has also been used extensively in Europe and Great Britain, but there it has been applied more often in market research, and different procedures have been favored (67).

## Standard Statistical Matching Framework

The standard matching framework is illustrated in Figure 1. In this framework, one has observations from two data sets (File A and File B). There is a limited set of variables common to both files (X-variables). File A also contains variables not available on File B (the Y-variables). Similarly, File B contains variables not available on File A (the Z-variables). A match of the two files results in at least one new data file (File C) in which each record contains information on all three sets of variables, that is, X, Y, and Z. If entities on one file are also on the other file, then File A and File B can be matched using exact matching procedures. An exact match involves linking each record on File A to the record on File B that has identical values for some selected set of the X-variables (the variables common to both files and in this case, usually unique identifiers). If, on the other hand, it is highly unlikely that any of the entities on one file are also on the other file, or if an exact match is not possible because the unique identifiers have been suppressed for confidentiality reasons, File A must be matched to File B using statistical matching procedures. A statistical match usually involves linking each record on File A to the record on File B with the most similar values on a selected subset of X-variables. In a statistical match, it is very unlikely that the pair of records that are matched (one from File A and the other from File B) will have identical values for all of the X-variables in the matching subset, and certainly not for the entire set of X-variables. In subsequent analyses using the matched file (File C), the values of the X-variables from the primary file are used. In the standard statistical matching framework, File A is considered the primary file and is referred to as the Host file, while File B is referred to as the Donor file. The resulting matched data file (File C), has records with values for the X and Y variables from File A and values for the Z-variables from File B. Often, the fact that File A and File B were combined using a statistical match, rather than an exact match, is ignored in subsequent usage of the matched file and it is treated as if it resulted from a single survey in which all three sets of variables were collected. The primary goals when performing a statistical match are to preserve on the matched file, to the maximum extent possible, the marginal distribution of each of the X-, Y-, and Z-variables as they appeared on the original data files, and to obtain joint distributions of the Y- and Z-variables that are reasonable estimates of the true joint distributions in the population.

## Unconstrained and Constrained Matching

There are two distinct types of statistical matching methods: *unconstrained* matching and *constrained* matching.

1. In *unconstrained* matching, each record in the Host file (File A) appears in the matched file (File C), but it is not required that all of the records in the Donor file (File B) be used in the match (10,27,36,46,50, 54,56,59,60,62,67). In unconstrained matching, each record on File A is matched with the record on File B that has the closest values on the subset of X-variables selected for use in the matching procedure. Typically in unconstrained matching, some records on File B will be used multiple times in the match, while other records will not be used at all. Limits sometimes are placed on the number of times a Donor record (File B record) can be used. When imposed, these limits help ensure that the (weighted) distributions of the Z-variables ''brought over'' to the Host file in the match are closely aligned with the distributions on the original file. Even so, one of the criticisms of unconstrained matching is that the marginal distributions of the Z-variables in the matched file can be quite different from those in the original file (41,46,56). Distortions of the Z and (X,Z) distributions

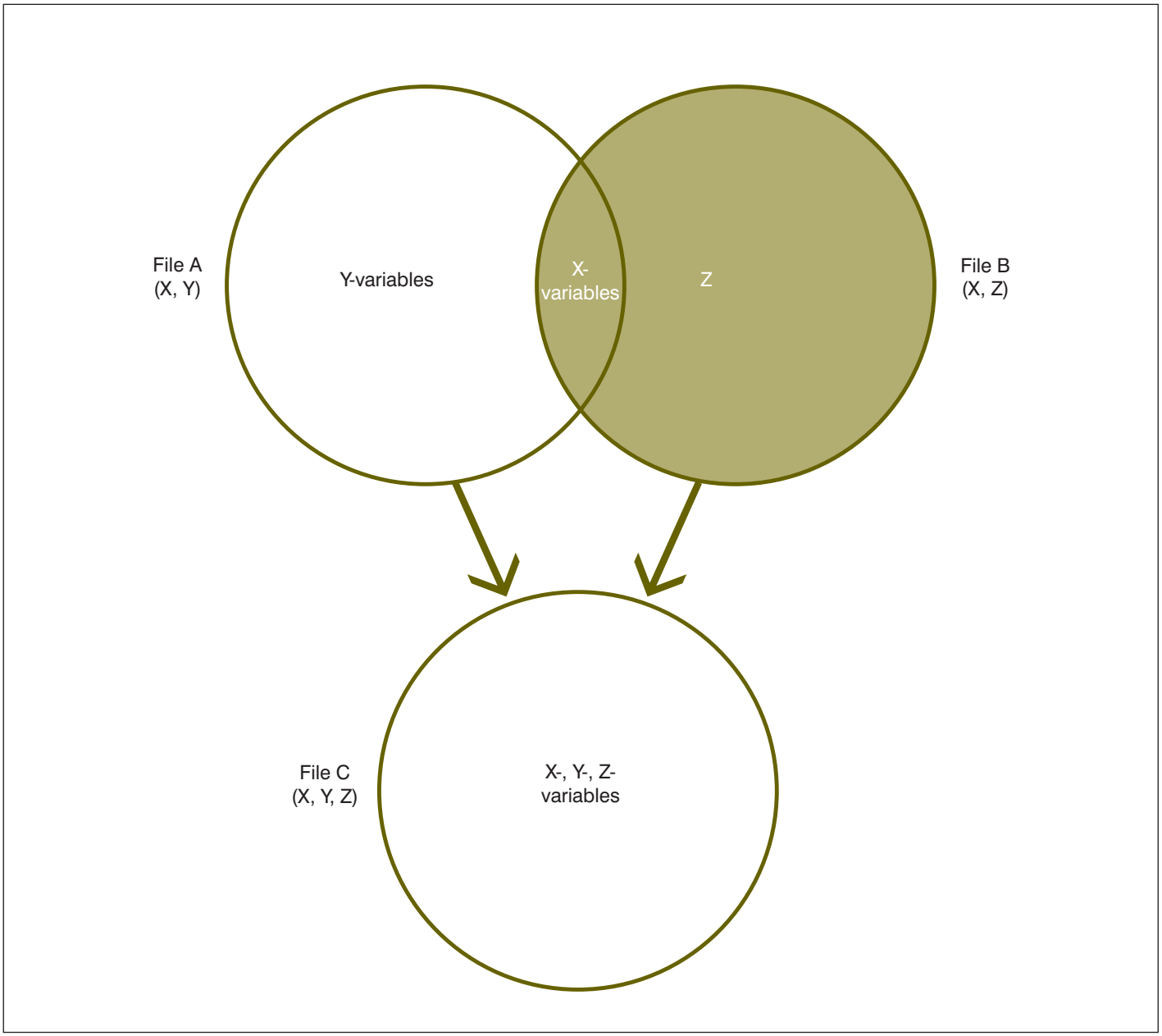
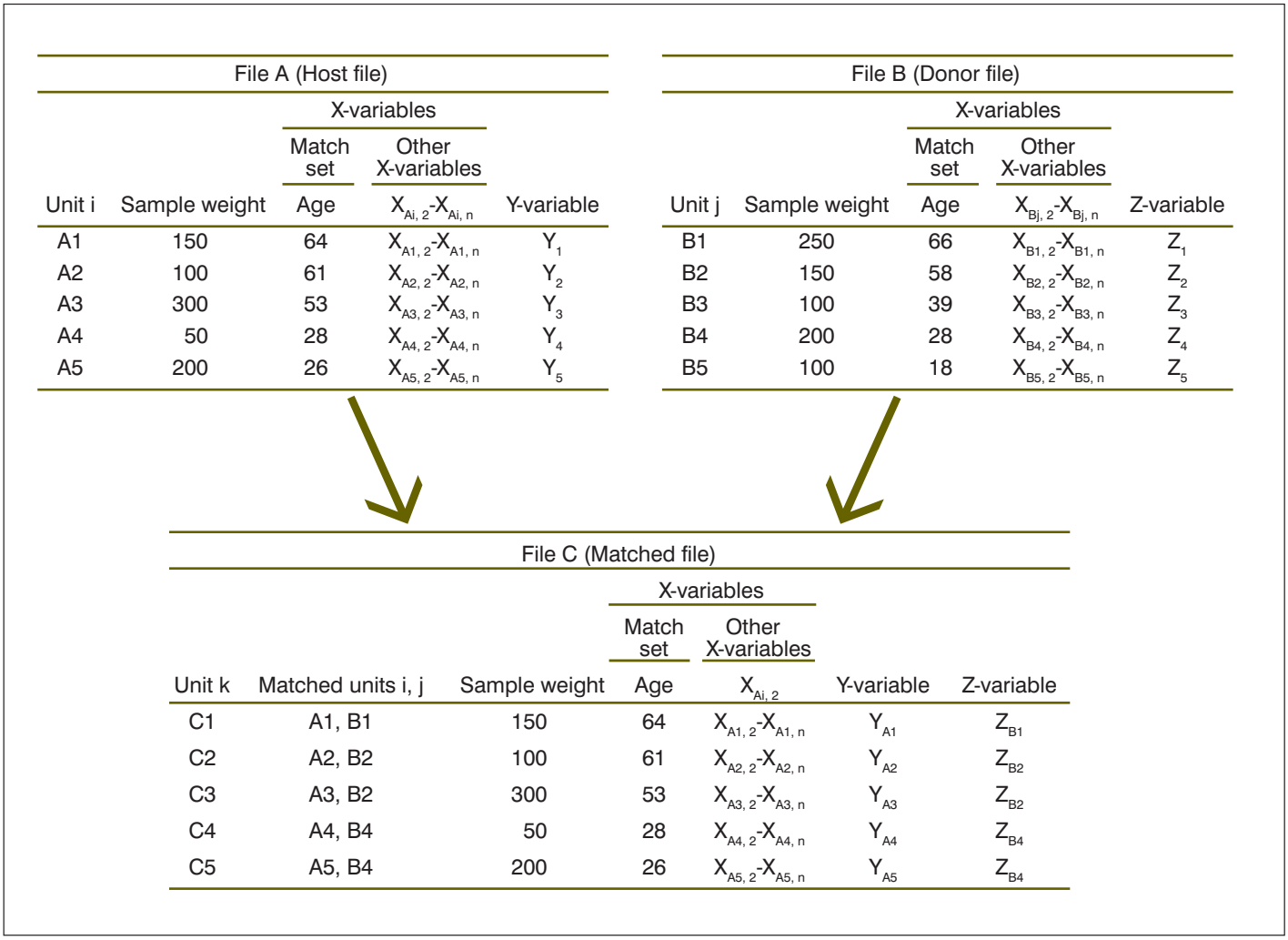



**Figure 1. Illustration of the standard statistical matching framework**

may occur partially because of highly different sample weights in the records (41). Rodgers reported that differences between the joint distributions of Z-variables in the matched file and their joint distributions in the Donor file tend to be greater when unconstrained matching is used than when constrained matching is used and thus, more error is introduced into regression models involving X-, Y-, and Z-variables (46).

To illustrate the unconstrained matching process, consider the hypothetical example shown in Figure 2. In this example, there are five units in the Host file (File A) and five units in the Donor file (File B), and the resulting matched file (File C) also has five units. The sample weights are not directly involved in the matching process; the sample weights associated with each of the units in the matched file are those found on the Host file (File A). In this example, only one variable from the set of X-variables (the variables common to both File A and File B), age, is used to assess "closeness" of File A and File B units; each File A unit is matched to the File B unit that is closest to it in age. Note that the sort order of the units within File A and File B does not affect which File A and File B units are matched. To begin the unconstrained matching process, the age difference between each File A unit and each File B unit is calculated. First, a match for A1, the first unit in File A, is chosen from the five File B units. B1 is closer in age to A1 than the other File B units are, so it is matched to A1 to form the first record in File C, C1 (A1 has age=64; B1 has



File A (Host file)

| Unit i | Sample weight | X-variables: Match set Age | X-variables: Other X-variables $X_{Ai,2}$-$X_{Ai,n}$ | Y-variable |
|---|---|---|---|---|
| A1 | 150 | 64 | $X_{A1,2}$-$X_{A1,n}$ | $Y_1$ |
| A2 | 100 | 61 | $X_{A2,2}$-$X_{A2,n}$ | $Y_2$ |
| A3 | 300 | 53 | $X_{A3,2}$-$X_{A3,n}$ | $Y_3$ |
| A4 | 50 | 28 | $X_{A4,2}$-$X_{A4,n}$ | $Y_4$ |
| A5 | 200 | 26 | $X_{A5,2}$-$X_{A5,n}$ | $Y_5$ |

File B (Donor file)

| Unit j | Sample weight | X-variables: Match set Age | X-variables: Other X-variables $X_{Bj,2}$-$X_{Bj,n}$ | Z-variable |
|---|---|---|---|---|
| B1 | 250 | 66 | $X_{B1,2}$-$X_{B1,n}$ | $Z_1$ |
| B2 | 150 | 58 | $X_{B2,2}$-$X_{B2,n}$ | $Z_2$ |
| B3 | 100 | 39 | $X_{B3,2}$-$X_{B3,n}$ | $Z_3$ |
| B4 | 200 | 28 | $X_{B4,2}$-$X_{B4,n}$ | $Z_4$ |
| B5 | 100 | 18 | $X_{B5,2}$-$X_{B5,n}$ | $Z_5$ |

File C (Matched file)

| Unit k | Matched units i, j | Sample weight | X-variables: Match set Age | X-variables: Other X-variables $X_{Ai,2}$ | Y-variable | Z-variable |
|---|---|---|---|---|---|---|
| C1 | A1, B1 | 150 | 64 | $X_{A1,2}$-$X_{A1,n}$ | $Y_{A1}$ | $Z_{B1}$ |
| C2 | A2, B2 | 100 | 61 | $X_{A2,2}$-$X_{A2,n}$ | $Y_{A2}$ | $Z_{B2}$ |
| C3 | A3, B2 | 300 | 53 | $X_{A3,2}$-$X_{A3,n}$ | $Y_{A3}$ | $Z_{B2}$ |
| C4 | A4, B4 | 50 | 28 | $X_{A4,2}$-$X_{A4,n}$ | $Y_{A4}$ | $Z_{B4}$ |
| C5 | A5, B4 | 200 | 26 | $X_{A5,2}$-$X_{A5,n}$ | $Y_{A5}$ | $Z_{B4}$ |

**Figure 2. Illustration of an unconstrained match of File A (Host file) and File B (Donor file) to produce a statistically matched file, File C**

age=66; C1 has age=64, the age of the Host unit). B2, the second unit in File B (with age=58) is closer in age to both the second and third units in File A (A2 with age=61 and A3 with age=53) than the other File B units, and therefore, B2 is matched to both of them. Similarly, B4 is the File B unit that is closest in age to A4 and is matched to it; B4 is the File B unit that is closest in age to A5 and is matched to it. Thus, two units in File B are used multiple times in the matching process, while two other units (B3 and B5) are not matched to any File A units.

2. In *constrained* matching, all of the records in *both* data files are represented in the matched file (File C) (24,35,39,41,44,46,48,50,59,60, 62,67). To accomplish this, records on *both* files may have to be used more than once because this type of matching involves making sure that the population weight attached to each record is ''used up'' in the match (when a record is used more than once its weight is ''split''). A necessary condition for performing a constrained match is that both input files have the same weighted population totals. A direct consequence of the constraints imposed on the weights is that the marginal distributions (and therefore, the means and variances) of the Y and Z variables in both input files are preserved on the matched file (File C). In applied work, it is often the case that the two input data files are from surveys taken over different time periods so that the weighted population totals are slightly different. In this case, it is common practice to ''scale'' one of the files (usually the Host file) so that the weighted population totals agree. When the input files are partitioned prior to marking (see below), the weights must be scaled within each partition cell. Such differential scaling can result in distortions of the marginal distributions of the Z-variables. When a complex survey design is present, the marginal distributions of the Z-variables may not be perfectly preserved on the matched file (67). One potential drawback of constrained matching is that records with an unacceptably large distance (defined in the section ''Choosing a close match'') between the X-variables may be matched. In addition, the number of records in a file created using constrained matching is usually larger

File A (Host file)

| | | X-variables | | |
|---|---|---|---|---|
| | | Match set | Other X-variables | |
| Unit i | Sample weight | Age | $X_{Ai,2}$-$X_{Ai,n}$ | Y-variable |
| A1 | 150 | 64 | $X_{A1,2}$-$X_{A1,n}$ | $Y_{A1}$ |
| A2 | 100 | 61 | $X_{A2,2}$-$X_{A2,n}$ | $Y_{A2}$ |
| A3 | 300 | 53 | $X_{A3,2}$-$X_{A3,n}$ | $Y_{A3}$ |
| A4 | 50 | 29 | $X_{A4,2}$-$X_{A4,n}$ | $Y_{A4}$ |
| A5 | 200 | 26 | $X_{A5,2}$-$X_{A5,n}$ | $Y_{A5}$ |

File B (Donor file)

| | | X-variables | | |
|---|---|---|---|---|
| | | Match set | Other X-variables | |
| Unit j | Sample weight | Age | $X_{Bj,2}$-$X_{Bj,n}$ | Z-variable |
| B1 | 250 | 66 | $X_{B1,2}$-$X_{B1,n}$ | $Z_{B1}$ |
| B2 | 150 | 58 | $X_{B2,2}$-$X_{B2,n}$ | $Z_{B2}$ |
| B3 | 100 | 39 | $X_{B3,2}$-$X_{B3,n}$ | $Z_{B3}$ |
| B4 | 200 | 28 | $X_{B4,2}$-$X_{B4,n}$ | $Z_{B4}$ |
| B5 | 100 | 18 | $X_{B5,2}$-$X_{B5,n}$ | $Z_{B5}$ |

File C (Matched file)

| | | | X-variables | | | |
|---|---|---|---|---|---|---|
| | | | Match set | Other X-variables | | |
| Unit k | Matched units i, j | Sample weight | Age | $X_{Ai,2}$-$X_{Ai,n}$ | Y-variable | Z-variable |
| C1 | A1, B1 | 150 | 64 | $X_{A1,2}$-$X_{A1,n}$ | $Y_{A1}$ | $Z_{B1}$ |
| C2 | A2, B1 | 100 | 61 | $X_{A2,2}$-$X_{A2,n}$ | $Y_{A2}$ | $Z_{B1}$ |
| C3 | A3, B2 | 150 | 53 | $X_{A3,2}$-$X_{A3,n}$ | $Y_{A3}$ | $Z_{B2}$ |
| C4 | A3, B3 | 100 | 53 | $X_{A3,2}$-$X_{A3,n}$ | $Y_{A3}$ | $Z_{B3}$ |
| C5 | A3, B4 | 50 | 53 | $X_{A3,2}$-$X_{A3,n}$ | $Y_{A3}$ | $Z_{B4}$ |
| C6 | A4, B4 | 50 | 29 | $X_{A4,2}$-$X_{A4,n}$ | $Y_{A4}$ | $Z_{B4}$ |
| C7 | A5, B4 | 100 | 26 | $X_{A5,2}$-$X_{A5,n}$ | $Y_{A5}$ | $Z_{B4}$ |
| C8 | A5, B5 | 100 | 26 | $X_{A5,2}$-$X_{A5,n}$ | $Y_{A5}$ | $Z_{B5}$ |

**Figure 3. Illustration of a constrained match of File A (Host file) and File B (Donor file) to produce a statistically matched file, File C**

(sometimes considerably larger) than the number in the Host file, which may be undesirable in some applications. Recently, Liu and Kovacevic have developed a complicated constrained matching procedure that utilizes an auxiliary data file and produces a matched file with a minimally inflated size (59,60,62).

Figure 3 illustrates the constrained matching process. In this hypothetical example, there are five units in the Host file (File A), five units in the Donor file (File B), and eight units in the resulting statistically matched file (File C). One X-variable from the set of common variables, age, is used to link units from File A with units from File B. Note that the sum of the sample weights for File A is identical to that for File B, so the weights do not have to be adjusted prior to matching. Prior to matching, both files are sorted on age, and from this point on, it is the rank order of the units and their sample weights that determine which File A and File B units are matched (70). The first unit in File C (C1) is obtained by matching the File A unit with a rank of 1, A1, with the File B unit with a rank of 1, B1. As A1 has a weight of 150 and B1 has a weight of 250, all of A1, but not all of B1, is ''used up'' in this match. To make the first match, B1 must be split into two records, one with a weight of 150 (to match the sample weight of A1) and one with a weight of 100. The second unit in the matched file, C2, is obtained by matching the File A unit with a rank of 2 with the remaining portion of B1. As A2 and the remaining portion of B1 both have a sample weight of 100, they are both used up in this match. The next match is between the next available File A unit, A3, and the next available File B unit, B2. In this match, B2, which has a sample weight of 150, is used up; but A3, which has a weight of 300, is not used up. Thus, the fourth match is between the remaining portion of

A3, which has a sample weight of 150 and the next available File B unit, B3, which has a sample weight of 100. Again the File B unit is used up, but A3 is not, a portion of A3 with a sample weight of 50 remains. Thus, the fifth match is between the remaining portion of A3 and the next available File B unit, B4. A3 is finally used up in this match, but B4 is not. Matching continues until all of the records in both files have been used up; this will always happen because the sum of the sample weights in the two files is identical. As the example illustrates, to accomplish the match, units on both the Host and Donor files must be split, with the result that the statistically matched file contains more units than the Host and Donor files do.

Unconstrained matching has been the more popular method because it is intuitive, relatively simple to implement, cost-effective, easy to replicate and update, and it makes fewer demands on system resources. However, a number of practitioners favor constrained matching because the risk of a poor match is lower with constrained matching and they believe that this outweighs the higher cost (39,41,48). With the advent of more powerful computers, cheaper memory, and faster numerical algorithms, the cost considerations have largely disappeared and constrained matching is used more often than it used to be.

# Choosing a "Close" Match

Regardless of whether an unconstrained or a constrained matching procedure is used, every Host record must be matched with a Donor record. The goal is to match each Host record with the Donor record that is similar to or in some sense "close" to it. The closeness of a pair of records is measured using one or more of the X-variables (those variables common to both files). The set of X-variables used to determine closeness has important consequences for the integrity of the matched data file (36,39,48,49,56). For example, it is recommended that more than one of the X-variables be used to match the Host and Donor records. Typically, some of the X-variables selected are demographic variables. Additionally, to help preserve the joint distributions of the Z-variables, some of the X-variables used for matching should be highly correlated with both the Y-variables and the Z-variables that will be involved in analyses performed using the matched file.

## Partitioning

One technique used to achieve a close match is *partitioning*. Partitioning, or blocking, involves dividing the records in both the Host and Donor files into mutually exclusive subgroups (cells) and permitting matches only between corresponding subgroups (i.e., only permitting a match between a Host and Donor record if their values for the X-variables used to define the subgroups place them in the same subgroup). Partitioning is used when matches between certain types of records should be avoided because the characteristics of the individuals are sufficiently dissimilar. For example, in the present context where the Z-variables include measures of health status and health care utilization, it would probably be unwise to allow a match between a man and a woman. Partitioning has the effect of narrowing the distance between records and allows for a "tighter" fit across the two data sets. However, when some of the cells in a partition must be merged because they are empty or have too few records (on either the Host or Donor files), poor matches may result. The variables used to define the partition should be variables deemed to be critical to the integrity of the match. Not only is the selection of the blocking variables important in partitioning, the order in which the variables are used in the partitioning scheme and the extent of the partitioning is also important. If cell sizes are too small, the sampling properties can be adversely affected. Higher minimum cell sizes should be used if the predictive power of the blocking variables is low, or if serious misalignment of the two files exists (two files are said to be misaligned when the weighted cell size of a partition cell on the Host file differs substantially from the weighted cell size of the corresponding cell on the Donor file). Large cell sizes increase the amount of computer time required and usually mean that the opportunity for deeper partitioning has not been realized.

## Distance Measures

Most statistical matches use a distance metric, such as the Mahalanobis or Euclidean distance functions, to assess how "close" two records are. A subset of the X-variables (those variables common to both the Host and Donor files) must be selected for inclusion in the distance metric (10,17,38,39,41,46,49,54,67). Typically, when more than one X-variable is involved in the distance metric, weights are assigned to the X-variables. These weights are determined subjectively or through modeling (e.g., regression models). With unconstrained matching, each Host record is matched to the record in the Donor file that is its "nearest neighbor" as measured by the metric. With constrained matching, each Host and each Donor record are assigned ranks based on the distance metric. The pairing of Host and Donor records is guided by their ranks and by their sample weights because the weights attached to each record must be "used up." Consequently, a Host record may not be matched to the Donor record that is its "nearest neighbor" (based on the values of the distance metric or X-variables) because that Donor record may already be matched to another Host record.

## Predictive Mean Matching

Another method used to match the most similar records is predictive mean matching. This term was coined by Little and usually refers to an imputation procedure for partially missing data within one survey data file (71). With predictive mean matching, a variable that is available on either the Host or Donor file (but not both) is identified for use as the dependent variable. Usually, the variable selected to be the dependent

variable is an important variable on the Donor file and is expected to be either an important variable in subsequent analyses of the matched file or highly correlated with variables that will be used in subsequent analyses. A regression is carried out using the selected dependent variable and a subset of the variables common to both files (the X-variables) as independent variables. Predicted values of the dependent variable are calculated for both the Host and Donor files using the regression coefficients and each file's respective X-variable values. Records from the Donor file are matched to records from the Host file using the predicted values. With unconstrained matching each Host record typically is matched to the Donor record with the most similar predicted value. With constrained matching, the Host and Donor files are sorted by the predicted variable, after which rank and sample weight drives the matching process as illustrated in Figure 3. Presumably, records with similar predicted values will have similar ranks and be matched, but it is likely that some Host records will not be matched to the Donor record with the closest predicted value. While the X-variables in the predictive mean matching regression can be thought of as playing the same sort of role as the variables in a classic distance metric, predictive mean matching differs from the distance metric approach because it involves Y or Z-variables as well as X-variables. Predictive mean matching is considered by some to have performed well in practice (65).

## Conditional Independence Assumption

In traditional statistical matching procedures, information on the Y-variables is completely ignored and all of the information about the Z-variables (the variables that are being brought over to the matched file from the Donor file) is communicated via the X-variables (the variables common to both File A and File B). As a result, an implicit assumption of these procedures is that the Y- and Z-variables are independent (or uncorrelated if normality is assumed) conditional on the X-variables. In other words, the relationships of the Y- and Z-variables can be completely inferred from the relationships of the Y- and X-variables and the Z- and X-variables:

$$P(Y,Z|X) = P(Y|X)\,P(Z|X).$$

This assumption is referred to as the conditional independence assumption (CIA). As has been extensively discussed in the literature, if the CIA does not hold, then estimates of (and inferences about) the Y-Z distributions (in the matched files) will be biased to a greater or lesser extent, which may lead to erroneous conclusions (13,14,19, 23,25,28,29,31,36,37,39,41,46, 49,50,52, 54,56,57,63). The extent to which the CIA is violated cannot be tested, nor can the resulting bias be estimated, because neither File A nor File B contains any information about the joint distributions of the Y- and Z-variables. This leads to uncertainty about inferences based on the matched file. Simulations and empirical studies have shown that bias resulting from violations of the CIA can be a problem (23,39,41,46,49,52, 54,56).

## Use of Auxiliary Information, Multiple Imputation, and Alternatives to the CIA

Paass suggested that auxiliary information about the Y-Z relationships be used in statistical matching as an alternative to reliance on the CIA (48,49). This approach requires a third data file, File AUX, with information on either (X,Y,Z) or (Y,Z). The basic idea is to add Z values to records on File A using information obtained from File A, File B, and File AUX on the joint distributions of the X-, Y-, and Z-variables. The auxiliary information can come from outdated data files, other samples, frequency tables, or regression equations. Paass described and developed a number of parametric and nonparametric approaches for the use of auxiliary information. Singh et al. proposed nonparametric and parametric methods, based on a log linear imputation method that extended Paass's work and work by Rubin (53,57). The methods involve using auxiliary information to impose categorical constraints on the matched file. Liu and Kovacevic extended these methods (59,60,62). Unfortunately, while empirical studies indicate that use of auxiliary information can improve the quality of a matched file, auxiliary data files with sufficient information on both the Y and Z variables generally are not available (5).

Another group of statistical matching procedures that have been developed to address the CIA problem are those involving the creation of multiple files corresponding to various assumptions about the unknowable Y,Z correlation. Kadane proposed assessing the potential impact of violations of the CIA on a statistically matched file by constructing numerous matched files using different estimates of the covariance matrix of the Y- and Z-variables so that the sensitivity of the results to nonzero values of the covariance of Y and Z can be explored (27,28,31). Rubin proposed a method he called ''file concatenation with adjusted weights and multiple imputation'' that involves concatenating File A and File B and then multiply imputing the missing values of Z-variables for records from File A and the missing values of Y-variables for records from File B. Rubin proposed that multiple matched files be obtained under the same model assumptions (e.g., that the partial association of the Y- and Z-variables is zero) to directly assess uncertainty due to sampling and also under different modeling assumptions (e.g., nonzero values for the partial association of the Y- and Z-variables) to directly assess sensitivity of the matched files to violations of the CIA (37,50). Moriarity and Scheuren have extended the work of Kadane and Rubin (66,68,69). Kamakura and Wedel also have extended Rubin's work; their mixture model approach was developed for use with categorical variables (61).

## Criticisms of Statistical Matching

Criticisms of statistical matching in general and of particular statistical matching procedures have been expressed by many (9,12,13,14,16–19, 24–27,29,31,36,37,39,41,46,48–50,52, 54,56,57,63,65,67). The primary criticism of most statistical matching methods is that their validity relies on the CIA, which is considered to be an unrealistic and untestable assumption (13,19,28,29,36,39,41,46,49,52,57, 63,65). Most of the improvements made to statistical matching procedures over the years have not addressed the CIA limitation; the CIA remains a strong implicit assumption of the procedures, with violations of it resulting in biased estimates. Only when information on the Y-Z relationships is available from an auxiliary data source can direct checks on estimates of the relationships among the Y- and Z-variables be made, and information about those relationships incorporated in the match. Statistical matching procedures that explore alternative assumptions to the CIA, although not providing direct measures of the accuracy of estimates of the Y-Z relationships, do provide an assessment of the sensitivity of those estimates to violations of the CIA. Simulations and empirical studies have shown that estimates of the relationships among Y- and Z-variables can be poor, and thus, inferences based on a statistically matched file can be risky (23,39,41,46, 49,52,54,56). Some of the weaknesses found by the validation results described later in this report are quite predictable given the literature on this topic.

A recurring criticism of statistical matching is that often the standard errors used to make inferences are not valid. There is a tendency to treat statistically matched files as though the set of variables on each record (the X-, Y-, and Z-variables) were collected from the same entity. However, this practice is incorrect because it is highly unlikely that an entity on File A is also on File B and thus, the matching procedure can never create a file that has the true Z-variable values for the File A records. Standard errors computed as though the matched file has always been a complete data set do not include uncertainty due to sampling variation and to matching and so are underestimated. Some of the uncertainty that must be incorporated in the standard errors for a matched file arises from variability in the population that is not captured in the matched file. Specifically, for a given set of values for the X-variables used for matching, multiple sets of values for the nonmatch variables (the other X-variables and the Y- and Z-variables) can be found in the population from which File A and File B were sampled. The File A entity with the given set of values for the X-variables used for matching is matched to only one File B record (in an unconstrained match, and possibly more than one in a constrained match), so the variability that exists in the population is not captured in the matched file. Error also is introduced into the matched file when a File B record matched to a File A record has different values for the X-variables used for matching than the File A record, and this error should be included in the standard error. Note that this particular type of error is not uniform across the matched file; it tends to be larger in sparse regions of X-variable distributions. One approach to obtain valid standard errors is to use one of the matching procedures that incorporate alternative assumptions to the CIA.

Another criticism of statistical matching is the heuristic nature of most statistical matching procedures. Many subjective decisions must be made throughout the matching process, some of which critically affect the quality of the matched file. For example, variables must be selected from the set of X-variables for use in any partitioning scheme and categories must be specified for each one; a distance metric or a model must be selected for use in the matching of the Host and Donor records and the variables to be included in the metric or model must be selected.

The theoretical framework on which statistical matching procedures are based is relatively undeveloped. The implicit and explicit assumptions of a given procedure are not always clearly identified and may make it difficult to evaluate the properties of the procedure. As a result, there is no consensus about which statistical matching procedures are best. There are no tests that can be used to determine whether two data files can be successfully matched, or to assess a matched file to determine whether it is statistically equivalent to a sample of (X,Y,Z) randomly drawn from the population of interest. Additionally, there is no way to empirically assess what types of analyses can be appropriately performed using a particular statistically matched file. Given the paucity of the theoretical framework and the impossibility of knowing the true distributions of (X,Y,Z) without an adequate auxiliary data file, a number of researchers have turned to empirical studies and simulations to evaluate the various approaches (23,39,41,46,49,52,54,56).

Beyond theoretical considerations, experience suggests that statistical matching is more tractable when the input data files are similar with respect to sample size, the population of interest, the time period over which the surveys are taken, sample stratification, weighting, and the type of questions that are asked.

# Description of Input Data Files

The original design of this project specified that a statistical match be performed with the 1995 National Health Interview Survey (NHIS) and either the Survey of Income and Program Participation (SIPP) or the March 1996 Current Population Survey (CPS). Because statistical matching works best when the data sets being matched are similar in design (this similarity relates to the sample frame, population of interest, size of the survey, and sampling methods), the similarities and differences among the designs of the three surveys were considered. The Advisory Group decided to match the NHIS and the CPS. Among the reasons the CPS was chosen rather than the SIPP were:

- Similarity: The CPS (March supplement) and NHIS have similar sample designs.
- File size: The CPS is similar in size to the NHIS, while the SIPP is a much smaller file than the NHIS. It is preferable for the two files being matched to be similar in size.
- Attrition: Because the SIPP interviews are conducted over multiple years, some information is missing for individuals who leave the household during that period.
- Complexity: The SIPP file structure is far more complex than the CPS given SIPP's longitudinal nature and the construction of the panel weights and this would result in more time being devoted to alignment. (Two files are aligned when the cell sizes of the partition cells of one file are the same as, or proportional to, the cell sizes of the corresponding partition cells of the other file).
- Uses of the file: As an important use of the resulting matched data file is microsimulation, the CPS is the natural choice because many existing microsimulation models use the CPS to obtain information on annual incomes and program participation. For example, ASPE uses the microsimulation model, Transfer Income Model (TRIM), with the CPS to calculate the effect of large-scale government health, tax, and transfer programs (e.g., welfare reform, tax reform, national health care reform) (4).

## 1996 March Current Population Survey

The Current Population Survey (CPS) is the source of official Government statistics on employment and unemployment, and is the Nation's primary source of information on characteristics of the general labor force and of the population as a whole. It is a monthly survey sponsored by the U.S. Bureau of Labor Statistics and conducted by the U.S. Census Bureau (1,2). The CPS uses a multistage probability sample design. The CPS sample is based on the civilian noninstitutionalized population of the United States. Data are reported for households, families, and individuals. In the CPS, any household member 15 years of age and over is eligible to act as a respondent for the household. Whenever possible, labor force characteristic data are collected from each eligible individual in the household; however, any knowledgeable adult in the household can provide the information.

The March CPS, referred to historically as the Annual Demographic File and now as the Annual Social and Economic Supplement, contains the basic demographic and labor force characteristics data collected at each monthly CPS, plus additional information on work experience and income sources (including noncash benefits such as food stamps, health insurance, Medicaid, and Medicare). Income and employment information are collected for all individuals in the household who are 15 years of age and over. The income information relates to all income received in the prior calendar year.

In the March 1996 survey, approximately 50,000 households, or 130,476 individuals with a weighted total of about 264 million people, were interviewed. The March 1996 CPS included a supplemental sample of about 3,000 Hispanic households identified in the previous November's basic CPS sample (for a total of almost 7,000 Hispanic households). If a person was identified as being of Hispanic origin from the November 1995 interview and was still residing at the same address in March 1996, the housing unit was eligible for the March survey. Post-stratification was carried out across age-race-Hispanic origin-sex groups.

## 1995 National Health Interview Survey

The National Health Interview Survey (NHIS) is the Nation's primary source of general health information for the civilian noninstitutionalized population of the United States. In addition to demographic and socioeconomic information, the NHIS collects information on a broad range of health topics including illnesses, injuries, activity limitations, chronic conditions, health insurance coverage, and utilization of health care. The NHIS is a continuous survey sponsored by NCHS and conducted by the U.S. Census Bureau (3). The NHIS has a stratified multistage probability design. Data are reported for households, families, and individuals. In the 1995 NHIS, amounts for wages and salaries and self-employment income were collected for individuals 18 years of age and over. Amounts for other income sources were collected for all persons in the family. Any knowledgeable adult in the household can serve as the respondent.

The 1995 NHIS has information on approximately 40,000 households, or 102,467 individuals, with a weighted total representative of about 261 million people (3,72). Both black and Hispanic households were oversampled in the 1995 NHIS. Post-stratification was carried out across the same age-race-Hispanic origin-sex groups used for the CPS post-stratification.

## Comparison of the Current Population Survey and the National Health Interview Survey

The CPS oversampled Hispanic persons while the NHIS oversampled both black and Hispanic persons. Differences in the unweighted percentages by sex, race, and Hispanic origin from the CPS and NHIS reflect the differences resulting from the oversampling (Table 1). Because the CPS and NHIS samples were post-stratified across the same age-race-Hispanic origin-sex groups, weighted counts for various subgroups are nearly identical (Table 2). The CPS and NHIS weighted totals differ somewhat (the CPS weighted counts are higher) because the CPS is post-stratified using 1996 population estimates while the NHIS is post-stratified using 1995 population estimates. A more detailed comparison of the files is provided in "Appendix I."

The time period covered by the two surveys differs. Data for the March 1996

CPS were collected during March 1996 and provide a ''snapshot'' of household and family composition in March 1996, with information on income and earnings asked retrospectively for the prior *calendar* year, e.g., January 1995–December 1995. The NHIS, on the other hand, is an ongoing survey, and data for the 1995 NHIS were collected over the course of the entire year. As questions on the 1995 NHIS ask about events in the past 2 weeks, the past month, or the past 12 months, the time period spanned by the 1995 NHIS is January 1994–December 1995.

The recall periods of the questions on the two surveys tend to differ. Questions on the 1996 CPS generally ask respondents for information about circumstances and events over a 12-month period. This is in contrast to the 1995 NHIS, on which many of the questions are subject to a much shorter recall period. It is likely that answers to certain questions are more accurate with this shorter recall period, but this creates issues with respect to comparability across the two data files for some variables, e.g., health insurance coverage.

No items on the CPS public-use files have missing data as all missing data have been replaced by imputed values. Some items on the NHIS public-use files have missing data. For example, about 10% of 1995 NHIS respondents are missing some or all health insurance coverage information and about 1% are missing respondent-assessed health status. So that all NHIS respondents would be available for the match, the missing values on the NHIS for the variables involved in the matching (specifically, in the partitioning and the predictive mean match regressions) were imputed prior to the match. For some items, imputed values were deterministically assigned to the missing values. For other items, the missing values were replaced using hot deck imputation. This will be discussed more fully in the ''Statistical Match Methodology'' section.

# Statistical Match Methodology

Three statistically matched data files were constructed, two with the CPS as Host and one with the NHIS as Host. All of the matches described in this report are person-level, constrained matches with partitioning of both the Host and Donor files using key variables, and a predictive mean matching algorithm used to link Host and Donor records. Initially just two matches were performed, one match with the CPS as Host and number of doctor visits (from the NHIS) used as the dependent variable in the predictive mean matching algorithm, and a second match with the NHIS as Host and total annual family income (from the CPS) used as the dependent variable in the predictive mean matching algorithm. Because of the low $R^2$ for the predictive mean match regression with number of doctor visits as the dependent variable (performed on the NHIS-Donor file), a second match with the CPS used as the Host file was performed. The second CPS-Host match used total annual family income (from the CPS) as the dependent variable in the predictive mean matching algorithm. Multiple matched files were constructed to maximize the usefulness of the resulting matched files and to provide a check on the validity of the matching methodology.

Early in the project, statistical matches were performed using unconstrained matching with partitioning and a Euclidean distance function. The distributions of the Z-variables in the resulting matched files were distorted, apparently because over 15% of the Donor records were not used in the match and whether or not they were used appeared to be nonrandom. Dr. Fritz Scheuren recommended that constrained matching with partitioning and a predictive mean matching algorithm be used instead. Therefore, this approach was adopted to produce the final matched files.

In addition to person-level variables, family-level variables were used in the file partitioning and predictive mean matching to ensure that the match was sensitive to family characteristics that may influence family income, access to health care, and health care utilization. One member of each family was designated to be the head of the family (referred to as head of family). Usually the member of a family who was the ''reference person'' (''householder'' on the CPS, ''reference person'' on the NHIS, see ''Appendix II'' for definitions) was designated to be the head of that family. However, if the reference person was not in the labor force, another adult family member who was in the labor force was designated to be the head of family. There are some small differences in the way families are defined on the March 1996 CPS and the 1995 NHIS. Because of the key role of the family-level variables in the statistical matching, it was important to make the CPS and NHIS families as equivalent as possible. Therefore, some modifications were made to the family groupings on both the CPS and the NHIS prior to matching: These modifications are summarized in the following text.

1. The 1995 NHIS did not collect any information for members of a family who were in the military; the March 1996 CPS did. Thus, for the 1995 NHIS, if the mother or father of a family was in the military, there is no record on the data file for this parent. If both parents of a family were in the military, there are no records on the 1995 NHIS data file for either parent. As a result, for a small number of families in the NHIS, some of the family-level variables were initially missing or ''incorrect'' because no records appeared on the file for one or both parents. To remedy this situation, families with one or both parents in the military were identified using information from the family relationship variables. Values for a limited number of variables were imputed for the

missing military adult(s) using deterministic rules. For example, for a two parent family with one parent in the military, the sex and marital status of the missing military adult was derived from that of the nonmilitary spouse. The age imputed for the military adult in this two-parent family was the age of his or her spouse. All of the military adults were assigned the educational level of the head of family, mid-level occupational pay category, and full-time employment status (35 or more hours per week). Records were not added to the file for these persons, rather their demographic and employment characteristics were considered when deriving head-of-family variables.

2. For both the 1996 CPS and the 1995 NHIS, all foster children in a household were assigned to the primary family of the household. A family is considered to be the ''primary family'' of a household if the reference person is a member of that family. See Appendix II.
3. In both the CPS and NHIS, a small number of families had no adult member (no person aged 18 years or over). For all of these families, there was at least one family member who was 16 or 17 years of age. One of these persons was designated to be the reference person and head of family.

Family-level characteristics and head-of-family characteristics were derived after these adjustments were made to family groupings and attached to the person records in each data file.

# Key Variables in the Statistical Match

## Total Annual Family Income

Total annual family income was used in the predictive mean matching scheme. While information on total annual family income is available on both the CPS and NHIS, it is measured differently in the two surveys. The March 1996 CPS collected information on income from specific detailed sources using multiple questions about income received during the previous calendar year from these various sources. The total annual family income variable available on the March 1996 CPS public-use data file is derived by the Census Bureau by summing income reported for specific income sources. The resulting total annual family income variable is continuous and for the March 1996 CPS has values ranging from –$9,999 to $713,797. There are only two questions on the core questionnaire of the 1995 NHIS about total annual family income. Both questions asked respondents to report income during the prior 12 months (rather than the previous calendar year as on the CPS). In one question, respondents were asked to indicate whether their income during the prior 12 months was above or below $20,000 per year. In the other question, they were asked to indicate into which of 27 categories their family income fell during the previous 12 months, with the lowest category being less than $1,000 and the highest category being $50,000 and above. For the statistical match, the missing total annual family income values on the NHIS public-use files were replaced with imputed values (73). Thus, unlike the CPS, only categorized total annual family income was available on the 1995 NHIS, and the upper income values were capped at $50,000. The distribution of categorized annual family income for the March 1996 CPS and the 1995 NHIS respondents is shown in Table 3. About 35% of the respondents on the CPS and 25% on the NHIS reported having annual family incomes of $50,000 or more. This difference does not reflect sampling bias; rather it reflects the way income data were collected in the two surveys.

## Health Insurance Coverage

Health insurance coverage was used to partition the two files and as a covariate in the predictive mean match regression models. For both of these purposes, three categories of health insurance coverage were used:

1. Not covered by health insurance
2. Covered by public insurance
3. Covered by private insurance or Medicare

Individuals with military health insurance were placed in the public coverage category. Individuals who reported having both public and private coverage were assigned to the private coverage category. Individuals who reported having Medicare coverage were placed in the private coverage category because Medicare coverage is not restricted to the low income population.

Both the CPS and NHIS include numerous questions about each family member's health insurance coverage. However, the questions used to ascertain health insurance coverage differ in the two surveys. Most importantly, the reference periods of the questions differ in the two surveys, with the 1996 CPS asking about coverage during the previous calendar year and the 1995 NHIS asking about coverage during the prior month. Thus, the uninsured group on the NHIS public-use file is more inclusive than the uninsured group on the CPS public-use file: NHIS respondents who reportedly had no health insurance during the month prior to their NHIS interview are identified as uninsured, regardless of their insurance status during the other 11 months of the prior 12 months; whereas, only CPS respondents who reportedly had no health insurance during all 12 months of 1995 are identified as uninsured.

There is no missing health insurance coverage information on the CPS public-use data file; any information that was not collected has been replaced on the public-use data file with imputed data. However, for the NHIS, missing health insurance data were not imputed and are missing on the public-use file. About 10% of the records on the 1995 NHIS are missing information for some or all of the health insurance questions. As health insurance coverage was a key variable in the statistical match, individuals with missing health insurance data could not be included in the match. Performing the match with health insurance data missing for some of the NHIS respondents caused problems (e.g., an unacceptably large percentage of the

Donor records were not matched to Host records in the unconstrained match), so the missing health insurance data on the NHIS were imputed for the final matches. Some of the missing data were imputed using deterministic imputation; most were imputed using a hot deck imputation (“Appendix III”).

Using the imputed health insurance data for the NHIS, Tables 4 and 5 compare health insurance coverage for CPS and NHIS respondents by respondent-assessed health status (excellent, very good, or good health and fair or poor health) and sex. Table 4 provides unweighted estimates; Table 5 provides weighted estimates. Generally, the percentage of persons reporting no health insurance coverage was slightly lower on the CPS than on the NHIS. However, for persons reporting good-to-excellent health, the weighted percentage with no health insurance was slightly higher on the CPS than on the NHIS. These findings for persons reporting no health insurance coverage are somewhat at odds with the expected results. Because of the different reference periods for the insurance questions on the two surveys, we expected that a substantially lower percentage of CPS respondents would report that they were uninsured compared with NHIS respondents. The percentage of persons reporting public health insurance coverage was slightly lower on the CPS than on the NHIS, while generally, the percentage reporting private health insurance or Medicare coverage was either about the same on the two surveys or slightly higher on the CPS than on the NHIS.

## Respondent-Assessed Health Status

Respondent-assessed health status was used both in the partitioning and the predictive mean match regressions. While there were no missing data for health status on the CPS public-use data file (on the CPS missing values are routinely replaced with imputed values), 1,190 NHIS respondents have missing health status. Their health status was assumed to be fair or poor.

## Partitioning

An extensive partitioning scheme was imposed on the CPS and NHIS data files. The eight variables selected as the blocking variables in the partitioning (sex, health insurance coverage, respondent-assessed health status, age, race-Hispanic origin, region of residence, urbanization level of residence, and educational attainment of the head of the family) have the following characteristics:

1. They are important predictors of income in the CPS or health status and health care utilization in the NHIS.
2. They are defined similarly in both surveys.
3. They are subject to little measurement error.

The eight blocking variables were applied in the order specified in Table A. The goal was to reduce CPS cell sizes to less than 250 and NHIS cell sizes to less than 200 but to maintain a minimum cell size of 30. As illustrated in Tables 4 and 5, which provide the unweighted and weighted cell sizes after partitioning on the first three blocking variables, there is considerable variability in cell sizes. As a result, partitioning was deeper for some groups (e.g., individuals with private health insurance who were in excellent or good health) than for others (e.g., individuals with no health insurance who were in fair or poor health).

The files initially were partitioned using the first five blocking variables (sex, health insurance coverage, respondent-assessed health status, age,

**Table A. Blocking variables used in the partitioning of the March 1996 Current Population Survey and 1995 National Health Interview Survey, listed in order of application**

| Blocking variable | Categories |
|---|---|
| Sex | Male<br>Female |
| Health insurance coverage | Private or Medicare<br>Public<br>No coverage |
| Respondent-assessed health status | Excellent, very good, or good/Fair or poor |
| Age | Eight age groups:<br>0–4 years<br>5–17 years<br>18–26 years<br>27–36 years<br>37–46 years<br>47–56 years<br>57–63 years<br>64 years and over |
| Race and Hispanic origin | White, not Hispanic<br>Black, not Hispanic<br>Other, not Hispanic<br>Hispanic |
| Region of residence | Northeast<br>Midwest<br>SouthWest |
| Urbanization level of residence | Large metro, central[1]<br>Large metro, suburban[2]<br>Medium or small metro[3]<br>Nonmetro[4] |
| Education of head of family | Less than high school<br>High school<br>Some college<br>College or more |

[1]A central county of a metropolitan area with a population of 1 million or more.
[2]A suburban county of a metropolitan area with a population of 1 million or more.
[3]A county in a metropolitan area of less than 1 million.
[4]A nonmetropolitan county.

and race and Hispanic origin). The size of some of the resulting cells was too small (less than 30 respondents). Therefore, for certain combinations of the blocking variables, the partitioning could not be completed using the original categorization of all five variables and some subgroups had to be combined to maintain the minimum cell size of 30. For example, for males with public health insurance coverage who reported being in fair/poor health, the 18–26 year age group was combined with the 27–36 year age group, and the 37–46 year age group was combined with the 46–57 year age group. Neither of the two resulting subgroups was partitioned further. For a few combinations of sex, health insurance, and health status the 0–4 year age group was combined with the 5–17 year age group, and the 57–63 year age group was combined with the 64 years and over age group. A number of the subgroups defined by the first four blocking variables (sex, health insurance, health status, and age) could not be partitioned further on the race and Hispanic origin variable, or could be partitioned on race and Hispanic origin only if some of the race and Hispanic origin categories were collapsed. For example, in some cases the non-Hispanic white, non-Hispanic black, and non-Hispanic other races categories were combined so the race and Hispanic origin grouping was reduced to two categories, non-Hispanic and Hispanic.

After partitioning using the first five blocking variables was completed, any remaining large cells (based on unweighted counts: CPS greater than or equal to 250 persons, NHIS greater than or equal to 200 persons) were partitioned further using one or more of the remaining three blocking variables (region, urbanization level, and education of head of family). The final partition structure consisted of 982 cells, 99% of which contained between 30 and 250 persons. Table 6 shows the frequency distributions of the cell sizes of the partitioning for the CPS and the NHIS.

## Regression on the Current Population Survey Total Annual Family Income

A weighted least squares regression model was fitted to the continuous variable *total annual family income* on the CPS. The independent variables selected for use in the model are strong predictors of annual family income as measured by standard statistical tests. Because the model is used to derive predicted values of annual family income for both the CPS and NHIS respondents, all of the independent variables must be X-variables (variables common to both files). To ensure that all members of a family would have the same predicted annual family income, only family-level variables and head of family variables were included in the model (see Table B for a list of the independent variables).

Categorized total annual family income is, of course, a very strong predictor of continuous annual family income. As categorized total annual family income is available on the NHIS and can be constructed from the continuous total annual family income variable on the CPS, categorized total family income was included in the regression models. Twelve income categories were used for the regressions: five categories for incomes less than $20,000, six categories for incomes $20,000 to $49,999, and one category for incomes $50,000 and over.

Income is highly correlated with age of the income earner, and the associations between income and the covariates vary with income level. Therefore, to maximize goodness of fit, regressions were performed within specified age-income subgroups. Separate regressions were performed for the age groups: head of family under 25 years of age, 25–44 years of age, 45–64 years of age, and 65 years and over. Within these age groups, separate models also were fitted to persons whose family incomes was less than $20,000, $20,000 to $49,999, and $50,000 and over.

For total annual family incomes less than $50,000, the $R^2$ values for the models were high ($0.93 \leq R^2 \leq 0.97$). Initial models for the $50,000 and over income subgroups fitted relatively poorly because they did not contain any income variables ($50,000 is the top income code in the NHIS). Therefore, predicted incomes for the $50,000 and over income subgroups were obtained using models fitted to respondents with family incomes of $40,000 or more. These models yielded higher $R^2$ values ($0.39 \leq R^2 \leq 0.55$) and stronger correlation between predicted income and actual income. Because of the very long right tail of the continuous income variable, the log transformation was applied to the continuous annual family income variable for the models used for the upper income category. Sample weights were incorporated in all of the regression models because regression coefficients obtained from weighted analyses differ from those obtained from unweighted analyses. The stratification and clustering aspects of the survey designs of the CPS and NHIS were not incorporated in the modeling because they do not affect the regression coefficients.

## Regression on the National Health Interview Survey Number of Doctor Visits

A weighted least squares regression model was fitted to *number of doctor visits* (in the past 12 months) from the NHIS. In the 1995 NHIS, the number of doctor visits ranges from 0 to 706, with loading on values such as 10, 20, 30, 40, and 50. About 24% of the 102,467 respondents reported 0 visits and less than 2% reported more than 50 visits. Therefore, for the regression modeling, the number of doctor visits was trimmed to 50. A mixture of person-level variables, head-of-family-level variables, and family-level variables was included in the model. See Table C for a list of the independent variables.

Because number of doctor visits is correlated with age, separate regressions were performed for six age groups (under 5 years, 5–17 years, 18–24 years, 25–44 years, 45–64 years, and 65 years and over). The $R^2$ values for these models were low ($0.12 \leq R^2 \leq 0.24$). Sample weights were incorporated in all models.

**Table B. Independent variables included in the predictive mean match regression model fitted to the March 1996 Current Population Survey with total annual family income as the dependent variable**

| Family-level variables | Head-of-family-level variables |
|---|---|
| Total annual family income[1] (dichotomous variables):<br>Less than $5,000<br>$5,000–$6,999<br>$7,000–$9,999<br>$10,000–$14,999<br>$15,000–$19,999<br>$20,000–$24,999<br>$25,000–$29,999<br>$30,000–$34,999<br>$35,000–$39,999<br>$40,000–$44,999<br>$45,000–$49,999<br>$50,000 or more | Age in years (continuous) |
| Family size (continuous) | Age in years, squared (continuous) |
| Home ownership (dichotomous, one variable):<br>Owner<br>Renter (reference group) | Sex (dichotomous, one variable):<br>Male<br>Female (reference group) |
| Number of employed adults (continuous) | Race and Hispanic origin (dichotomous, three variables):<br>Black, not Hispanic<br>Other, not Hispanic<br>Hispanic<br>White, not Hispanic (reference group) |
| Region (dichotomous, three variables):<br>Northeast<br>Midwest<br>West<br>South (reference group) | Education (dichotomous, three variables):<br>College or more<br>Some college<br>High school<br>Less than high school (reference group) |
| Urbanization level (dichotomous, three variables):<br>Large metro, central[2]<br>Large metro, suburban[3]<br>Medium or small metro[4]<br>Nonmetro[5] (reference group) | Marital status (dichotomous, two variables):<br>Married<br>Separated, widowed, or divorced<br>Never married (reference group) |
| | Respondent-assessed health status (dichotomous, four variables):<br>Poor<br>Fair<br>Good<br>Very good<br>Excellent (reference group) |
| | Health insurance coverage (dichotomous, two variables):<br>Private insurance or Medicare<br>Public insurance<br>No insurance coverage (reference group) |
| | Employment status (dichotomous, two variables):<br>Employed<br>Unemployed<br>Not in labor force (reference group) |
| | Hours worked per week (dichotomous, one variable):<br>35 or more hours per week<br>Less than 35 hours per week (reference group) |
| | Occupational pay category[6] (dichotomous, three variables):<br>High paying jobs<br>Mid paying jobs<br>Low paying jobs<br>Not in labor force (reference group) |

[1]Regression models for persons with family incomes of less than $20,000 included four dichotomous income variables with the less than $5,000 group as the reference group. Regression models for persons with family incomes of $20,000 to $49,999 included five dichotomous income variables with the $20,000 to $24,999 group as the reference group. To obtain estimates for persons with family incomes of $50,000 and over regression models were fitted to persons with family incomes of $40,000 and over; these models included two dichotomous variables with the $40,000–$44,999 group as the reference group.

[2]A central county of a metropolitan area with a population of 1 million or more.

[3]A suburban county of a metropolitan area with a population of 1 million or more.

[4]A county in a metropolitan area of less than 1 million.

[5]A nonmetropolitan county.

[6]Occupational pay category—occupation codes were grouped into three categories: high-paying (includes executive, administrative, and managerial, and professional specialty occupations, technicians, and related support occupations); mid-paying (includes sales, administrative support, protective service, precision production, craft and repair, transportation and material moving, and military occupations); and low-paying (includes private household, service, farming, forestry, and fishing occupations, machine operators, assemblers, and inspectors, handlers, equipment cleaners, helpers, day laborers, and unknown occupations).

NOTE: The predicted values of the dependent variable, total annual family income, that were obtained from the fitted models were used in two predictive mean matches, one with the NHIS as the host file and the CPS as the Donor file, and the other with the CPS as the Host file and the NHIS as the Donor file.

**Table C. Independent variables included in the predictive mean match regression model fitted to the 1995 National Health Interview Survey with number of doctor visits as the dependent variable**

| Family-level variables | Head-of-family-level variables | Person-level variables |
|---|---|---|
| Total annual family income (dichotomous variables)<br>Less than $5,000<br>$5,0000–$6,999<br>$7,000–$9,999<br>$10,000–$14,999<br>$15,000–$19,999<br>$20,000–$24,999<br>$25,000–$29,999<br>$30,000–$34,999<br>$35,000–$39,999<br>$40,000–$44,999<br>$45,000–$49,999<br>$50,000 or more | Education (dichotomous, three variables)<br>College or more<br>Some college<br>High school<br>Less than high school (reference group) | Age in years (continuous) |
| Family size (continuous) | Marital status (dichotomous, two variables)<br>Married<br>Separated, widowed, divorced<br>Never married (reference group) | Age in years, squared (continuous) |
| Home ownership (dichotomous, one variable)<br>Owner<br>Renter (reference group) | Employment status (dichotomous, two variables)<br>Employed<br>Unemployed<br>Not in labor force (reference group) | Sex (dichotomous, one variable)<br>Male<br>Female (reference group) |
| Number of employed adults (continues) | Hours worked per week (dichotomous, one variable)<br>35 or more hours per week<br>Less than 35 hours per week (reference group) | Race and Hispanic origin (dichotomous, three variables)<br>Black, not Hispanic<br>Other, not Hispanic<br>Hispanic<br>White, not Hispanic (reference group) |
| Region (dichotomous, three variables)<br>Northeast<br>Midwest<br>West<br>South (reference group) | Occupational pay category[5] (dichotomous, three variables)<br>High paying jobs<br>Mid paying jobs<br>Low paying jobs<br>Not in labor force (reference group) | Respondent-assessed health status (dichotomous, four variables)<br>Poor<br>Fair<br>Good<br>Very good<br>Excellent (reference group) |
| Urbanization level (dichotomous, three variables)<br>Large metro, central[1]<br>Large metro, suburban[2]<br>Medium or small metro[3]<br>Nonmetro (reference group)[4] | | Health insurance coverage (dichotomous, two variables)<br>Private insurance or Medicare<br>Public insurance<br>No insurance coverage (reference group) |

[1]A central county of a metropolitan area with a population of 1 million or more.

[2]A suburban county of a metropolitan area with a population of 1 million or more.

[3]A county in a metropolitan area of less than 1 million.

[4]A nonmetropolitan county.

[5]Occupation codes were grouped into three categories: high-paying (includes executive, administrative, managerial, and professional specialty occupations; technicians, related support occupations); mid-paying (includes sales, administrative support, protective service, precision production, craft and repair, transportation and material moving, and military occupations); and low-paying (includes private household, service, farming, forestry, and fishing occupations, machine operators, assemblers, inspectors, handlers, equipment cleaners, helpers, day laborers, and unknown occupations).

NOTE: The predicted values for the dependent variable, number of doctor visits, that were obtained from the fitted models were used in the predictive mean match with the Current Population Survey as the Host file and National Health Interview Survey as the Donor file.

## Predictive Mean Matching

Using the parameter estimates obtained from the regression models, predicted values for the predictive mean match variable (total annual family income or number of doctor visits) were derived for both CPS and NHIS respondents. Next, because a constrained matching approach was used, each of the partition cells for the Donor file was aligned with the corresponding Host cell by scaling the weights in each Donor cell so that their sum equaled the sum of the weights in the corresponding Host cell. Aligning the cell weights is a necessary step as it ensures that all records will be fully used in the match. After aligning the cells, the Host and Donor records within each cell were sorted on the predicted values of the predictive mean match variable and assigned ranks. The presumption is that records with the closest predicted values (and thus, presumably the closest values of the X-variables) have a similar rank order. Finally, based on rank order, each record in the Host file was matched to the closest Donor record(s) that had not already been fully matched (had all of its sample weight used up). Both Host and Donor records were ''split'' as necessary so that all of the weight of each Host and Donor record was used up. At the end of this process, each Host record had been matched to one or more Donor records and each Donor record had been matched to one or more Host records. Figure 3 in the ''Overview of Statistical Matching Section'' illustrates this process.

### Match with Current Population Survey as Host and Predicted Number of Doctor Visits as Predictive Mean Match Variable

The partition cells were aligned by scaling the weights for each cell in the NHIS file so that their sum equaled the sum of the weights in the corresponding cell for the CPS file. After alignment, the Host and Donor records within each cell were sorted on predicted number of doctor visits. Finally, each CPS-Host record was matched to the closest NHIS-Donor record(s) (based on the rank order) that had not already been matched, as described previously. The matched file will be referred to as the CPS-Host Match 1.

### Match with Current Population Survey as Host and Predicted Total Annual Family Income as Predictive Mean Match Variable

The partition cells were aligned by scaling the weights for each cell in the NHIS file so that their sum equaled the sum of the weights in the corresponding cell for the CPS file. After alignment, the Host and Donor records within each cell were sorted on predicted total annual family income. Finally, each CPS-Host record was matched to the closest NHIS-Donor record(s) (based on the rank order) that had not already been matched, as described previously. The matched file will be referred to as the CPS-Host Match 2.

### Match with the National Health Interview Survey as Host and Predicted Total Annual Family Income as Predictive Mean Match Variable

The sample weights within each NHIS and CPS partition cell were aligned, that is, the weights for each cell in the CPS file were scaled so that their sum equaled the sum of the weights in the corresponding cell of the NHIS file. After aligning the cells, the Host and Donor records within each cell were sorted on predicted total annual family income. Each record in the NHIS-Host file was matched to the closest CPS-Donor record(s) (based on rank order) that had not already been matched as described previously.

# Results

Each of the three matched files has 231,961 observations. This number is approximately equal to the combined number of records on the CPS and NHIS and reflects the fact that records were split to enforce the constraints that all records had to be included in the match and all of the sample weights had to be used up.

The steps taken to evaluate the matches were:

1. To compare the means of selected continuous variables and the distribution of selected categorical variables on the matched file with the means and distributions of those variables on the Donor file, for the full sample and for subgroups of interest.
2. To compare cross-tabulations involving several variables on the matched and Donor files.

The set of evaluation tables presented in this report for the CPS-Host matched files is more extensive than the set presented for the NHIS-Host matched file because a CPS-Host matched file is currently being used in the Transfer Income Model (TRIM) (4). However, insights gleaned from the evaluation of the CPS-Host matched files should be directly applicable to the NHIS-Host matched file.

## Variance Estimation

The variance estimates presented in the tables for both the matched and Donor files incorporate the sample weights and survey design of the CPS or NHIS, but the variance estimates for the matched files do not incorporate the variance due to the statistical matching. As discussed in the ‘‘Overview of Statistical Matching’’ section of this report, standard errors computed for a statistically matched file that do not include an estimate of the additional variance due to the statistical match will underestimate the true variance. Unfortunately, the constrained matching procedure employed for these matches does not provide any measure of this additional component of the variance. Statistical matching procedures that explore alternatives to the CIA can provide some measure of this additional variance; however, methodology that adjusts for this type of uncertainty has not been developed. Additionally, note that none of the variance estimates presented in the tables (neither those for the matched files nor those for the Donor files) account for the variance due to imputation of missing data.

SUDAAN was used to incorporate the survey design and sample weights of the CPS and NHIS when producing variance estimates for the matched and Donor files (74). The pseudo-strata and pseudo-PSU variables available on the NHIS public-use file were used for the variance estimation on the NHIS-Donor and NHIS-Host matched files. Due to privacy restrictions in effect on all Census Bureau surveys, strata and PSU variables are not included on the CPS public-use data file. The Census Bureau suggests that variance estimates be obtained using a generalized variance function method that is described in the CPS file documentation (which does not involve strata and PSU variables), but we chose not to do so for two reasons. First, the generalized variance function method only produces estimates of variance, not estimates of covariance, and the preliminary evaluation of the match involved using both variance and covariance estimates. Second, Davern et al found that standard errors for rates and means produced using the generalized variance function method are considerably smaller than those produced using the sample design information available on internal Census Bureau CPS files (75,76). They show that standard errors produced using strata and PSU variables constructed from information available on the CPS public-use files, while smaller than those obtained using the survey design variables from the internal CPS files, are larger than the standard errors obtained using the generalized variance approach. Therefore, it was decided that for variance estimation on the CPS-Donor and CPS-Host matched files it was preferable to follow Davern et al’s recommendations to construct pseudo-strata and pseudo-PSU variables that could then be used to obtain variance estimates from SUDAAN. Our procedure was similar but not identical to the method used by Davern et al. Davern et al used the lowest level of geography available on each CPS record (state, county, and metropolitan statistical area (MSA) to define artificial pseudo-strata; we also constructed artificial pseudo-strata for CPS respondents by concatenating the available state, county, and MSA codes for each record (resulting in 408 pseudo-strata). Whereas Davern et al treated each household within an

artificial pseudo-stratum as an artificial pseudo-PSU, we constructed two artificial pseudo-PSUs within each artificial pseudo-stratum by assigning households to one of two PSUs based on whether the household ID was odd or even. (Note that the 1995 NHIS public-use file variance estimation structure has 187 pseudo-strata with two pseudo-PSUs per pseudo-stratum.) The variance estimates we obtained for the CPS-Donor and CPS-Host matched files using the pseudo-strata and pseudo-PSU appeared reasonable. They are generally smaller than NHIS variance estimates, but this was expected, given that the CPS has a larger sample size and more PSUs than the NHIS. However, bear in mind that the work by Davern et al showed that variance estimates obtained in this manner for the CPS underestimate the true variance and that this underestimation is compounded for the CPS-Host matched files because the variance due to matching is not included. For both the CPS-Host matched files and the NHIS-Host matched files, the ability to detect true differences is reduced; too many differences will be identified as statistically significant.

## Identifying Significant Differences

In the tables in this section, differences between the values found on the matched and Donor files are considered statistically significantly different if their 95% confidence intervals do not overlap. While this approach is a less precise method for assessing statistical significance than a statistical test (because it may fail to detect differences with *p*-values close to the α-level), it was judged to be adequate given the underestimation of variances from the matched files.

## Current Population Survey-Host Matches

Tables 7–24 report results from the two matches with the CPS as the Host file and the NHIS as the Donor file. Table 7 shows means for two Z-variables, that is, variables brought over to the matched files from the Donor file. As this table illustrates, generally, the means of Z-variables were similar on the matched and Donor files. The other tables show percentages from cross tabulations of X- and Z-variables, that is, variables common to both files and variables brought over to the matched files from the Donor file. Tables displayed illustrate the impact on the match results of defining subgroups using various X-variables or Z-variables: variables such as sex that were always used in the partitioning, variables such as education that were rarely used in the partitioning, and variables such as poverty level that were derived from a key variable in the predictive mean match regression. As Tables 8–24 illustrate, the percentages obtained from the various cross tabulations on the CPS-Host matched files generally were very similar to those from the NHIS-Donor file, but there were some statistically significant differences also. The evaluation highlights the importance of using variables in the partitioning that are related to the variables that will be involved in subsequent analyses of the matched file.

Table 7 shows the mean number of doctor visits per year and the mean number of bed days per year on the CPS-Host matched files and the NHIS-Donor file (both variables brought over to the CPS-Host matched file from the NHIS-Donor file). As can be seen, the means generally were similar on the matched and Donor files. However, for persons 65 years and over, the number of doctor visits and bed days in the past 12 months were overestimated on the CPS-Host matched file.

Table 8 shows the percentage of the civilian noninstitutionalized U.S. population with no doctor visits within the past 12 months for selected subgroups. Number of doctor visits is the health care variable of interest that is brought over from the NHIS-Donor file. The population subgroups in this table are defined by X-variables that played various roles in the partitioning. The first three, namely, sex, age, and race and Hispanic origin were always used in the partitioning, although some age groups occasionally had to be combined and some race and Hispanic origin groups often had to be combined. The fourth variable, education of head of family, was almost never used in the partitioning scheme because cell sizes usually were too small to permit partitioning on this variable. The percentages obtained for the sex, age, and race and Hispanic origin subgroups from the CPS-Host matched files are quite similar to each other and to those from the NHIS-Donor file. However, those obtained for the education subgroups from the CPS-Host matched files are significantly different from the percentages obtained from the NHIS-Donor file. Further, the pattern across the education subgroups observed on the NHIS Donor file (i.e., that the percentage with no doctor visits is significantly lower for the more than high school education subgroup than for the less than high school and high school subgroups) is not observed on the matched files. Note that even for the sex subgroups (the only partition variable that was always used and for which no categories were ever combined) the CPS-Host and NHIS-Donor estimates are not identical. This is a consequence of the cell alignment that had to be done prior to matching because of the differences in the CPS and NHIS weighted cell sizes.

Table 9 shows the percentage of persons 18–64 years of age in each of four activity limitation status levels. Activity limitations status is the health variable of interest brought over from the NHIS-Donor file. In this table, the population subgroups are defined by the X-variable age. For both 18–44-year-olds and 45–64-year-olds, the percentages from the two CPS-Host matched files are nearly identical to each other and are very similar to the percentages from the NHIS-Donor file. This was expected given that the subgroups for which percentages were obtained are defined by a partition variable that was always used, namely, age. The tabulations on the CPS-Host matched files used the CPS-Host values of the partition variable.

Tables 10 and 11 show the percentage of the population with no usual source of health care by age, by age and health insurance coverage, and by age and respondent-assessed health

status. Usual source of care is the health variable brought over from the NHIS-Donor file. Two partition variables are used in each table to define the population subgroups. The percentage with no usual source of health care is similar in the matched and Donor files for persons under 45 years of age, but statistically significantly different for those 45–64 years of age. Across the age and health insurance subgroups and the age and health status subgroups, the percentages with no usual source of health care from the two CPS-Host matched files (one matched using number of doctor visits, the other matched using total annual family income) generally are very similar to each other and also generally very similar to the percentages from the NHIS-Donor file. There is only one statistically significant difference: the CPS-Host matched files significantly overestimate the percentage of persons aged 45–64 years in good to excellent health who have no usual source of care. The tabulations shown in these two tables involve one Z-variable (a variable brought over to the matched file from the Donor file), usual source of health care, and two partition variables (age and health insurance and age and respondent-assessed health status, respectively). In each table, the subgroups of interest were defined using two partition variables that were always used in the partitioning. The tabulations on the two CPS-Host matched files used the CPS-Host values of the partition variables.

Tables 12 and 13 show the percentage of the population with no doctor visits within the past 12 months by age and health insurance coverage and by age and respondent-assessed health status. The percentages from the two CPS-Host matched files are nearly identical to each other and, with two exceptions, very similar to the percentages from the NHIS-Donor file. The largest discrepancy is the percentage of children in fair or poor health who have no doctor visits in the past 12 months. On the NHIS-Donor file this percentage was estimated to be 9.6, while on the two CPS-Host matched files it was estimated to be 16.6 and 16.7. The difference cannot be attributed to differences in the distribution of health status on the CPS and NHIS, as on both files about 10% of children are reported to have fair or poor health. The other discrepancy observed is for persons 45–64 years of age in good to excellent health; the percentage with no doctor visits is overestimated in the CPS-Host matched files. As for Tables 10 and 11, the percentages shown in these two tables involve one Z-variable (in this case, number of doctor visits) and two partition variables, age and health insurance coverage and age and health status. In each table, the subgroups for which percentages were obtained were defined by the two partition variables, both of which were always used in the partitioning. The cross tabulations on the CPS-Host matched files used the CPS-Host values of the partition variables.

Table 14 shows the percentage of the U.S. civilian noninstitutionalized population with no usual source of health care by age, race, and Hispanic origin. Generally, the percentages obtained from the CPS-Host matched files are similar to each other and do not differ significantly from those obtained from the NHIS-Donor file. The pattern observed on the NHIS-Donor file of non-Hispanic whites having the lowest percentage with no usual source of care and Hispanics having the highest percentage is maintained in both CPS-Host matched files. The estimates shown in this table involve one Z-variable (usual source of health care) and two partition variables, age and race and Hispanic origin. Age was always used in the partitioning, but race and Hispanic origin was not always used and when used often had some categories combined. The cross tabulations on the CPS-Host matched files used the CPS-Host values of age and race and Hispanic origin.

Table 15 shows the percentage of the population with no usual source of health care by age and education of the head of family. For all of the age by education subgroups, the percentages from the two CPS-Host matched files are similar to each other, although more disparate than seen in the earlier cross tabulations. For the two lower education subgroups (less than high school and high school) the percentages from the two CPS-Host matched files are consistently lower than those from the NHIS-Donor file, although not statistically significantly lower. For the highest education subgroup (more than high school), the percentages from the two CPS-Host matched files are higher than the percentages from the NHIS-Donor file (statistically significantly higher for persons under 65 years of age). As a result of the CPS-Host matched file estimates for the two lower education subgroups being lower than the NHIS-Donor file estimates and the CPS-Host matched file estimates for the highest education subgroup being higher than the NHIS-Donor file estimates, the differences across the education subgroups observed in the NHIS-Donor file are less pronounced in the CPS-Host matched files. As in the earlier tables, the percentages shown in this table involve one Z-variable and two partition variables (age and education of the head of the family). While the partition variable age is among the set that was always used in the partitioning, education of the head of family was one of the partition variables that was rarely used. The cross tabulations on the CPS-Host matched files used the CPS-Host values of the partition variables.

Table 16 shows the percentage of the population with no doctor visits within the past 12 months by age and education of the head of family. For all of the age by education subgroups, the percentages from the two CPS-Host matched files are very similar to each other. For persons under 65 years of age, the percentages obtained from the CPS-Host matched files generally differ significantly from those obtained from the NHIS-Donor file (for the two lower education subgroups, the percentages are lower; for the highest education subgroup, they are higher). As a result, the significant associations across the education subgroups observed on the NHIS-Donor file are substantially weakened on the CPS-Host matched files for the two youngest age groups and lost for the 45–64 year age group. For persons 65 years and over, the

percentages obtained from the CPS-Host matched files do not differ significantly from those obtained from the NHIS-Donor file. As for the earlier tables, the percentages shown in this table involve one Z-variable (number of doctor visits) and two partition variables (age and education of the head of family). As in the previous table, the subgroups for which percentages were obtained were defined by the two partition variables. While the partition variable age is among the set that was always used in the partitioning, education of the head of the family was one of the partition variables that was rarely used. The cross tabulations on the CPS-Host matched files used the CPS-Host values of the partition variables.

Table 17 shows the percentage with no usual source of health care by age and percent of poverty level. For all of the age groups, the percentages obtained from the two CPS-Host matched files are similar to each other and for the three lower poverty groups; they also are similar (except for the 200% or more poverty level) to the percentages from the NHIS-Donor file. For the three younger age groups, the percentage of persons at 200% or more of poverty who have no usual source of health care is statistically significantly over-estimated on the CPS-Host matched files. For the three younger age groups, the trends by percent of poverty level are similar in the CPS-Host matched files and the NHIS-Donor file (although slightly weaker). For the oldest age group, persons 65 years and over, the percentage of persons below 100% of poverty with no usual source of care is substantially underestimated on the two CPS-Host matched files, and as a result, the relationship between poverty and usual source of care seen in the NHIS-Donor file is lost in the two CPS-Host matched files. While the percentages obtained from the two CPS-Host matched files do not differ significantly from each other, the percentages from the CPS-Host matched file that was matched using number of doctor visits tend to be more similar to the percentages from the NHIS-Donor file. The subgroups for which percentages were obtained were defined by two X-variables, age, which was always used in the partitioning and percent of poverty level, which was derived using one of the key predictive mean match regression variables (total annual family income). The tabulations on the two CPS-Host matched files used the CPS-Host values of age and percent of poverty level. Note that the poverty level distribution by age on the March 1996 CPS and 1995 NHIS files differs; the CPS has lower estimates for the percentage of the population below 200% of poverty and a higher estimate of the percentage at 200% or more of poverty (Table V in "Appendix I").

Table 18 shows the relationship between age, poverty level, and the percentage with no doctor visits within the past 12 months among the civilian noninstitutionalized U.S. population. The percentages of persons in the three lower poverty level groups with no doctor visits within the past 12 months are somewhat lower on the CPS-Host matched files than on the NHIS-Donor file, although the differences are statistically significant only for persons 18–44 years at 100—less than 150% of poverty and for persons 65 years and over at 100—less than 150% of poverty and at 150—less than 200% of poverty. The percentages of persons at 200% or more of poverty with no doctor visits during the past 12 months is higher on the CPS-Host matched files than on the NHIS-Donor file; these differences are statistically significant for persons in the three younger age groups. As a result, some differences across the poverty levels that are significant on the NHIS-Donor file are not significant on the CPS-Host matched files, and others are weaker. For example, on the NHIS-Donor file, the percentage of children with no doctor visits is significantly higher for the two lower poverty groups compared with those at 200% or more of the poverty level, whereas on the CPS-Host matched files the differences are not significant. As for Table 16, the subgroups for which data are shown in this table are defined by one partition variable (age) and one X-variable (percent of poverty level) derived using one of the predictive mean match regression variables, namely, total annual family income. The cross tabulations on the two CPS-Host matched files used the CPS-Host values of age and percent of poverty level.

Table 19 presents health insurance coverage among persons 18–64 years of age who report that they cannot perform major activities. The percentages from the two CPS-Host matched files are similar to those from the NHIS-Donor file. This cross tabulation involved, as did earlier tables, one Z-variable and two of the primary partition variables. However, in this table percentages are obtained for one of the partition variables (health insurance coverage) rather than for the Z-variable and the subgroups are defined using the other partition variable (age) and the Z-variable (activity limitations status). The success of this particular cross tabulation may reflect the fact that health insurance was one of the partition variables as well as a variable in the predictive mean match regression models.

Table 20 shows the distribution across the four poverty status groups of working-age adults who report that they cannot perform major activities. The distribution of persons who cannot perform major activities across the poverty categories is fairly similar on the two CPS-Host matched files. However, the distribution observed on the two matched files differs from that observed on the NHIS-Donor file. For persons who cannot perform major activities, both of the CPS-Host matched files substantially underestimate the percentage of persons who live below 149% of poverty and overestimate the percentage at 200% or more of poverty. For example, of 18–44-year-olds who cannot perform major activities, 26.4% and 25.5% live below 100% of poverty according to the two CPS-Host matched files compared with 35.0% according to the NHIS. Further, according to the two CPS-Host matched files, almost one-half of 18–44-year-olds who cannot perform major activities live at 200% or more of poverty, whereas according to the NHIS, only about one-third does. The discrepancies between the CPS-Host matched files and the NHIS-Donor files seen in this table are larger than any in the previous tables. Clearly some of the NHIS respondents who cannot perform

major activities and who are living below 150% of the poverty level have been matched to CPS respondents who are living at higher income levels. In this table, as in the previous table, percentages are obtained for an X-variable (in this case, percent of poverty level) rather than for a Z-variable, and the population subgroups of interest are defined using a partition variable (age) and a Z-variable (activity limitations status). In this table, the X-variable for which percentages are estimated (percent of poverty level) is not a partition variable but is derived using total annual family income, one of the key variables in the predictive mean match regression models. The cross tabulations on the two CPS-Host matched files used the CPS-Host values of age and percent of poverty level.

Table 21 shows the percentage of working-age adults who receive Supplemental Security Income (SSI) according to age and activity limitation status. There appears to be some discrepancy in the way SSI recipiency is reported on the CPS and NHIS; the percentage of persons who receive SSI is higher on the CPS-Host matched files than on the NHIS-Donor file (1.7% on the two CPS-Host matched files compared with 1.3% on the NHIS among persons 18–44 years of age, 2.8% on the CPS-Host matched files compared with 2.5% on the NHIS-Donor file among persons 45–64 years of age). The higher rate of SSI recipiency among CPS respondents is reflected in the higher rate of SSI recipiency among CPS respondents aged 18–44 years and 45–64 years who report no activity limitations compared with NHIS respondents, although the differential is more pronounced. The rates of SSI recipiency among working-age adults who report limitations in some major activities or in other activities are similar on the CPS-Host matched files and the NHIS-Donor file. However, the rates of SSI recipiency among persons who cannot perform major activities obtained from the two matched files differ substantially from those obtained from the NHIS-Donor file. The discrepancy is not only large, it is not in the expected direction. From the NHIS we estimate that 24.1% of 18–44 year olds who cannot perform major activities receive SSI compared with only 11.4% and 11.0% on the CPS-Host matched files. In this table, as in the two previous tables, percentages are obtained for an X-variable rather than for a Z-variable and the subgroups are defined using the partition variable age and a Z-variable (activity limitations status). The X-variable for which percentages are estimated (SSI recipiency) is not a partition variable, nor a predictive mean matching variable, nor is it strongly correlated with a predictive mean matching variable. The cross tabulations on the two CPS-Host matched files used the CPS-Host values of age and SSI.

As in Table 20, the substantial differences between the CPS-Host matched files and the NHIS-Donor file seen in Table 21 indicate that mismatching has occurred. The percentages shown in Table 22 support this conclusion. Table 22 shows that 22.0% of CPS respondents aged 18–44 years and 18.1% of CPS respondents aged 45–64 years who reported that they did not work or were limited in their work within the prior calendar year due to a health problem or disability also reported that they received SSI. These percentages are very similar to the percentage of NHIS respondents who reported that they cannot perform major activities and also that they receive SSI and very different from the corresponding percentages for respondents on the CPS-Host matched files. The percentages shown in Tables 21 and 22 suggest that disabled NHIS respondents are being matched to nondisabled CPS respondents; this also may explain the finding in Table 20 that disabled persons with low income, are being matched to persons with higher income.

Comparison of race-specific estimates of SSI recipiency among working-age adults who cannot perform major activities shows substantial differences between the two CPS-Host matched files and the NHIS-Donor file (Table 23). Differences were expected given the findings presented in Tables 20 and 21. Of interest here was whether the differences were of similar magnitude across the race-ethnicity groups. The estimates of recipiency obtained from the NHIS are higher for all three of the race-ethnicity groups than those obtained from the CPS-Host matched files, some significantly higher. The largest differences in recipiency rates occur among non-Hispanic black persons ages 18–44 years. Large differences were also seen for non-Hispanic white persons. However, the recipiency rates for Hispanics did not differ significantly on the matched and donor files. For example, on the NHIS-Donor file, 33.0% of non-Hispanic black persons 18–44 years of age who cannot perform major activities receive SSI compared with only 14.7% and 16.5% on the two CPS-Host matched files. For Hispanics 18–44 years of age, 16.2% of the respondents on the NHIS who cannot perform major activities reported receiving SSI compared with 12.0% and 11.6% on the two CPS-Host matched files. Thus, the magnitude of the differences between the estimates from the NHIS-Donor file and the CPS-Host matched files differed across the race-ethnicity groups. In general, the estimates of SSI recipiency obtained from the two CPS-Host matched files are similar. Four variables are involved in this cross tabulation rather than two or three as in the previous tables. Again, the percentages are obtained for an X-variable rather than for the Z-variable. The subgroups are defined using the partition variable age that was always used in the partitioning, the partition variable race and Hispanic origin that was used less often, and a Z-variable (activity limitations status). The X-variable for which percentages are estimated (SSI recipiency) is not a partition variable, nor a predictive mean matching variable, nor is it strongly correlated with a predictive mean matching variable. The cross tabulations on the two CPS-Host matched files used the CPS-Host values of age, race and Hispanic origin, and SSI.

Table 24 shows the percentage of persons with no usual source of health care by age and level of activity limitations. The percentages obtained from the two CPS-Host matched files

are fairly similar to each other and to those obtained for the NHIS-Donor file. The only statistically significant difference between the matched and Donor estimates occurs for persons 45–64 years of age with no activity limitations. In this cross tabulation, percentages are obtained for a Z-variable (usual source of health care) and the subgroups are defined by a partition variable (age) and another Z-variable (activity limitations status).

## National Health Interview Survey-Host Match

Tables 25 and 26 report results from the match with the NHIS as the Host file, the CPS as the Donor file, and predicted total annual family income used to rank records within partition cells. For the full sample, the means of the selected variables on the matched file are quite similar to those on the Donor file. For the age subgroups, the means on the matched file also generally are similar to those on the Donor file.

Table 26 presents the percentage of persons with no health insurance coverage according to age and poverty status. The percentage of the population without health insurance coverage is similar in the NHIS-Host matched file and the CPS-Donor file for persons in the two younger age groups who are living below 200% of poverty. However, comparison of the estimated percentage uninsured for persons 45–64 years of age or for persons at 200% or more of poverty shows some statistically significant differences between the two files. For persons 45–64 years of age living at 100% or more of poverty, the percentage uninsured is substantially lower on the NHIS-Host matched file than on the CPS-Donor file. In this example, the percent of poverty level (the ratio of family income to poverty threshold) is calculated using the total annual family income variable brought over from the CPS-Donor file in the match.

# Future Current Population Survey—National Health Interview Survey Statistical Matches and Suggestions for Further Research

The purpose of this project was to determine the feasibility of statistically matching the CPS and NHIS. The statistical matches described in this report are considered successful enough to warrant performance of statistical matches of the CPS and NHIS annually to incorporate health variables into the Transfer Income model, version 3 (TRIM3) (4). TRIM3 is used to examine how major governmental tax, transfer, and health programs affect the U.S. population and to understand the potential outcomes of changes to these programs. The addition of health-related information to the TRIM3 system enhances the model's ability to simulate various programs and possible new policies. Currently, TRIM3 modelers are particularly interested in adding information on functional limitations and on private nongroup health insurance premiums to the TRIM system via a CPS-NHIS statistical match.

Disability status is of particular interest for inclusion in TRIM3 because of its relevance to program eligibility or benefits (77). For example, a nonelderly person is eligible for Supplemental Security Income (SSI) benefits only if he or she is disabled; a child aged 13 or over is eligible for federally-funded child care subsidies only if he or she has a special need such as a disability. A household that includes a disabled person receives extra exemptions in determining household income for purposes of the Food Stamp Program; and disability may confer eligibility for a federal income tax credit. Unfortunately, on the CPS, disability in children can only be identified if they receive SSI, disability in working-age individuals can only be identified if the individual reports receiving SSI or reports not working due to illness or disability, and disability in the elderly cannot be identified. Without additional health variables from the NHIS, TRIM3 must use this limited disability definition for all purposes. While it is appropriate for some purposes, it is not appropriate for others.

The Office of the Assistant Secretary for Planning and Evaluation (ASPE) also is interested in adding private nongroup health insurance premium data to TRIM3 to facilitate evaluation of various programs and policies impacted by health insurance. For example, the availability of private nongroup premium data to TRIM3 would enable ASPE to conduct ''what if'' simulations, for instance to test the possible impact of a hypothetical or proposed health insurance tax credit. The availability of private nongroup premiums for TRIM3 would also allow creation of a comprehensive measure of income (health costs and taxes subtracted, transfers added) that could be used in looking at how persons spend down into poverty. The March CPS includes substantial information about health insurance coverage, but no information about health insurance premiums. Various imputations and matches have been used to add health insurance premiums into the TRIM3 database. However, the method that has most recently been used to impute private nongroup premiums—a ''look up table'' of premiums of Web-based quotes from insurance companies, varying by key demographic characteristics—may provide an inaccurate picture of actual premiums because it is difficult to incorporate the variations arising from different plan choices and different health conditions.

The NHIS variables that the Urban Institute and ASPE researchers consider to be the most likely candidates for use in TRIM3 simulations or tabulations are as follows (76):

- Private nongroup health insurance—For up to two private nongroup policies that an individual holds, the type of plan (family or individual) and the annual premium amount.
- Disability variables:
  - For children 4 years of age and under: limitations in play activities (kind, amount).
  - For children 17 years of age and under: receipt of special education or early intervention.
  - For all persons 4 years of age and over: whether person needs help with personal care, and if so, the number of specific needs (bathing, dressing, eating, transferring, toileting, getting around the house).
  - For persons 18 years of age and over: whether person needs help handling routine needs.
- Health condition variables: For persons with at least one limitation, the condition or health problem causing the limitation (vision problem, heart problem, stroke, hypertension, diabetes, cancer, depression, anxiety, or emotional problem, and alcohol or drug-related problems).
- Health care utilization:
  - Number of overnight hospital stays within the year.
  - Whether received care from a health professional 10 or more times during the year.

Building on the experience gained in the CPS-NHIS Statistical Matching Project, the Urban Institute has carried out statistical matches between some of the more recent CPS and NHIS surveys in order to add NHIS variables to the TRIM3 system (77). Consistent with this project's demonstration of the importance of using key analysis variables as partition variables, the partition scheme used in the more recent statistical matches addresses ASPE's current interest in nongroup health insurance. The new matches have been done in such a way that CPS respondents who are the policyholder of a private nongroup family policy are matched only to NHIS respondents who also are the policyholder of a private nongroup family policy. Further, given the knowledge gained in this project about the performance of different matching techniques, the newer matches have used the predictive mean match regression approach to assess closeness (rather than a distance function) of Host and Donor records. However, despite the findings of this study, and earlier studies in the literature, that constrained matching results in a better match than unconstrained matching, the current CPS-NHIS matches have been done using unconstrained matching because the TRIM3 system cannot accommodate the splitting of records that inevitably occurs with constrained matching.

The results of the statistical matches of the CPS and NHIS described in this report suggest a number of areas in which future research could result in significant improvements in the quality of subsequent matches.

## Partitioning

To some extent, the partitioning carried out for this project was an exercise driven by intuition and by trial and error. Methods for selecting the partition variables, determining their order, determining cut points for continuous variables and groupings for categorical variables, and assessing how ''deep'' the partitioning should go would be highly useful. Exploration of the utility of classification and regression tree procedures, which have been implemented in various software packages, in developing the partitioning scheme may be worthwhile. Development of statistical tests or diagnostics to aid the partitioning also is desirable.

## Multivariate Measures

The matching strategy relied on using the predicted values of only one variable. It seems reasonable to expect improved performance if a number of variables are used.

## Conditional Independence Assumption

It would be useful to gain a clearer understanding of when the conditional independence assumption (CIA) holds and when it does not, when violations of the CIA will result in serious problems in the statistical match, and what can be done to remedy problems resulting from violations of the assumption. Clearly in the CPS-Host matches presented in this report, the CIA did not hold for SSI and this resulted in problems with cross tabulations involving this variable. It may be that partitioning on an intervening variable would have remedied the problems resulting from violation of the CIA. The Medical Expenditures Panel Survey (MEPS) and the MEPS-NHIS linkage could prove extremely useful for addressing these methodological questions. These files have information on health status, health care utilization, and family resources. The files are unsuitable as a replacement for the CPS-Host matched file or the NHIS-Host matched file because of their complexity and because of the small sample size of MEPS. Additionally, although the MEPS is unsuitable as a replacement for a statistically matched CPS-NHIS file, it may be suitable as a source of auxiliary data that could be used to improve the quality of the statistical match. As discussed in the ''Overview of Statistical Matching'' section of this report, a number of statistical matching procedures have been developed that use auxiliary data to reduce reliance on the CIA (48,49, 53,57–59,62).

## Constrained Matching Compared With Unconstrained Matching

Problems were encountered during the course of the statistical matching project when unconstrained matching was used, primarily that a significant portion of the Donor sample failed to match with the Host sample. The solution to this was to perform a constrained match. However, constrained matching results in a data file with multiple records for each respondent in the Host file (because all of the sample weights must be ''used up'' in a constrained match, a Host record may

be matched to multiple Donor records, with its sample weight ''split'' to match the sample weights of the Donor records). This is not a problem when performing person-level analyses, but is problematic when performing family-level or household-level analyses because family or household members now appear multiple times in the data file (with different covaraiate values on each record). Each possible combination of the replicates of the individuals in the family/household results in a new family/household. Thus, using a fully constrained statistically matched file with the TRIM3 system has proved problematic because TRIM operates on families and households as well as on persons. As a result, current matched files for use in TRIM3 are being constructed using unconstrained matching. Work is currently being performed by TRIM project staff, in consultation with ASPE staff, to reduce the portion of the Donor file that remains unmatched in new CPS-NHIS matches. Additional work should be done to evaluate the benefits of modifying the TRIM3 system to incorporate the results of a constrained match.

## Application of Other Statistical Matching Procedures

Moriarity and Scheuren developed a new statistical matching procedure after the CPS-NHIS match occurred (66,68,69). To date, research on the new procedure has been limited to simulations involving large simple random samples from multivariate normal distributions. An advantage of the new procedure is that it provides information about the uncertainty due to the statistical matching process that is not available from the other procedures. Adaptation of this new procedure to a match of the CPS and NHIS, if feasible, could provide important additional information about the uncertainty introduced by the matching process.

# Summary

The results reported here suggest that the fully-constrained predictive mean match of the March 1996 CPS and 1995 NHIS data files had some success in combining the information on the two files, but that relationships among some variables on the matched file may be quite different from the relationships between those variables on the Donor file. The results of this evaluation suggest that the partition variables, and to a lesser extent, the variables involved in the predictive mean match regression play an important role in determining whether relationships among variables of interest on the matched file correctly represent relationships among those variables in the population. Thus, the results of the evaluation suggest that the partition variables should be chosen based on the expected uses of the matched file; preferably they should be key analytic variables or variables that are highly correlated with key analytic variables. The evaluation also suggested that estimates for small subgroups that are not defined using partition variables or variables strongly correlated with partition variables may be especially subject to error. The evaluation reinforces the need to proceed carefully when exploring relationships among variables brought over to the matched file from the Donor file and Host variables not found on the Donor file. When possible, relationships among Donor variables and Host variables on the matched file should be checked against the relationships among those variables on the Donor file, or on an auxiliary file.

In the statistical match of the March 1996 CPS and the 1995 NHIS, means and percentages obtained for individual variables brought over to the matched file from the Donor file generally were very similar to the means and percentages of those variables on the Donor file, both for the full sample and for subgroups defined by one of the partition variables. This result was expected for the full sample, given that a constrained match was performed. Estimates also were expected to be close for subgroups derived from partition variables that were always used in the partitioning. For example, sex was always used as a partition variable, so sex-specific means and percentages of variables brought over to the matched files should be, and were, almost identical to the means and percentages obtained for those variables on the Donor file. Categorized age was always used in the partitioning, but at times several age groups were combined for the partitioning. Thus, some differences in means and percentages of variables on the matched and Donor file might be expected when computed by age (particularly for age groupings narrower than those used in the partitioning). Note however, that if age groups broader than those used in the partitioning are used to form the analysis subgroups, such as 18–44 years, estimates from the matched file should be nearly identical to those from the Donor file. This was found to be the case. Estimates on the matched and Donor files also might be expected to differ somewhat when subgroups are defined by a partition variable that was not always used in the partitioning and when used its categories often had to be combined. Race and Hispanic origin was one such partition variable. As the cross tabulation of age, race and Hispanic origin, and number of doctor visits shows, however, even for partition variables that have experienced some category collapsing, the estimates obtained from the matched file can still be similar to those obtained from the Donor file. It seemed likely that in cross tabulations involving partition variables that were not used extensively in the partitioning (such as region, urbanization level, and education of the head of family) estimates from the matched and Donor files would be more disparate. This did appear to be the case as the differences between the matched and Donor file estimates seen in the cross tabulation of education of head of family and number of doctor visits were statistically significant.

The distributions of Z-variables on the matched and Donor files tended to be similar for subgroups defined by multiple partition variables (such as age

and health insurance coverage), provided that those partition variables were extensively used in the partitioning scheme. For example, estimates of the percentage with no usual source of health care (the Z-variable) for age-health insurance coverage subgroups (both partition variables) were similar on the matched and Donor files. The similarity of matched and Donor estimates obtained from many cross tabulations of this type was reassuring. However, the evaluation showed that while reasonable estimates can be expected from the matched file when subgroups in a cross tabulation are based on multiple partition variables, substantial differences can still occur. For example, in the tabulation involving subgroups formed using the partition variables age and respondent-assessed health status and the Donor variable number of doctor visits in the past year, the percentage of children in fair to poor health who had no doctor visits was significantly overestimated on the CPS-Host matched files. The distributions of Z-variables on the matched and Donor files tended to differ when some of the variables used to define the subgroups were partition variables that were rarely used in the partitioning scheme. Specific examples are the tabulations with subgroups defined by age and education of head of family. Not only were the differences between the matched and Donor file estimates larger and more frequent than the differences seen in tabulations where the subgroups were defined using extensively used partition variables such as health insurance coverage, but the significant trends observed on the Donor file were consistently lost on the matched files.

When some of the variables used to define subgroups involved in a tabulation were not partition variables but were used in the predictive mean match regression model or were strongly correlated with predictive mean match regression variables, the estimates obtained from the matched and Donor files still tended to be similar. An example of this is the tabulations that involved percent of poverty level. While poverty status was not a variable in the predictive mean match regression model, total annual family income, used to calculate percent of poverty level, is. In the cross tabulation of age, percent of poverty level, and doctor visits, the percentage with no visits within the past 12 months generally was similar in the matched and Donor files across the age-percent of poverty subgroups. The only statistically significant differences between the percentages obtained from the CPS-Host matched files and those obtained from the NHIS-Donor file occurred for persons under 65 years of age who were at 200% or more of poverty. The problems with this category may reflect differences in the distribution of total annual family income on the CPS and NHIS and the fact that the total annual family income variable on the NHIS was top coded at $50,000. Despite the limited number of statistically significant differences between the matched and Donor estimates, some of the associations with poverty status observed in the Donor file were not significant on the CPS-Host matched files. For example, on the NHIS-Donor file, the percentage of children with no doctor visits is significantly higher for children below 200% of poverty compared with those having higher income, whereas on the CPS-Host matched files the differences were not significant. Thus, distortions of relationships among variables seem more likely when some of the variables used to define the subgroups of interest are not partition variables.

When some of the variables used to define subgroups involved in a tabulation are Z-variables, the estimates obtained from the matched files could be quite similar to those obtained from the Donor file but they could also be quite dissimilar. For example, activity limitations status, one of the variables brought over to the matched files from the Donor file, was crossed with age to form subgroups that were used in several tabulations. In one of these tabulations, the percentages without health insurance coverage from the matched and Donor files were nearly identical, perhaps because health insurance coverage is a partition variable. However, in another tabulation involving age-activity limitation status subgroups, some matched and Donor file estimates of the percentage in each poverty category differed substantially for the ''cannot perform major activities'' subgroups. In fact, the discrepancies between the matched and Donor files seen in this tabulation were some of the largest seen in the evaluation. It seemed clear that mismatching had occurred; specifically, that some of the NHIS respondents who cannot perform major activities and who are living below 150% of the poverty level were matched to CPS respondents living at higher income levels. In a third tabulation involving age and activity limitations subgroups, some estimates of the percentage receiving SSI obtained from the matched and Donor files differed considerably. This tabulation indicates that disabled respondents have been matched with nondisabled respondents. The similarity of the percentage of CPS working-age adults unable to work due to disability who receive SSI and the percentage of NHIS working-age adults unable to perform major activities who receive SSI was further confirmation of the mismatching. The mismatching and its impact on the matched estimates revealed by the activity limitations and SSI cross tabulations may have occurred because neither activity limitations nor SSI was a partition variable and neither is strongly correlated with any of the partition variables. Mismatching of working-age disabled respondents could probably be reduced if a variable such as labor force status was used in the partitioning. The impact of mismatching may have been particularly noticeable for SSI as the percentage of persons who are so disabled that they cannot perform major activities is very small and the percentage who receives SSI also is small. Matching even a relatively small number of ''disabled SSI recipients'' with nondisabled persons would, therefore, result in large distortions of the estimates. These three tabulations provide further evidence that the partition variables play a key role in determining the success of the match and point out the potential difficulty of obtaining estimates from the matched file for small subgroups.

It is worth noting that the standard errors calculated for the matched files are underestimated because they do not account for the uncertainty introduced by the statistical match itself. As a result, too many differences will be found to be statistically significant, and thus, the power to detect true differences is, to some extent, reduced in the matched files. The power to detect true differences may be further reduced in the CPS-Host matched files because of the less adequate information available on the CPS public-use files for variance estimation. As was noted earlier, the standard errors for the CPS-Host matched files tended to be lower than those for the NHIS-Donor file. Whether this is attributable to the survey design, the match, or the method used to obtain the standard errors is not clear.

Performing the statistical match of the CPS-Host and NHIS-Donor files using both predicted number of doctor visits and predicted total annual family income was a useful way to evaluate the matching strategy. The rationale for using number of doctor visits as the predictive mean matching variable when the NHIS is the Donor file is that 1) the set of common variables that predict health care variables probably differs from the set that predicts economic variables and 2) there are correlations among health care variables. Thus, one might expect that a match achieved using a health care variable as the predictive mean match variable would be better than a match achieved using an economic variable as the predictive mean match variable. However, we found that the cross tabulations for the CPS-Host matched file with number of doctor visits used as the matching variable generally were quite similar to those for the CPS-Host matched file with total annual family income used as the matching variable. It is possible that this result occurred because of the much poorer fit of the doctor visit regression models compared with the fit of the income regression models. When differences between the two CPS-Host matched file estimates were more substantial, there was no consistency in which file had estimates closer to NHIS-Donor estimates. The results of the matches suggest that partitioning variables have a far stronger impact on the success of the match than does choice of the variable used for the predictive mean match.

The results of the evaluation indicate that the matched files can be used to obtain estimates of population means and percentages for the full sample and for population subgroups. However, there were discrepancies between some of the estimates obtained on the matched file and the corresponding estimates on the Donor file, some of which were statistically significant and some of which resulted in the elimination of a relationship between variables. There was some evidence to indicate that when subgroups of interest can be identified prior to performing the statistical match, the variables needed to define those subgroups should be used as partition variables in the match. In addition, if analyses involving specific variables can be anticipated, for example, analyses involving measures of disability, then the inclusion of variables highly correlated with those key analysis variables in the partitioning scheme is desirable. Given that for some cross tabulations we found substantial differences between estimates from the matched file and estimates from the Donor file and changes in the relationships among variables, users are advised to cross-check estimates obtained from the matched file, insofar as possible, against estimates obtained from the Donor file or from an auxiliary file. Of course, the purpose of doing a statistical match is to obtain estimates that involve some Host-only and some Donor-only variables. This evaluation makes it clear that in order to have confidence in these estimates, the variables involved that are not common to both files should be highly correlated with a partition variable and the variables common to both files should either be partition variables or highly correlated with one. Finally, although some of the cross tabulations in this report successfully use three or more variables, great caution should be used when estimating multivariate relationships.

# References


1. Bureau of the Census. Current Population Survey, March 1996 [machine-readable data file] conducted by the Bureau of the Census for the Bureau of Labor Statistics. Washington: Bureau of the Census (producer and distributor). 1996.
2. Bureau of the Census. 1996 Current Population Survey, March 1996 Technical Documentation, prepared by Administrative and Customer Services Division, Microdata Access Branch, Bureau of the Census. Washington: Bureau of the Census. 1996.
3. National Center for Health Statistics. Data file documentation, (machine readable data file and documentation, CD-ROM Series 10, no. 10C), National Center for Health Statistics, Hyattsville, MD. 1998.
4. Transfer Income Model. The Urban Institute. Available from: http://trim3.urban.org/T3Welcome.php.
5. Ingram DD, O'Hare J, Scheuren F, Turek J. Statistical matching: A new validation case study. Proceedings of the Section on Survey Research Methods. Am Stat Assoc 746–51. 2001.
6. Ingram DD, Moriarity C. Statistical match of the 1995 National Health Interview Survey and the March 1996 Current Population Survey. Proceedings of the Section on Survey Research Methods. Am Stat Assoc. 2003.
7. Budd EC, Radner DB. The OBE size distribution series: Methods and tentative results for 1964. Am Econ Rev LLX: 435–9. 1969.
8. Budd EC. The creation of a microdata file estimating the size distribution of income. Review Income Wealth 17:317–33. 1971.
9. Budd EC. Comments (on constructing a new database from existing microdata sets: the 1966 Merge File, by BA Okner and Sims' comments on Okner's article). Annals Economic Social Measurement 1:349–54. 1972.
10. Okner BA. Constructing a new data base from existing microdata sets: the 1966 Merge File. Annals Economic Social Measurement 1:325–42. 1972.
11. Okner BA. Reply and comments (to Peck and Sims). Annals Economic Social Measurement 1:359–62. 1972.
12. Peck JK. Comments ((on Comments by Sims) Annals Economic Social Measurement 1:347–78. 1972.

13. Sims CA. Comments (on Constructing New Data Bases From Existing Microdata Sets: The 1966 Merge File, by B.A. Okner). Annals Economic Social Measurement 1:343–5. 1972.
14. Sims CA. Rejoinder (to Budd and Peck). Annals Economic Social Measurement 1:355–7. 1972.
15. Budd EC, Radner DB, Hinrichs JC. Size distribution of family personal income: Methodology and estimates for 1964. Bureau of Economic Analysis Staff Paper No. 21. U.S. Department of Commerce. 1973.
16. Alter HE. Creation of a synthetic data set by linking records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey: 1970. Annals Economic Social Measurement 2:373–94. 1974.
17. Okner BA. Data matching and merging: An overview. Annals Economic Social Measurement 3:347–52. 1974.
18. Ruggles N, Ruggles R. A strategy for merging and matching microdata sets. Annals Economic Social Measurement 3:353–71. 1974.
19. Sims CA. Comments (on articles by Alter and Ruggles and Ruggles). Annals Economic Social Measurement 3:395–7. 1974.
20. Turner JS, Gillian GE. Reducing and merging microdata files. OTA Paper 7, Office of Tax Analysis, U.S. Treasury Department. U.S. Government Printing Office. Washington, DC. 1975.
21. Beebout H, Doyle P, Kendall A. Estimation of food stamp participation and cost for 1977: A microsimulation approach (Final report). MPR Working Paper #E-48. Mathematica Policy Research, Inc. 1976.
22. King JA. The distributional impact of energy policies: Development and application of the Phase I Comprehensive Human Resources Data System (Task 8 Final Report). Project Report Series MPR/PR 77–13. Mathematica Policy Research, Inc. 1977.
23. Ruggles N, Ruggles R, Wolff E. Merging microdata: Rationale, practice, and testing. Annals Economic Social Measurement 6:429–44. 1977.
24. Barr RS, Turner JS. A new, linear programming approach to microdata file merging. 1978 Compendium of Tax Research. Office of Tax Analysis, Department of the Treasury. U.S. Government Printing Office, Washington, DC. 131–49. 1978.
25. Fellegi IP. Discussion. 1977 Proceedings of the Social Statistics Section, Am Stat Assoc 762–4. 1978.
26. Goldman AJ. Comments (in 1978 Compendium of Tax Research sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. 1978.
27. Kadane JB. Some statistical problems in merging data files. 1978 Compendium of Tax Research. Office of Tax Analysis, U.S. Department of the Treasury. U.S. Government Printing Office, Washington, DC. 159–71. 1978. Reprinted: J Official Statistics, 17:423–33. 2001.
28. Kadane JB. Reply (to Sims' Comments (on Some statistical problems in merging data files). 1978 Compendium of Tax Research, Office of Tax Analysis, U.S. Department of the Treasury. U.S. Government Printing Office, Washington, DC. 177–9. 1978.
29. Radner DB. Age and family income. Paper presented at the NBER Workshop on Policy Analysis with Social Security Research Files. Williamsburg, VA, March 15–17, 1978 (in Policy Analysis with Social Security Research Files, the proceedings of the workshop). 1978.
30. Radner DB, Muller HJ. Alternative types of record matching: Costs and benefits. 1977 Proceedings of the Social Statistics Section. Am Stat Assoc 756–61. 1978.
31. Sims, CA. Comments (on Some statistical problems in merging data files by JB Kadane). 1978 Compendium of Tax Research, Office of Tax Analysis, U.S. Department of the Treasury. U.S. Government Printing Office, Washington, DC. 172–7. 1978.
32. Colledge MJ, Johnson JB, Pare R, Sande IG. Large scale imputation of survey data. 1978 Proceedings of the Section on Survey Research Methods. Am Stat Assoc 431–6. 1979.
33. Hollenbeck K, Doyle P. Distributional characteristics of a merged microdata file. Proceedings of the Section on Survey Research Methods Am Stat Assoc 418–20. 1979.
34. Radner DB. The development of statistical matching in economics. 1978 Proceedings of the Social Statistics Section. Am Stat Assoc 503–8. 1979.
35. Barr RS, Turner JS. Merging the 1977 statistics of income and the March 1978 Current Population Survey. Report prepared for the Office of Tax Analysis, U.S. Department of the Treasury. Washington, DC. 1980.
36. Radner DB, Allen R, Gonzalez ME, Jabine TB, Muller HJ. Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, U.S. Department of Commerce. Washington, DC. U.S. Government Printing Office. 1980.
37. Rubin DB. File concatenation with adjusted weights and multiple imputations: A solution to the file matching problem different in principle from the constrained optimization approach. Unpublished manuscript. Social Security Administration. 1980.
38. Radner DB. An example of the use of statistical matching in the estimation and analysis of the size distribution of income. Review Income Wealth 211–42. 1981.
39. Barr RS, Stewart WH, Turner JS. An empirical evaluation of statistical matching strategies. Unpublished manuscript, Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX. 1982.
40. Klevmarken NA. Missing variables and two-state least squares estimation from more than one data set. 1981 Proceedings of the Business and Economics Statistics Section. Am Stat Assoc 156–61. 1982.
41. Rodgers WL, DeVol E. An evaluation of statistical matching. 1981 Proceedings of the Section on Survey Research Methods. Am Stat Assoc 128–32. 1982.
42. Kelley RP. A preliminary study of error structure of statistical matching. 1983 Proceedings of the Social Statistics Section. Am Stat Assoc 206–8. 1983.
43. Radner DB. Adjusted estimates of the size distribution of family money income. J Business Economic Statistics 1:136–46. 1983.
44. Rubin DB. Discussion (of statistical record matching for files. by DB Rubin). Incomplete Data in Sample Surveys (Vol 3): Proceedings of the Symposium, (eds. Madow WG, Olkin I). Academic Press, New York 203–5. 1983.
45. Woodbury MA. Statistical record matching for files. Incomplete data in sample surveys (Vol 3): Proceedings of the Symposium, (eds. Madow WG, Olkin I). Academic Press, New York 173–81. 1983.
46. Rodgers, WL. An evaluation of statistical matching. J Business Economic Statistics 2:91–102. 1984.
47. Gavin NI. An application of statistical matching with the Survey of Income and Education and the 1976 National

Health Interview Survey. Health Services Research 20:183–98. 1985.
48. Paass G. Statistical record linkage methodology: state of the art and future prospects. Proceedings of the 100th Session of the International Statistical Institute. International Statistical Institute, Amsterdam. 1985.
49. Paass G. Statistical match: evaluation of existing procedures and improvements by using additional information. Microanalytic Simulation Models Support Social and Financial Policy, (eds. Orcutt GH, Merz J, Quinke H). Elsevier Science, Amsterdam 401–22. 1986.
50. Rubin, D.B. Statistical matching using file concatenation with adjusted weights and multiple imputations. J Business Economic Statistics 4:87–94. 1986.
51. Wolfson M, Gribble C, Bordt M, Murphy B, Rowe G. The social policy simulation database: An example of survey and administrative data integration. Proceedings, Statistical Uses of Administrative Data, an International Symposium 201–29. 1987.
52. Barry JT. An investigation of statistical matching. J Applied Statistics 15:275–83. 1988.
53. Singh AC, Armstrong J, Lemaître GE. Statistical matching using log-linear imputation. Proceedings of the Section on Survey Research Methods. Am Stat Assoc 672–77. 1988.
54. Armstrong J. An evaluation of statistical matching methods, Working Paper No. BSMD 90–003E, Methodology Branch, Statistics Canada, Ottawa. 1989.
55. Wolfson M, Gribble M, Bordt M, Murphy B, Rowe G. The social policy simulation data base and model: An example of survey and administrative data integration. Survey of Current Business 69: 36–41. 1989.
56. Cohen ML. Statistical matching and microsimulation models. Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling, Vol. II, Technical Papers (eds. Citro CF, Hanushek EA). National Academy Press, Washington, DC. 1991.
57. Draper D, et al. Combining information: Statistical issues and opportunities for research. Contemporary Statistics, No. 1, National Academy Press and the American Statistical Association. 1992.
58. Singh AC, Mantel HJ, Kinack MD, Rowe G. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. Survey Methodology 19:59–79. 1993.
59. Kovacevic MS, Liu T. Statistical matching data files: A simulation study. Proceedings of the Section on Survey Research Methods. Am Stat Assoc 487–94. 1994.
60. Liu TP, Kovacevic MS. Categorically constrained matching. Proceedings of the Survey Methods Section, Statistical Society of Canada 123–33. 1996.
61. Kamakura WA, Wedel M. Statistical data fusion. J Marketing Research 34:485–98. 1997.
62. Liu TP, Kovacevic MS. An empirical study on categorically constrained matching. Proceedings of the Survey Methods Section, Statistical Society of Canada 167–78. 1997.
63. Liu TP. A categorical constraints guided matching algorithm. Proceedings of the Survey Methods Section, Statistical Society of Canada 249–56. 1998.
64. Rensen RH. Use of statistical matching techniques in calibration estimation. Survey Methodology 24:171–83. 1998.
65. O'Hare JF. Impute or match: Strategies for microsimulation modeling. Presented at the Conference on Combinatorics, Portland, ME. Advances in Economic Analysis, North Holland, 1999.
66. Moriarity C, Scheuren F. Statistical matching: A paradigm for assessing the uncertainty in the procedure. J Official Statistics 17:407–22. 2001.
67. Rässler S. Statistical matching: A frequentist theory, practical applications and alternative Bayesian approaches. Lecture Notes in Statistics. New York: Springer Verlag. 2002.
68. Moriarity C, Scheuren F. A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. J Business Economic Statistics 21:65–73. 2003.
69. Moriarity C, Scheuren F. Regression-Based Statistical Matching: Recent Developments. 2004 American Statistical Association Proceedings, 4050–7. ASA. 2004.
70. Goel PK, Ramalingan T. The Matching Methodology: Some Statistical Properties. Lecture Notes in Statistics. New York: Springer Verlag. 1989.
71. Little, RJA. Missing Data in Census Bureau Surveys, Proceedings of the Second Annual Research Conference, Washington, DC, U.S. Bureau of the Census, 442–54. 1986.
72. Benson V, Marano MA. Current estimates for the National Health Interview Survey, 1995. National Center for Health Statistics. Vital Health Stat 10(199). 1998.
73. National Center for Health Statistics. National Health Interview Survey Imputed annual family income. CD ROM. series 10(9). 1999.
74. Shah BV, Barnwell BG, Bieler GS. SUDAAN Software for the Statistical Analysis of Correlated Data User's Manual Release 9.0.0. Research Triangle Park, North Carolina: Research Triangle Institute. 2004.
75. Davern M, Davidson G, Blewett L, Jones A, Lepkowski J. Estimating standard errors for regression coefficients using the Current Population Survey's public-use file. Proceedings of the Joint Statistical Meetings, 1516–23. Am Stat Assoc (Alexandria, VA). 2004.
76. Davern M, Jones A, Lepkowski J, Davidson G, Blewett LA. Statistical error estimation and the Current Population Survey: Various approaches produce unstable estimates. Presented at the Joint Statistical Meetings, Section on Survey Research Methods, San Francisco, CA. August 3–7. 2003.
77. O'Hare J, Morton J, Giannarelli L. Final report on the match of the 2001 NHIS data with CY 2001 TRIM-CPS data. ASPE Technical Memorandum. 2003.

**Table 1. Unweighted record counts and percent distribution for the March 1996 Current Population Survey and 1995 National Health Interview Survey, by sex, race, and Hispanic origin**

| Sex, race, and Hispanic origin | March 1996 CPS | | 1995 NHIS | |
|---|---|---|---|---|
| | Number | Percent distribution | Number | Percent distribution |
| Male | 62,424 | 100.0 | 48,809 | 100.0 |
| White, not Hispanic | 44,774 | 71.7 | 31,165 | 63.9 |
| Black, not Hispanic | 5,642 | 9.0 | 6,007 | 12.3 |
| Other, not Hispanic | 2,749 | 4.4 | 1,812 | 3.7 |
| Hispanic | 9,259 | 14.8 | 9,825 | 20.1 |
| Female | 68,052 | 100.0 | 53,658 | 100.0 |
| White, not Hispanic | 47,662 | 70.0 | 33,856 | 63.1 |
| Black, not Hispanic | 7,251 | 10.7 | 7,356 | 13.7 |
| Other, not Hispanic | 3,037 | 4.5 | 1,958 | 3.6 |
| Hispanic | 10,102 | 14.8 | 10,488 | 19.5 |
| Total | 130,476 | 100.0 | 102,467 | 100.0 |
| White, not Hispanic | 92,436 | 70.9 | 65,021 | 63.5 |
| Black, not Hispanic | 12,893 | 9.9 | 13,363 | 13.0 |
| Other, not Hispanic | 5,786 | 4.4 | 3,770 | 3.7 |
| Hispanic | 19,361 | 14.8 | 20,313 | 19.8 |

NOTES: CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

**Table 2. Weighted record counts and percent distribution for the March 1996 Current Population Survey and 1995 National Health Interview Survey, by sex, race, and Hispanic origin**

| Sex, race, and Hispanic origin | March 1996 CPS | | 1995 NHIS | |
|---|---|---|---|---|
| | Number | Percent distribution | Number | Percent distribution |
| Male | 129,143,329 | 100.0 | 127,576,540 | 100.0 |
| White, not Hispanic | 93,712,074 | 72.6 | 93,187,672 | 73.0 |
| Black, not Hispanic | 15,443,049 | 12.0 | 14,838,138 | 11.6 |
| Other, not Hispanic | 5,610,072 | 4.3 | 5,571,265 | 4.4 |
| Hispanic | 14,378,134 | 11.1 | 13,979,465 | 11.0 |
| Female | 135,170,835 | 100.0 | 134,326,384 | 100.0 |
| White, not Hispanic | 97,558,944 | 72.2 | 97,771,132 | 72.8 |
| Black, not Hispanic | 17,630,116 | 13.0 | 17,089,992 | 12.7 |
| Other, not Hispanic | 5,922,158 | 4.4 | 5,713,421 | 4.3 |
| Hispanic | 14,059,617 | 10.4 | 13,751,839 | 10.2 |
| Total | 264,314,164 | 100.0 | 261,902,924 | 100.0 |
| White, not Hispanic | 191,271,018 | 72.4 | 190,958,804 | 72.9 |
| Black, not Hispanic | 33,073,165 | 12.5 | 31,928,130 | 12.2 |
| Other, not Hispanic | 11,532,230 | 4.4 | 11,284,686 | 4.3 |
| Hispanic | 28,437,751 | 10.8 | 27,731,304 | 10.6 |

NOTES: CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

**Table 3. Percent distribution of total annual family income: March 1996 Current Population Survey and 1995 National Health Interview Survey**

| Total annual family income | Unweighted percent distribution | | Unweighted percent distribution | |
|---|---|---|---|---|
| | March 1996 CPS | 1995 NHIS | March 1996 CPS | 1995 NHIS |
| Total income | 100.0 | 100.0 | 100.0 | 100.0 |
| Less than $5,000 | 3.8 | 3.9 | 3.8 | 3.5 |
| $5,000–$6,999 | 2.7 | 2.9 | 2.6 | 2.6 |
| $7,000–$9,999 | 4.4 | 4.9 | 4.2 | 4.3 |
| $10,000–$14,999 | 7.8 | 10.0 | 7.6 | 8.9 |
| $15,000–$19,999 | 7.8 | 10.1 | 7.6 | 9.4 |
| $20,000–$24,999 | 7.4 | 9.4 | 7.2 | 9.0 |
| $25,000–$34,999 | 13.9 | 15.9 | 13.7 | 16.1 |
| $35,000–$49,999 | 17.4 | 18.0 | 17.2 | 18.9 |
| More than $50,000 | 34.9 | 25.1 | 36.1 | 27.4 |

NOTES: CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

**Table 4. Unweighted numbers and percent distribution of persons with selected types of health insurance coverage, by sex and respondent-assessed health status: March 1996 Current Population Survey and 1995 National Health Interview Survey**

| Sex and health insurance coverage | Health status | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Excellent/very good/good | | | | Fair/poor | | | |
| | March 1996 CPS | | 1995 NHIS | | March 1996 CPS | | 1995 NHIS | |
| | Number | Percent distribution | Number | Percent distribution | Number | Percent distribution | Number | Percent distribution |
| Male | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 |
| No insurance | 9,314 | 16.6 | 8,035 | 18.4 | 903 | 14.0 | 948 | 18.2 |
| Public insurance | 4,715 | 8.4 | 4,251 | 9.7 | 850 | 13.2 | 788 | 15.1 |
| Private insurance or Medicare | 41,947 | 74.9 | 31,317 | 71.8 | 4,695 | 72.8 | 3,470 | 66.7 |
| Female | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 |
| No insurance | 8,765 | 14.8 | 7,288 | 15.6 | 1,082 | 12.5 | 1,094 | 15.9 |
| Public insurance | 6,253 | 10.5 | 5,884 | 12.6 | 1,262 | 14.6 | 1,310 | 19.0 |
| Private insurance or Medicare | 44,403 | 74.7 | 33,613 | 71.8 | 6,287 | 72.8 | 4,469 | 65.0 |

. . . Data not applicable.

NOTES: In the March 1996 CPS, health insurance coverage is for the previous calendar year; in the 1995 NHIS, it is for the prior month. CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

**Table 5. Weighted numbers and percent distributions of persons with selected types of health insurance coverage, by sex and respondent-assessed health status: All ages, March 1996 Current Population Survey and 1995 National Health Interview Survey**

| Sex and health insurance coverage | Health status | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Excellent/very good/good | | | | Fair/poor | | | |
| | March 1996 CPS | | 1995 NHIS | | March 1996 CPS | | 1995 NHIS | |
| | Number | Percent distribution | Number | Percent distribution | Number | Percent distribution | Number | Percent distribution |
| Male | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 |
| No insurance | 1,971,255 | 17.0 | 18,901,991 | 16.5 | 1,934,844 | 14.3 | 2,127,156 | 16.5 |
| Public insurance | 9,217,423 | 8.0 | 9,317,592 | 8.1 | 1,842,331 | 13.6 | 1,767,477 | 13.7 |
| Private insurance or Medicare | 86,676,220 | 75.0 | 86,490,839 | 75.4 | 9,759,957 | 72.1 | 8,971,485 | 69.7 |
| Female | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . | 100.0 |
| No insurance | 16,831,253 | 14.3 | 16,014,994 | 13.6 | 2,102,921 | 12.2 | 2,250,237 | 13.8 |
| Public insurance | 11,747,442 | 10.0 | 12,692,034 | 10.8 | 2,516,107 | 14.5 | 2,790,629 | 17.1 |
| Private insurance or Medicare | 89,293,168 | 75.8 | 89,334,702 | 75.7 | 12,679,944 | 73.3 | 11,243,788 | 69.1 |

. . . Data not applicable.

NOTES: In the March 1996 CPS, health insurance coverage is for the previous calendar year; in the 1995 NHIS, it is for the prior month. CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

**Table 6. Frequency distributions of the unweighted cell sizes from the partitioning of the March 1996 Current Population Survey and 1995 National Health Interview Survey**

| Unweighted cell size | Number of cells | |
|---|---|---|
| | March 1996 CPS | 1995 NHIS |
| All cells | 982 | 982 |
| 30–49 | 57 | 110 |
| 50–99 | 292 | 415 |
| 100–149 | 303 | 298 |
| 150–199 | 180 | 104 |
| 200–249 | 102 | 41 |
| 250 and more | 48 | 14 |

NOTE: CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 7. Mean and standard error of number of doctor visits and bed days within the past 12 months, by age: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Mean | Standard error | Mean | Standard error | Mean | Standard error |
| | | | Number of doctor visits | | | |
| All persons | 4.0 | 0.04 | 4.1 | 0.03 | 4.1 | 0.03 |
| Under 18 years | 3.1 | 0.05 | 3.0 | 0.03 | 3.0 | 0.03 |
| 18–64 years | 3.9 | 0.05 | 3.9 | 0.03 | 3.9 | 0.03 |
| 65 years and over | 6.8 | 0.17 | *7.6 | 0.16 | *7.6 | 0.16 |
| | | | Number of bed days | | | |
| All persons | 5.3 | 0.10 | 5.5 | 0.07 | 5.6 | 0.07 |
| Under 18 years | 2.4 | 0.05 | *2.2 | 0.03 | *2.2 | 0.03 |
| 18–64 years | 5.4 | 0.12 | 5.4 | 0.09 | 5.4 | 0.09 |
| 65 years and over | 11.2 | 0.51 | *13.8 | 0.45 | *13.8 | 0.42 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Number of doctor visits within the past 12 months and number of bed days within the past 12 months are variables brought over to the CPS-Host files from the NHIS-Donor file. The age variable used is that originally found on each file (CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file). CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 8. Percentage of persons with no doctor visits within the past 12 months for selected subgroups: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Selected subgroups | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Sex | | | | | | |
| Male | 29.5 | 0.3 | 29.8 | 0.2 | 29.7 | 0.2 |
| Female | 18.5 | 0.2 | 18.7 | 0.1 | 18.7 | 0.2 |
| Age | | | | | | |
| Under 18 years | 19.5 | 0.4 | 19.7 | 0.2 | 19.8 | 0.3 |
| 18–44 years | 30.5 | 0.3 | 30.6 | 0.2 | 30.6 | 0.2 |
| 45–64 years | 22.9 | 0.4 | 24.1 | 0.2 | 24.2 | 0.3 |
| 65 years and over | 12.5 | 0.4 | 11.7 | 0.3 | 11.7 | 0.3 |
| Race and Hispanic origin | | | | | | |
| White, not Hispanic | 22.4 | 0.2 | 22.7 | 0.1 | 22.6 | 0.1 |
| Black, not Hispanic | 23.7 | 0.5 | 24.0 | 0.4 | 24.2 | 0.3 |
| Other, not Hispanic | 30.1 | 1.1 | 31.5 | 0.5 | 29.9 | 0.9 |
| Hispanic | 31.8 | 0.5 | 30.8 | 0.4 | 31.1 | 0.3 |
| Education of head of family | | | | | | |
| Less than high school | #27.9 | 0.4 | *24.7 | 0.3 | *24.7 | 0.3 |
| High school | #26.6 | 0.4 | *24.3 | 0.2 | *24.0 | 0.2 |
| More than high school | #20.7 | 0.3 | *23.8 | 0.2 | *23.9 | 0.2 |

# The pattern observed across these NHIS-Donor subgroups is not observed across the CPS-Host matched file subgroups.

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits during the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Number of doctor visits within the past 12 months is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 9. Percent distribution of activity limitation status among working-age adults: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and activity limitation status | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent distribution | Standard error | Percent distribution | Standard error | Percent distribution | Standard error |
| 18–44 years | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . |
| Cannot perform major activities | 3.2 | 0.1 | 3.0 | 0.1 | 3.2 | 0.1 |
| Limited in some major activities | 3.9 | 0.1 | 3.8 | 0.1 | 3.9 | 0.1 |
| Limited in other activities | 3.0 | 0.1 | 2.9 | 0.1 | 2.9 | 0.1 |
| No activity limitations | 90.0 | 0.2 | 90.3 | 0.1 | 90.1 | 0.1 |
| 45–64 years | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . |
| Cannot perform major activities | 9.5 | 0.3 | 8.8 | 0.2 | 8.9 | 0.1 |
| Limited in some major activities | 7.9 | 0.2 | 7.5 | 0.1 | 7.6 | 0.1 |
| Limited in other activities | 5.3 | 0.2 | 5.4 | 0.1 | 5.5 | 0.1 |
| No activity limitations | 77.3 | 0.4 | 78.2 | 0.3 | 78.0 | 0.3 |

. . . Data not applicable.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Activity limitation status is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. The age variable is that originally found on each file: CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

**Table 10. Percentage of persons with no usual source of health care, by age and health insurance coverage: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and health insurance coverage[1] | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS-Host Match 2 file[3] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 6.5 | 0.2 | 6.5 | 0.1 | 6.6 | 0.1 |
| Not insured | 22.1 | 1.1 | 22.5 | 0.6 | 22.5 | 0.6 |
| Insured | 3.9 | 0.2 | 4.0 | 0.1 | 4.0 | 0.1 |
| 18–44 years | 20.3 | 0.4 | 20.3 | 0.2 | 20.4 | 0.2 |
| Not insured | 43.8 | 0.8 | 44.2 | 0.5 | 44.3 | 0.5 |
| Insured | 14.1 | 0.3 | 13.8 | 0.2 | 13.9 | 0.2 |
| 45–64 years | 11.1 | 0.3 | *12.4 | 0.2 | *12.3 | 0.2 |
| Not insured | 34.9 | 1.1 | 37.1 | 0.9 | 36.6 | 0.8 |
| Insured | 8.0 | 0.3 | 8.6 | 0.2 | 8.5 | 0.2 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]The reference period for health insurance coverage different in the National Health Interview Survey (NHIS) and Current Population Survey (CPS). In the March 1996 CPS, health insurance coverage is ascertained for the previous calendar year; in the 1995 NHIS it is ascertained for the prior month.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[3]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTE: Usual source of health care is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 11. Percentage of persons with no usual source of health care, by age and respondent-assessed health status: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and respondent-assessed health status | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 6.5 | 0.2 | 6.5 | 0.1 | 6.6 | 0.1 |
| Fair or poor | 9.2 | 1.1 | 8.5 | 0.7 | 9.2 | 0.8 |
| Good to excellent | 6.4 | 0.2 | 6.5 | 0.1 | 6.5 | 0.1 |
| 18–44 years | 20.3 | 0.4 | 20.3 | 0.2 | 20.4 | 0.2 |
| Fair or poor | 19.6 | 1.0 | 18.3 | 0.7 | 18.3 | 0.7 |
| Good to excellent | 20.4 | 0.4 | 20.4 | 0.2 | 20.5 | 0.2 |
| 45–64 years | 11.1 | 0.3 | 12.4 | 0.2 | 12.3 | 0.2 |
| Fair or poor | 9.0 | 0.5 | 9.9 | 0.4 | 9.7 | 0.4 |
| Good to excellent | 11.5 | 0.3 | *12.9 | 0.2 | *12.8 | 0.2 |
| 65 years and over | 5.7 | 0.3 | 5.4 | 0.2 | 5.5 | 0.2 |
| Fair or poor | 3.9 | 0.4 | 3.7 | 0.3 | 3.6 | 0.2 |
| Good to excellent | 6.4 | 0.3 | 6.5 | 0.2 | 6.6 | 0.2 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Usual source of health care is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 12. Percentage of persons with no doctor visits within the past 12 months, by age and health insurance coverage: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and health insurance coverage[1] | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS-Host Match 2 file[3] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 19.5 | 0.4 | 20.0 | 0.2 | 19.8 | 0.3 |
| Not insured | 33.1 | 0.9 | 33.6 | 0.6 | 33.2 | 0.6 |
| Insured | 17.2 | 0.4 | 17.8 | 0.2 | 17.6 | 0.3 |
| 18–44 years | 30.5 | 0.3 | 30.6 | 0.2 | 30.6 | 0.2 |
| Not insured | 46.6 | 0.6 | 46.5 | 0.4 | 47.0 | 0.5 |
| Insured | 26.2 | 0.3 | 26.2 | 0.2 | 26.0 | 0.2 |
| 45–64 years | 22.9 | 0.4 | 24.1 | 0.2 | 24.2 | 0.3 |
| Not insured | 38.5 | 1.0 | 41.3 | 0.8 | 40.6 | 0.8 |
| Insured | 20.8 | 0.4 | 21.4 | 0.2 | 21.6 | 0.2 |

[1]The reference period for health insurance coverage differs in the NHIS and CPS. In the March 1996 CPS, health insurance coverage is ascertained for the previous calendar year; in the 1995 NHIS it is ascertained for the prior month.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[3]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Number of doctor visits within the past 12 months is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file, for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 13. Percentage of persons with no doctor visits within the past 12 months, by age and respondent-assessed health status: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and respondent-assessed health status | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 19.5 | 0.4 | 20.0 | 0.2 | 19.8 | 0.3 |
| Fair or poor | 9.6 | 1.0 | *16.6 | 0.9 | *16.7 | 1.0 |
| Good to excellent | 19.7 | 0.4 | 20.1 | 0.2 | 19.9 | 0.3 |
| 18–44 years | 30.5 | 0.3 | 30.6 | 0.2 | 30.6 | 0.2 |
| Fair or poor | 16.9 | 0.8 | 16.9 | 0.7 | 16.1 | 0.6 |
| Good to excellent | 31.5 | 0.3 | 31.6 | 0.2 | 31.6 | 0.2 |
| 45–64 years | 22.9 | 0.4 | 24.1 | 0.2 | 24.2 | 0.3 |
| Fair or poor | 10.1 | 0.5 | 10.4 | 0.5 | 11.0 | 0.4 |
| Good to excellent | 25.4 | 0.4 | *26.9 | 0.3 | *26.9 | 0.3 |
| 65 years and over | 12.5 | 0.4 | 11.7 | 0.3 | 11.7 | 0.3 |
| Fair or poor | 6.5 | 0.4 | 6.0 | 0.3 | 6.0 | 0.3 |
| Good to excellent | 14.8 | 0.4 | 15.3 | 0.3 | 15.2 | 0.4 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Number of doctor visits within the past 12 months is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 14. Percentage of persons with no usual source of health care, by age, race, and Hispanic origin: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age, race, and Hispanic origin | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 6.5 | 0.2 | 6.5 | 0.1 | 6.6 | 0.1 |
| White, not Hispanic | 4.9 | 0.3 | 4.6 | 0.1 | 4.6 | 0.1 |
| Black, not Hispanic | 7.3 | 0.6 | 7.9 | 0.4 | 8.0 | 0.5 |
| Other, not Hispanic | 7.1 | 1.4 | 8.4 | 0.6 | 8.7 | 0.7 |
| Hispanic | 12.8 | 0.8 | 13.1 | 0.4 | 13.1 | 0.4 |
| 18–44 years | 20.3 | 0.4 | 20.3 | 0.2 | 20.4 | 0.2 |
| White, not Hispanic | 18.5 | 0.4 | 17.8 | 0.2 | 18.1 | 0.2 |
| Black, not Hispanic | 21.5 | 0.8 | 23.7 | 0.6 | 23.3 | 0.3 |
| Other, not Hispanic | 22.8 | 1.5 | 21.8 | 0.9 | 20.2 | 0.9 |
| Hispanic | 29.8 | 0.8 | 31.3 | 0.6 | 31.1 | 0.6 |
| 45–64 years | 11.1 | 0.3 | 12.4 | 0.2 | 12.3 | 0.2 |
| White, not Hispanic | 10.3 | 0.3 | 11.3 | 0.2 | 11.4 | 0.2 |
| Black, not Hispanic | 10.8 | 0.8 | *13.7 | 0.7 | *14.8 | 0.7 |
| Other, not Hispanic | 13.4 | 1.9 | 18.9 | 1.6 | 16.0 | 1.2 |
| Hispanic | 19.2 | 1.1 | 18.6 | 0.8 | 16.7 | 0.8 |
| 65 years and over | 5.7 | 0.3 | 5.4 | 0.2 | 5.4 | 0.2 |
| White, not Hispanic | 5.4 | 0.3 | 5.2 | 0.2 | 5.2 | 0.2 |
| Black, not Hispanic | 6.9 | 1.0 | 6.1 | 0.6 | 6.1 | 0.6 |
| Other, not Hispanic | 8.1 | 2.8 | 7.8 | 1.2 | 8.1 | 1.4 |
| Hispanic | 7.5 | 0.8 | 7.5 | 0.8 | 7.8 | 0.7 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Usual source of health care is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 15. Percentage of persons with no usual source of health care by age and education of head of family: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and education of head of family | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 6.5 | 0.2 | 6.5 | 0.1 | 6.6 | 0.1 |
| Less than high school | #12.7 | 0.8 | 10.8 | 0.4 | 11.6 | 0.5 |
| High school | #7.6 | 0.5 | 6.8 | 0.2 | 7 | 0.3 |
| More than high school | #3.9 | 0.2 | *5.0 | 0.2 | *4.8 | 0.2 |
| 18–44 years | 20.3 | 0.4 | 20.3 | 0.2 | 20.4 | 0.2 |
| Less than high school | #32.3 | 0.9 | 29.5 | 0.6 | 29.5 | 0.6 |
| High school | #21.9 | 0.6 | 21.7 | 0.4 | 21.7 | 0.4 |
| More than high school | #16.9 | 0.4 | *17.7 | 0.2 | *17.8 | 0.2 |
| 45–64 years | 11.1 | 0.3 | *12.4 | 0.2 | *12.3 | 0.2 |
| Less than high school | #17.2 | 0.9 | 16.6 | 0.5 | 15.5 | 0.6 |
| High school | #12.1 | 0.5 | 12.7 | 0.4 | 12.4 | 0.4 |
| More than high school | #8.9 | 0.4 | *11.2 | 0.2 | *11.4 | 0.3 |
| 65 years and over | 5.7 | 0.3 | 5.4 | 0.2 | 5.5 | 0.2 |
| Less than high school | #7.3 | 0.5 | 5.1 | 0.3 | 5.1 | 0.3 |
| High school | #5.2 | 0.4 | 5.2 | 0.3 | 5.3 | 0.3 |
| More than high school | #4.9 | 0.5 | 5.8 | 0.3 | 5.9 | 0.3 |

# The associations observed across the education subgroups in the NHIS-Donor file are weakened or lost in the CPS-Host matched files.

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Usual source of health care is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 16. Percentage of persons with no doctor visits in the past 12 months, by age and education of head of family: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and education of head of family | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 19.5 | 0.4 | 20.0 | 0.2 | 19.8 | 0.3 |
| Less than high school | #25.9 | 0.9 | *22.7 | 0.5 | 22.8 | 0.5 |
| High school | #22.9 | 0.6 | *20.5 | 0.4 | 20.5 | 0.4 |
| More than high school | #15.2 | 0.5 | *18.9 | 0.3 | 18.5 | 0.3 |
| 18–44 years | 30.5 | 0.3 | 30.6 | 0.2 | 30.6 | 0.2 |
| Less than high school | #41.2 | 0.9 | *35.4 | 0.6 | *35.3 | 0.6 |
| High school | #34.1 | 0.5 | *31.0 | 0.4 | *30.7 | 0.4 |
| More than high school | #26.0 | 0.3 | *29.4 | 0.3 | *29.4 | 0.3 |
| 45–64 years | 22.9 | 0.4 | *24.1 | 0.2 | *24.2 | 0.3 |
| Less than high school | #26.5 | 0.9 | *23.3 | 0.7 | *23.3 | 0.7 |
| High school | #24.7 | 0.6 | 24.7 | 0.5 | 23.8 | 0.5 |
| More than high school | #20.5 | 0.5 | *24.0 | 0.3 | *24.6 | 0.3 |
| 65 years and over | 12.5 | 0.4 | 11.7 | 0.3 | 11.7 | 0.3 |
| Less than high school | 12.6 | 0.6 | 11.1 | 0.5 | 11.2 | 0.5 |
| High school | 13.8 | 0.8 | 11.8 | 0.4 | 11.6 | 0.4 |
| More than high school | 10.9 | 0.5 | 12.1 | 0.5 | 12.1 | 0.4 |

# The associations observed across the education subgroups in the NHIS-Donor file are weakened or lost in the CPS-Host matched files.

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Number of doctor visits within the past 12 months is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 17. Percentage of persons with no usual source of health care, by age and percent of poverty level: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and percent of poverty level[1] | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS-Host Match 2 file[3] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 6.5 | 0.2 | 6.5 | 0.1 | 6.5 | 0.1 |
| Below 100% | 10.9 | 0.6 | 9.5 | 0.4 | 9.9 | 0.4 |
| 100% to less than 150% | 9.7 | 0.7 | 9.6 | 0.4 | 10.3 | 0.5 |
| 150% to less than 200% | 6.9 | 0.8 | 7.4 | 0.5 | 8.5 | 0.5 |
| 200% or more | 3.6 | 0.2 | *4.7 | 0.1 | *4.2 | 0.1 |
| 18–44 years | 20.3 | 0.4 | 20.3 | 0.3 | 20.4 | 0.2 |
| Below 100% | 28.5 | 0.9 | 27.4 | 0.6 | 26.0 | 0.6 |
| 100% to less than 150% | 26.5 | 0.9 | 26.2 | 0.7 | 25.6 | 0.6 |
| 150% to less than 200% | 24.3 | 0.9 | 24.1 | 0.6 | 24.1 | 0.6 |
| 200% or more | 16.7 | 0.4 | *17.8 | 0.2 | *18.2 | 0.2 |
| 45–64 years | 11.1 | 0.3 | 12.4 | 0.2 | 12.3 | 0.2 |
| Below 100% | 19.0 | 1.1 | 18.1 | 0.8 | 16.8 | 0.7 |
| 100% to less than 150% | 17.1 | 1.0 | 17.2 | 0.8 | 16.0 | 0.8 |
| 150% to less than 200% | 15.7 | 1.1 | 16.6 | 0.9 | 15.3 | 0.8 |
| 200% or more | 9.2 | 0.3 | *11.0 | 0.2 | *11.2 | 0.2 |
| 65 years and over | 5.7 | 0.3 | 5.4 | 0.2 | 5.5 | 0.2 |
| Below 100% | #9.3 | 0.7 | *5.5 | 0.4 | *5.5 | 0.5 |
| 100% to less than 150% | #6.2 | 0.6 | 4.9 | 0.4 | 5.5 | 0.4 |
| 150% to less than 200% | #5.5 | 0.6 | 4.9 | 0.4 | 5.6 | 0.5 |
| 200% or more | #4.6 | 0.4 | 5.7 | 0.2 | 5.4 | 0.2 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

# The associations observed across the education subgroups in the NHIS-Donor file are weakened or lost in the CPS-Host matched files.

[1]Percentage of poverty level is the ratio of family income to poverty thresholds.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[3]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Usual source of health care is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. The percentage of poverty level on the CPS-Host files was calculated using CPS total annual family income; on the NHIS-Donor file it was calculated using NHIS total annual family income. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 18. Percentage of persons with no doctor visits within the past 12 months, by age and percent of poverty level: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and percent of poverty level[1] | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS-Host Match 2 file[3] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 19.5 | 0.4 | 20.0 | 0.2 | 19.8 | 0.3 |
| Below 100% | #23.5 | 0.9 | 21.4 | 0.5 | 20.9 | 0.5 |
| 100% to less than 150% | #24.7 | 0.9 | 22.6 | 0.6 | 22.1 | 0.6 |
| 150% to less than 200% | #22.8 | 1.1 | 20.9 | 0.7 | 22.2 | 0.7 |
| 200% or more | #15.5 | 0.4 | *18.7 | 0.3 | *18.5 | 0.3 |
| 18–44 years | 30.5 | 0.3 | 30.6 | 0.2 | 30.6 | 0.2 |
| Below 100% | #33.1 | 0.9 | 31.9 | 0.6 | 32.0 | 0.6 |
| 100% to less than 150% | #36.3 | 0.8 | *33.4 | 0.6 | *32.2 | 0.6 |
| 150% to less than 200% | #34.9 | 0.9 | 32.5 | 0.7 | 32.1 | 0.8 |
| 200% or more | #28.1 | 0.3 | *29.8 | 0.2 | *29.9 | 0.2 |
| 45–64 years | 22.9 | 0.4 | *24.1 | 0.2 | *24.2 | 0.2 |
| Below 100% | #23.4 | 1.0 | 23.7 | 0.7 | 24.4 | 0.8 |
| 100% to less than 150% | #27.1 | 1.2 | 24.6 | 0.8 | 24.6 | 0.9 |
| 150% to less than 200% | #26.2 | 1.2 | 23.3 | 1.0 | 22.8 | 0.9 |
| 200% or more | #22.0 | 0.4 | *24.2 | 0.3 | *24.2 | 0.3 |
| 65 years and over | 12.5 | 0.4 | 11.7 | 0.3 | 11.7 | 0.3 |
| Below 100% | #14.7 | 0.9 | 10.3 | 0.6 | 11.1 | 0.6 |
| 100 to less than 150% | #13.1 | 0.7 | *10.3 | 0.6 | *11.0 | 0.6 |
| 150 to less than 200% | #13.1 | 0.9 | *10.8 | 0.6 | *12.7 | 0.7 |
| 200% or more | #11.5 | 0.5 | 12.7 | 0.4 | 11.7 | 0.4 |

# The associations observed across the education subgroups in the NHIS-Donor file are weakened or lost in the CPS-Host matched files.

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Percent of poverty level is the ratio of family income to poverty thresholds.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[3]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTE: Number of doctor visits within the past 12 months is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. The percentage of poverty level on the CPS-Host files was calculated using CPS total annual family income; on the NHIS-Donor file it was calculated using NHIS total annual family income. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 19. Percent distribution of health insurance coverage among persons who report that they cannot perform major activities, by age: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and health insurance coverage[1] | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS-Host Match 2 file[3] | |
|---|---|---|---|---|---|---|
| | Percent distribution | Standard error | Percent distribution | Standard error | Percent distribution | Standard error |
| 18–44 years | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . |
| Not insured | 19.1 | 1.2 | 19.1 | 1.0 | 19.4 | 0.9 |
| Insured | 80.9 | 1.2 | 80.9 | 1.0 | 80.6 | 0.9 |
| 45–64 years | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . |
| Not insured | 13.7 | 0.8 | 15.5 | 0.7 | 15.5 | 0.8 |
| Insured | 86.3 | 0.8 | 84.5 | 0.7 | 84.5 | 0.8 |

. . . Data not applicable.

[1]The reference period for health insurance coverage differed in the NHIS and CPS. In the March 1996 CPS, health insurance coverage is ascertained for the previous calendar year; in the 1995 NHIS it is ascertained for the prior month.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[3]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Activity limitation status is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All of the other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 20. Percent distribution of poverty status among working-age adults who reported that they cannot perform major activities: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and percent of poverty level[1] | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS-Host Match 2 file[3] | |
|---|---|---|---|---|---|---|
| | Percent distribution | Standard error | Percent distribution | Standard error | Percent distribution | Standard error |
| 18–44 years of age | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . |
| Below 100% | 35.0 | 1.4 | *26.4 | 1.2 | *25.5 | 1.3 |
| 100% to less than 150% | 21.4 | 1.4 | *14.8 | 1.0 | *15.3 | 1.0 |
| 150% to less than 200% | 9.2 | 0.9 | *10.9 | 0.9 | *9.5 | 0.8 |
| 200% or more | 34.4 | 1.4 | *48.0 | 1.4 | *49.7 | 1.3 |
| 45–64 years of age | 100.0 | . . . | 100.0 | . . . | 100.0 | . . . |
| Below 100% | 29.6 | 1.3 | *22.5 | 0.9 | *20.7 | 0.8 |
| 100% to less than 150% | 19.0 | 0.9 | *12.4 | 0.7 | *11.4 | 0.7 |
| 150% to less than 200% | 13.0 | 0.9 | *9.9 | 0.6 | *9.0 | 0.6 |
| 200% or more | 38.5 | 1.5 | *55.2 | 1.0 | *58.9 | 1.0 |

. . . Data not applicable.

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Percent of poverty level is the ratio of family income to poverty thresholds.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[3]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Activity limitation status is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. The percentage of poverty level on the CPS-Host files was calculated using CPS total annual family income; on the NHIS-Donor file it was calculated using NHIS total annual family income. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 21. Percentage of working-age adults who receive Supplemental Security Income, by age and activity limitation status: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and activity limitation status | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| 18–44 years | 1.3 | 0.1 | *1.7 | 0.1 | *1.7 | 0.1 |
| Cannot perform major activities | 24.1 | 1.4 | *11.4 | 0.9 | *11.0 | 0.8 |
| Limited in some major activities | 5.1 | 0.6 | 5.4 | 0.5 | 4.1 | 0.4 |
| Limited in other activities | 3.2 | 0.5 | 3.7 | 0.5 | 4.0 | 0.5 |
| No limitations | 0.3 | 0.0 | *1.1 | 0.1 | *1.2 | 0.1 |
| 45–64 years | 2.5 | 0.1 | 2.8 | 0.1 | 2.8 | 0.1 |
| Cannot perform major activities | 17.3 | 1.0 | *13.1 | 0.7 | *12.3 | 0.7 |
| Limited in some major activities | 5.0 | 0.6 | 6.1 | 0.5 | 6.4 | 0.5 |
| Limited in other activities | 4.6 | 0.6 | 4.6 | 0.5 | 4.4 | 0.4 |
| No limitations | 0.3 | 0.0 | *1.2 | 0.1 | *1.3 | 0.1 |

0.0 Quantity more than zero but less than 0.05.

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Activity limitation status is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 22. Supplemental Security Income recipiency among National Health Interview Survey and Current Population Survey respondents 18–64 years of age who cannot perform major activities and among Current Population Survey respondents 18–64 years of age who did not work during the prior calendar year due to disability or illness, by age: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched file, and March 1996 Current Population Survey**

| Age | Cannot perform major activities | | | | Did not work in prior year due to disability or illness[1] | |
|---|---|---|---|---|---|---|
| | 1995 NHIS | | March 1996 CPS-Host Match 1 file[2] | | March 1996 CPS | |
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| 18–44 years | 24.1 | 1.4 | *11.4 | 0.9 | 22.0 | 1.0 |
| 45–64 years | 17.3 | 1.0 | *13.1 | 0.7 | 18.1 | 0.7 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]"Yes" answer to the CPS question: Does . . . have a health problem or a disability which prevents work or which limits the kind or amount of work?

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

NOTES: Activity limitation status is from the 1995 NHIS. Age and SSI are those originally found on each file; for example, CPS age for the CPS file and CPS-Host matched file and NHIS age for the NHIS file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 23. Supplemental Security Income recipiency among working-age adults who report that they cannot perform major activities, by age, race, and Hispanic origin: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age, race, and Hispanic origin | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| 18–44 years | 24.1 | 1.4 | 11.4 | 0.8 | 11.0 | 0.8 |
| White, not Hispanic | 22.8 | 1.9 | 10.2 | 1.1 | 9.2 | 0.9 |
| Black, not Hispanic | 33.0 | 2.7 | 14.7 | 1.8 | 16.5 | 2.5 |
| Hispanic | 16.2 | 2.5 | 12.0 | 2.0 | 11.6 | 2.0 |
| 45–64 years | 17.3 | 1.0 | 13.1 | 0.7 | 13.1 | 0.7 |
| White, not Hispanic | 13.7 | 1.1 | 10.6 | 0.8 | 9.9 | 0.8 |
| Black, not Hispanic | 28.7 | 2.8 | 21.4 | 2.0 | 23.8 | 2.1 |
| Hispanic | 22.7 | 2.3 | 17.0 | 1.7 | 17.3 | 1.7 |

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Activity limitation status is the variable brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched files and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 24. Percentage of persons with no usual source of health care, by age and activity limitation status: 1995 National Health Interview Survey and March 1996 Current Population Survey-Host matched files**

| Age and activity limitation status | 1995 NHIS | | March 1996 CPS-Host Match 1 file[1] | | March 1996 CPS-Host Match 2 file[2] | |
|---|---|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error | Percent | Standard error |
| Under 18 years | 6.5 | 0.2 | 6.5 | 0.1 | 6.6 | 0.1 |
| Cannot perform major activities | 6.1 | 1.8 | 5.7 | 1.3 | 6.6 | 1.6 |
| Limited in some major activities | 6.6 | 1.1 | 6.5 | 0.7 | 6.7 | 0.9 |
| Limited in other activities | 3.9 | 1.2 | 3.1 | 0.7 | 3.2 | 0.7 |
| No limitations | 6.5 | 0.3 | 6.6 | 0.1 | 6.6 | 0.1 |
| 18–44 years | 20.3 | 0.4 | 20.3 | 0.2 | 20.4 | 0.2 |
| Cannot perform major activities | 13.4 | 1.1 | 12.7 | 0.9 | 13.1 | 0.9 |
| Limited in some major activities | 17.2 | 1.1 | 16.3 | 0.8 | 16.3 | 0.9 |
| Limited in other activities | 16.5 | 1.4 | 15.9 | 0.9 | 16.9 | 0.9 |
| No limitations | 20.8 | 0.4 | 20.9 | 0.2 | 20.9 | 0.2 |
| 45–64 years | 11.1 | 0.3 | *12.4 | 0.2 | *12.3 | 0.2 |
| Cannot perform major activities | 7.9 | 0.8 | 8.6 | 0.5 | 8.1 | 0.5 |
| Limited in some major activities | 7.7 | 0.8 | 8.9 | 0.6 | 8.8 | 0.6 |
| Limited in other activities | 7.8 | 0.8 | 8.5 | 0.6 | 7.8 | 0.7 |
| No limitations | 12.1 | 0.3 | *13.4 | 0.2 | *13.4 | 0.2 |
| 65 years and over | 5.7 | 0.3 | 5.4 | 0.2 | 5.5 | 0.2 |
| Cannot perform major activities | 3.8 | 0.5 | 3.6 | 0.4 | 3.8 | 0.4 |
| Limited in some major activities | 4.4 | 0.7 | 4.0 | 0.4 | 4.2 | 0.4 |
| Limited in other activities | 4.1 | 0.5 | 3.8 | 0.4 | 3.9 | 0.3 |
| No limitations | 6.6 | 0.3 | 6.5 | 0.2 | 6.6 | 0.2 |

* The matched file value differs significantly from the Donor file value, $p < 0.05$.

[1]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

[2]Predicted total annual family income is the variable used to perform the predictive mean match.

NOTES: Activity limitation status and usual source of health care are the variables brought over to the CPS-Host matched files from the NHIS-Donor file. All other variables are those originally found on each file; for example, CPS age for the CPS-Host matched file and NHIS age for the NHIS-Donor file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 25. Mean and standard error of selected sources of income (in dollars), by age: March 1996 Current Population Survey and 1995 National Health Interview Survey-Host matched file**

| Age and sources of income | March 1996 CPS | | 1995 NHIS-Host matched file[1] | |
|---|---|---|---|---|
| | Mean | Standard error | Mean | Standard error |
| All persons | | | | |
| Family income[2] | 48,503 | 256 | 48,406 | 401 |
| Wages and salary | 13,030 | 69 | 13,076 | 110 |
| Dividends | 237 | 8 | 240 | 6 |
| 18–64 years | | | | |
| Family income[2] | 51,612 | 268 | 51,736 | 415 |
| Wages and salary | 20,741 | 106 | 20,786 | 156 |
| Dividends | 255 | 11 | 261 | 8 |
| 65 years and over | | | | |
| Family income[2] | 32,668 | 399 | 32,686 | 499 |
| Social Security | 7,178 | 43 | 7,093 | 34 |
| Dividends | 667 | 31 | 664 | 25 |

[1]Predicted total annual family income is the variable used to perform the predictive mean match.

[2]Total annual family income.

NOTES: Total annual family income, wages and salary, Social Security, and dividends are the variables brought over to the NHIS-Host matched file from the CPS-Donor file. The age variable used is that originally found on each file; CPS age for the CPS-Donor file, NHIS age for the NHIS-Host matched file. CPS is Current Population Survey and NHIS is National Health Interview Survey.

**Table 26. Percentage of persons with no health insurance coverage, by age and percent of poverty level: March 1996 Current Population Survey and 1995 National Health Interview Survey-Host matched file**

| Age and percent of poverty level[1] | March 1996 CPS | | 1995 NHIS-Host matched file[2] | |
|---|---|---|---|---|
| | Percent | Standard error | Percent | Standard error |
| Under 18 years | 13.8 | 0.3 | 14.1 | 0.4 |
| Below 100% | 21.2 | 1.0 | 21.2 | 0.9 |
| 100% to less than 150% | 24.3 | 1.0 | 23.1 | 0.9 |
| 150% to less than 200% | 20.0 | 1.1 | 18.4 | 1.0 |
| 200% or more | 7.7 | 0.3 | 8.7 | 0.3 |
| 18–44 years | 21.8 | 0.2 | 21.3 | 0.4 |
| Below 100% | 44.1 | 0.8 | 43.1 | 0.8 |
| 100% to less than 150% | 39.5 | 0.9 | 33.3 | 0.9 |
| 150% to less than 200% | 33.8 | 0.9 | 31.0 | 1.0 |
| 200% or more | 13.7 | 0.2 | 13.9 | 0.3 |
| 45–64 years | 13.5 | 0.2 | 11.8 | 0.3 |
| Below 100% | 36.6 | 1.1 | 30.1 | 1.1 |
| 100% to less than 150% | 31.4 | 1.2 | 23.8 | 1.2 |
| 150% to less than 200% | 25.3 | 1.3 | 18.9 | 1.0 |
| 200% or more | 8.5 | 0.3 | 7.3 | 0.3 |

[1]Percent of poverty level is the ratio of family income to poverty thresholds.

[2]Predicted number of doctor visits within the past 12 months is the variable used to perform the predictive mean match.

NOTES: Percent of poverty level was calculated using the CPS total annual family income variable brought over to the NHIS-Host matched from the CPS-Donor file. All other variables used are those originally found on each file; for example, CPS age for the CPS-Donor file, NHIS age for the NHIS-Host matched file. Note that the reference period for health insurance coverage differs in the NHIS and CPS. In the March 1996 CPS, health insurance coverage is ascertained for the previous calendar year; in the 1995 NHIS it is ascertained for the prior month. The percent of poverty level is calculated for the CPS file using the CPS total annual family income variable, while for the NHIS-Host file it is calculated using the NHIS total annual family income variable. CPS is Current Population Survey and NHIS is National Health Interview Survey.

# Appendix I

## Detailed Tables Comparing the Current Population Survey and the National Health Interview Survey

**Table I. Unweighted record counts on the March 1996 Current Population Survey, by sex, age, race, and Hispanic origin**

| Sex and age | White, not Hispanic | Black, not Hispanic | Other, not Hispanic | Hispanic | Total |
|---|---|---|---|---|---|
| All persons | | | | | |
| All ages | 92,436 | 12,893 | 5,786 | 19,361 | 130,476 |
| Under 18 years | 23,318 | 4,517 | 1,811 | 7,045 | 36,691 |
| 18–24 years | 7,362 | 1,269 | 620 | 2,264 | 11,515 |
| 25–34 years | 13,134 | 1,811 | 916 | 3,462 | 19,323 |
| 35–44 years | 15,249 | 1,894 | 930 | 2,906 | 20,979 |
| 45–54 years | 12,096 | 1,353 | 683 | 1,596 | 15,728 |
| 55–64 years | 8,084 | 936 | 425 | 1,053 | 10,498 |
| 65 years and over | 13,193 | 1,113 | 401 | 1,035 | 15,742 |
| Male | | | | | |
| All ages | 44,774 | 5,642 | 2,749 | 9,259 | 62,424 |
| Under 18 years | 11,864 | 2,225 | 912 | 3,520 | 18,521 |
| 18–24 years | 3,590 | 528 | 299 | 1,105 | 5,522 |
| 25–34 years | 6,412 | 728 | 439 | 1,666 | 9,245 |
| 35–44 years | 7,503 | 774 | 431 | 1,371 | 10,079 |
| 45–54 years | 5,930 | 597 | 327 | 719 | 7,573 |
| 55–64 years | 3,924 | 391 | 180 | 463 | 4,958 |
| 65 years and over | 5,551 | 399 | 161 | 415 | 6,526 |
| Female | | | | | |
| All ages | 47,662 | 7,251 | 3,037 | 10,102 | 68,052 |
| Under 18 years | 11,454 | 2,292 | 899 | 3,525 | 18,170 |
| 18–24 years | 3,772 | 741 | 321 | 1,159 | 5,993 |
| 25–34 years | 6,722 | 1,083 | 477 | 1,796 | 10,078 |
| 35–44 years | 7,746 | 1,120 | 499 | 1,535 | 10,900 |
| 45–54 years | 6,166 | 756 | 356 | 877 | 8,155 |
| 55–64 years | 4,160 | 545 | 245 | 590 | 5,540 |
| 65 years and over | 7,642 | 714 | 240 | 620 | 9,216 |

**Table II. Weighted record counts on the March 1996 Current Population Survey, by sex, age, race, and Hispanic origin**

| Sex and age | White, not Hispanic | Black, not Hispanic | Other, not Hispanic | Hispanic | Total |
|---|---|---|---|---|---|
| **All persons** | | | | | |
| All ages | 91,271,018 | 33,073,165 | 11,532,230 | 28,437,751 | 264,314,164 |
| Under 18 years | 46,009,561 | 11,215,844 | 3,615,737 | 10,306,592 | 71,147,733 |
| 18–24 years | 16,451,079 | 3,540,043 | 1,261,826 | 3,590,370 | 24,843,318 |
| 25–34 years | 28,357,405 | 5,211,261 | 1,994,764 | 5,355,169 | 40,918,599 |
| 35–44 years | 31,888,074 | 5,194,115 | 1,930,649 | 4,064,818 | 43,077,655 |
| 45–54 years | 24,750,189 | 3,393,798 | 1,268,964 | 2,171,157 | 31,584,109 |
| 55–64 years | 16,780,941 | 2,073,555 | 738,073 | 1,491,795 | 21,084,364 |
| 65 years and over | 27,033,770 | 2,444,549 | 722,219 | 1,457,850 | 31,658,388 |
| **Male** | | | | | |
| All ages | 93,712,074 | 15,443,049 | 5,610,072 | 14,378,134 | 129,143,329 |
| Under 18 years | 23,606,955 | 5,675,931 | 1,826,756 | 5,292,300 | 36,401,942 |
| 18–24 years | 8,284,771 | 1,643,899 | 617,148 | 1,856,509 | 12,402,327 |
| 25–34 years | 14,131,528 | 2,342,761 | 1,004,045 | 2,911,831 | 20,390,165 |
| 35–44 years | 15,962,087 | 2,384,198 | 904,804 | 2,021,650 | 21,272,740 |
| 45–54 years | 12,109,704 | 1,586,929 | 626,463 | 1,000,549 | 15,323,646 |
| 55–64 years | 8,172,009 | 916,250 | 322,590 | 681,380 | 10,092,229 |
| 65 years and over | 11,445,020 | 893,081 | 308,267 | 613,914 | 13,260,281 |
| **Female** | | | | | |
| All ages | 97,558,944 | 17,630,116 | 5,922,158 | 14,059,617 | 135,170,835 |
| Under 18 years | 22,402,606 | 5,539,913 | 1,788,980 | 5,014,292 | 34,745,791 |
| 18–24 years | 8,166,308 | 1,896,144 | 644,678 | 1,733,861 | 12,440,991 |
| 25–34 years | 14,225,877 | 2,868,500 | 990,719 | 2,443,338 | 20,528,434 |
| 35–44 years | 15,925,986 | 2,809,916 | 1,025,845 | 2,043,168 | 21,804,915 |
| 45–54 years | 12,640,485 | 1,806,869 | 642,501 | 1,170,608 | 16,260,463 |
| 55–64 years | 8,608,931 | 1,157,305 | 415,483 | 810,415 | 10,992,135 |
| 65 years and over | 15,588,751 | 1,551,468 | 413,952 | 843,936 | 18,398,107 |

**Table III. Unweighted record counts on the 1995 National Health Interview Survey, by sex, age, race, and Hispanic origin**

| Sex and age | White, not Hispanic | Black, not Hispanic | Other, not Hispanic | Hispanic | Total |
|---|---|---|---|---|---|
| All persons | | | | | |
| All ages | 65,021 | 13,363 | 3,770 | 20,313 | 102,467 |
| Under 18 years | 15,968 | 4,686 | 1,030 | 8,027 | 29,711 |
| 18–24 years | 5,094 | 1,238 | 403 | 2,363 | 9,098 |
| 25–34 years | 9,221 | 1,967 | 693 | 3,457 | 15,338 |
| 35–44 years | 10,944 | 2,008 | 618 | 2,795 | 16,365 |
| 45–54 years | 8,489 | 1,412 | 479 | 1,595 | 11,975 |
| 55–64 years | 5,822 | 911 | 265 | 1,027 | 8,025 |
| 65 years and over | 9,483 | 1,141 | 282 | 1,049 | 11,955 |
| Male | | | | | |
| All ages | 31,155 | 6,007 | 1,812 | 9,825 | 48,809 |
| Under 18 years | 8,065 | 2,376 | 546 | 4,042 | 15,029 |
| 18–24 years | 2,462 | 532 | 202 | 1,153 | 4,349 |
| 25–34 years | 4,463 | 810 | 331 | 1,646 | 7,250 |
| 35–44 years | 5,265 | 841 | 280 | 1,342 | 7,728 |
| 45–54 years | 4,180 | 592 | 212 | 725 | 5,709 |
| 55–64 years | 2,756 | 393 | 122 | 490 | 3,761 |
| 65 years and over | 3,974 | 463 | 119 | 427 | 4,983 |
| Female | | | | | |
| All ages | 33,856 | 7,356 | 1,958 | 10,488 | 53,658 |
| Under 18 years | 7,903 | 2,310 | 484 | 3,985 | 14,682 |
| 18–24 years | 2,632 | 706 | 201 | 1,210 | 4,749 |
| 25–34 years | 4,758 | 1,157 | 362 | 1,811 | 8,088 |
| 35–44 years | 5,679 | 1,167 | 338 | 1,453 | 8,637 |
| 45–54 years | 4,309 | 820 | 267 | 870 | 6,266 |
| 55–64 years | 3,066 | 518 | 143 | 537 | 4,264 |
| 65 years and over | 5,509 | 678 | 163 | 622 | 6,972 |

**Table IV. Weighted record counts on the 1995 National Health Interview Survey, by sex, age, race, and Hispanic origin**

| Sex and age | White, not Hispanic | Black, not Hispanic | Other, not Hispanic | Hispanic | Total |
|---|---|---|---|---|---|
| **All persons** | | | | | |
| All ages | 91,597,612 | 32,010,133 | 11,362,497 | 26,932,682 | 261,902,924 |
| Under 18 years | 47,006,717 | 10,864,771 | 3,059,403 | 9,743,745 | 70,674,636 |
| 18–24 years | 16,852,088 | 3,488,097 | 1,306,513 | 3,282,641 | 24,929,339 |
| 25–34 years | 28,476,930 | 5,057,504 | 2,200,887 | 5,056,807 | 40,792,128 |
| 35–44 years | 31,721,186 | 4,958,162 | 1,824,294 | 3,820,446 | 42,324,088 |
| 45–54 years | 24,176,550 | 3,118,243 | 1,380,840 | 2,241,790 | 30,917,423 |
| 55–64 years | 16,640,734 | 2,011,784 | 786,173 | 1,360,205 | 20,798,896 |
| 65 years and over | 26,723,407 | 2,511,572 | 804,387 | 1,427,048 | 31,466,414 |
| **Male** | | | | | |
| All ages | 93,514,997 | 14,889,113 | 5,614,825 | 13,557,605 | 127,576,540 |
| Under 18 years | 24,034,432 | 5,493,528 | 1,660,836 | 4,975,296 | 36,164,092 |
| 18–24 years | 8,404,650 | 1,616,725 | 676,910 | 1,695,136 | 12,393,421 |
| 25–34 years | 14,093,182 | 2,255,079 | 1,070,001 | 2,655,398 | 20,073,660 |
| 35–44 years | 15,775,951 | 2,260,015 | 869,899 | 1,926,348 | 20,832,213 |
| 45–54 years | 11,976,326 | 1,403,900 | 616,707 | 1,082,097 | 15,079,030 |
| 55–64 years | 8,011,301 | 873,776 | 378,185 | 631,371 | 9,894,633 |
| 65 years and over | 11,219,155 | 986,090 | 342,287 | 591,959 | 13,139,491 |
| **Female** | | | | | |
| All ages | 98,082,615 | 17,121,020 | 5,747,672 | 13,375,077 | 134,326,384 |
| Under 18 years | 22,972,285 | 5,371,243 | 1,398,567 | 4,768,449 | 34,510,544 |
| 18–24 years | 8,447,438 | 1,871,372 | 629,603 | 1,587,505 | 12,535,918 |
| 25–34 years | 14,383,748 | 2,802,425 | 1,130,886 | 2,401,409 | 20,718,468 |
| 35–44 years | 15,945,235 | 2,698,147 | 954,395 | 1,894,098 | 21,491,875 |
| 45–54 years | 12,200,224 | 1,714,343 | 764,133 | 1,159,693 | 15,838,393 |
| 55–64 years | 8,629,433 | 1,138,008 | 407,988 | 728,834 | 10,904,263 |
| 65 years and over | 15,504,252 | 1,525,482 | 462,100 | 835,089 | 18,326,923 |

**Table V. Poverty status and percent distribution, by age: March 1996 Current Population Survey and the 1995 National Health Interview Survey**

| Age group and percent of poverty level[1] | March 1996 CPS | | 1995 NHIS | |
|---|---|---|---|---|
| | Percent distribution | Standard error | Percent distribution | Standard error |
| Under 18 years | 100.0 | . . . | 100.0 | . . . |
| Below 100% | 19.8 | 0.4 | 21.6 | 0.5 |
| 100% to less than 150% | 12.7 | 0.3 | 15.2 | 0.4 |
| 150% to less than 200% | 9.6 | 0.3 | 10.6 | 0.4 |
| 200% or more | 57.9 | 0.5 | 52.6 | 0.8 |
| 18–44 years | 100.0 | . . . | 100.0 | . . . |
| Below 100% | 12.4 | 0.2 | 14.5 | 0.5 |
| 100% to less than 150% | 9.4 | 0.2 | 11.7 | 0.3 |
| 150% to less than 200% | 9.1 | 0.2 | 10.0 | 0.3 |
| 200% or more | 69.1 | 0.3 | 63.8 | 0.6 |
| 45–64 years | 100.0 | . . . | 100.0 | . . . |
| Below 100% | 8.5 | 0.2 | 8.2 | 0.3 |
| 100% to less than 150% | 6.4 | 0.2 | 7.8 | 0.3 |
| 150% to less than 200% | 6.7 | 0.2 | 8.1 | 0.3 |
| 200% or more | 78.5 | 0.4 | 75.9 | 0.6 |
| 65 years and over | 100.0 | . . . | 100.0 | . . . |
| Below 100% | 12.8 | 0.4 | 14.6 | 0.5 |
| 100% to less than 150% | 15.8 | 0.3 | 16.6 | 0.5 |
| 150% to less than 200% | 15.0 | 0.4 | 15.9 | 0.5 |
| 200% or more | 56.4 | 0.6 | 52.9 | 0.7 |

. . . Data not applicable.

[1]Percent of poverty level is the ratio of family income to poverty thresholds. The percent of poverty level is calculated for the CPS file using the CPS total annual family income variable, while for the NHIS file it is calculated using the NHIS total annual family income variable.

NOTES: CPS is Current Population Survey and NHIS is National Health Interview Survey. Percentages may not sum to 100 due to rounding.

# Appendix II

## Definitions of Current Population Survey and National Health Interview Survey Variables

The Current Population Survey (CPS) and the National Health Interview Survey (NHIS) collect data for households, families, and persons. For both the March 1996 CPS and the 1995 NHIS, information about each member of the household was recorded (except that the NHIS did not collect information on household members who were in the military). For both surveys, one adult member of the household was selected and the relationship of all other household members to the selected member was recorded. In this report, as in the NHIS, this focal adult is referred to as the ''reference person''; in the CPS, this focal adult is referred to as the ''householder.'' For both surveys, a household may contain one or more families. The primary family of the household is the family that includes the reference person. In this report, the reference person generally is considered to be the head of the primary family in the household. The operational definitions of household, family, and reference person in the March 1996 CPS and the 1995 NHIS are given in the following text.

### Household

*Household* is defined similarly in the March 1996 CPS and the 1995 NHIS.

- CPS—A household consists of all the persons who occupy a house, an apartment, or other group of rooms, or a room, which constitutes a housing unit. A group of rooms or a single room is regarded as a housing unit when it is occupied as separate living quarters; that is, when the occupants do not live and eat with any other person in the structure, and when there is direct access from the outside or through a common hall. The count of households excludes persons living in group quarters, such as rooming houses, military barracks, and institutions. Inmates of institutions (mental hospitals, rest homes, correctional institutions, etc.) are not included in the survey.
- NHIS—A household consists of the entire group of persons living in the sample unit. It may consist of several persons living together or one person living alone. It includes the reference person and any relative living in the unit, as well as roomers, domestics, or other persons not related to the reference person. (Note that like the CPS, the NHIS sample does not include institutionalized persons.)

### Family

*Family*—The definition of a family is similar in the March 1996 CPS and the 1995 NHIS, as demonstrated by the following definitions. However, foster children and other unrelated children living in a household are coded as secondary families (of size 1) in the CPS, while they tend to be coded as members of the primary family in the NHIS.

- CPS—A family is a group of two persons or more (one of whom is the householder) residing together and related by birth, marriage, or adoption. All such persons (including related subfamily members) are considered as members of one family. Beginning with the 1980 CPS, unrelated subfamilies (referred to in the past as secondary families) are no longer included in the count of families, nor are the members of unrelated subfamilies included in the count of family members.
- NHIS—A family is a group of two or more related persons living together in the same household; for example, the reference person, his or her spouse, foster son, daughter, son-in-law, their children, and the wife's uncle. Additional groups of persons living in the household, who are related to each other, but not to the reference person, are considered to be separate families; for example, a lodger and his or her family, a household employee and his or her spouse. Hence, there may be more than one family living in a household.

### Reference person

One *reference person*, or householder, is identified for each household. The relationship of each household member to this selected member is recorded.

- CPS—The householder refers to the person (or one of the persons) in whose name the housing unit is owned or rented (maintained) or, if there is no such person, any adult member, excluding roomers, boarders, or paid employees. If the house is owned or rented jointly by a married couple, the householder may be either the husband or the wife. The person designated as the householder on the file is the ''reference person'' on the CPS-260 control card to whom the relationship of all other household members, if any, is recorded.
- NHIS—The reference person is the person or one of the persons who owns or rents the sample unit, that is, the first person mentioned by the respondent in answer to question 1a on the Household Composition page. For persons occupying the sample unit without payment of cash rent, the reference person is the first adult household member named by the respondent. This person must be a household member of the sample unit.

# Appendix III

## Imputation of Missing Health Insurance Coverage in the 1995 National Health Interview Survey

Health insurance coverage was a key variable in the statistical match, serving both as a partition variable and as an independent variable in the predictive mean match regression models. None of the respondents on the CPS public-use files had missing data for health insurance coverage as, on the CPS, missing data are routinely replaced with imputed values. However, 10,459 (about 10%) of the 1995 NHIS respondents have missing data for some or all of the health insurance questions. As a result, none of these respondents could be assigned to one of the three health insurance coverage categories used in the statistical match (private or Medicare coverage, public coverage, no coverage) and therefore could not be included in the match. Of these, 4,310 have some health insurance information available. For example, an individual may have reported not having public health insurance but have missing data for the private health insurance question. Exclusion of NHIS respondents with missing health insurance coverage information from the statistical match resulted in an unacceptably large percentage of NHIS respondents not being matched in the early unconstrained statistical matches, so the missing health insurance data on the NHIS were imputed. Some of the missing data were imputed using deterministic rules; most were imputed using hot deck imputation.

The first step in the imputation was to assign values to some of the missing items using the following deterministic rules:

1. If there was no indication of any person in the family being in the military, ''unknown'' military coverage was set to ''No military coverage.''
2. For persons less than 65 years of age, ''unknown'' Medicare coverage was set to ''No Medicare coverage.''
3. For persons receiving Aid for Families with Dependent Children, ''unknown'' Medicaid/public coverage was set to ''Yes, public coverage.''
4. For persons aged 65 years and over, if they reported having ''No, Medicare coverage'' and ''No, Medicaid/other public coverage,'' ''unknown'' private coverage was set to ''Yes, private coverage.''
5. For persons aged 65 years and over, if private/Medicare coverage was ''unknown'' and they did not report ''No Medicare coverage,'' Medicare coverage was set to ''Yes, Medicare coverage.''
6. For persons aged 65 years and over, if they reported having no private coverage and no Medicare coverage, ''unknown'' Medicaid/other public coverage was set to ''Yes, Medicaid/other public coverage.''
7. For persons reporting Medicaid/other public coverage, ''unknown'' private coverage was set to ''No private coverage.''
8. For persons with private coverage, ''unknown'' Medicaid/other public coverage was set to ''No Medicaid/other public coverage.''
9. For persons with single service coverage only or no private coverage and with no Medicare coverage, private insurance coverage was set to ''No private coverage.''
10. For persons in families assigned to the 200% or more of poverty category, ''unknown'' Medicaid/other public coverage was set to ''No, Medicaid/other public coverage.''

Use of these deterministic rules reduced the number of persons who could not be assigned to one of the three health insurance coverage categories from 10,467 to 8,488. As a result of applying the rules, all persons aged 65 and over were assigned to one of the three coverage categories. Therefore, persons aged 65 and over were not included in the remainder of the imputation. After applying the deterministic rules, some information on health insurance coverage was available for all of the 8,488 persons who could still not be assigned to a coverage category.

The second step in the imputation was to implement hot deck procedures. Initially, respondents were partitioned into three groups based on the availability or unavailability of health insurance coverage data for members of their family:

- Group 1—All family members have known health insurance coverage.
- Group 2—Some, but not all, family members have missing health insurance coverage.
- Group 3—All family members have missing health insurance coverage.

The hot deck imputation was performed in two stages. First, missing health insurance coverage data were imputed for respondents in Group 2 (some, but not all, family members have missing data). Finally, the imputation was carried out for respondents in Group 3 (all family members have missing data).

### Hot Deck Imputation for Respondents in Group 2

Respondents in Group 1 (all family members have complete health insurance coverage data) were combined with respondents in Group 2 (some, but not all, family members have missing health insurance coverage data). The combined respondents were partitioned into cells using the variables: census region (four levels), urbanization level (four levels), poverty category (four levels), the dominant race and Hispanic origin (non-Hispanic white, non-Hispanic black, non-Hispanic other, Hispanic) in the family, and by whether the household had multiple families or not. All cells with more families with incomplete data than families with complete data were collapsed with an adjacent cell. This affected only a few single-family/multiple-family household cells.

After sorting the records in each cell by household ID, family ID, and person ID, imputation was done in three passes, using a hot-deck procedure. No attempt was made to limit the number

of times a donor was used. As each record with missing data was encountered, a search was done over previous records in the cell for a match on sex and age category. If no matches with nonmissing health insurance status were found, a second search was done by sex only. If, again, no matches with nonmissing health insurance status were found, a third search for any record with nonmissing health insurance status was done. When a match was found, data items from the donor were imputed to the donee; for example, if the indicator flag for missing private insurance information was ''on,'' the private insurance information from the donor was imputed. After imputation was completed, the three category health insurance coverage variable (private or Medicare, public, or none) was computed. Any respondent who still had undetermined health insurance status had their original data restored, records in the cell were resorted in the opposite order, and the above procedure was repeated. No respondents had undetermined health insurance status after this step was carried out.

## Hot Deck Imputation for Respondents in Group 3

The hot deck imputation carried out for respondents in Group 3 (all family members have missing health insurance coverage data) was very similar to that for respondents in Group 2. Respondents with no missing health insurance coverage data (including those in Group 2 who had just had their missing health insurance coverage data imputed) were combined with respondents in Group 3 (all family members have missing health insurance coverage data). Imputation cells were formed in the same manner as in the hot deck imputation for Group 2. The hot-deck imputation procedure followed was identical to that described above for Group 2 with one important difference. Rather than impute values for each missing health insurance coverage item, each respondent's health insurance coverage category (private or Medicare, public, no insurance) was imputed. No checking was done to see if there was conflict between imputed health insurance category and any health insurance data that were present; however, most respondents in Group 3 are missing all health insurance coverage data.

# Vital and Health Statistics series descriptions

SERIES 1. **Programs and Collection Procedures**—These reports describe the data collection programs of the National Center for Health Statistics. They include descriptions of the methods used to collect and process the data, definitions, and other material necessary for understanding the data.

SERIES 2. **Data Evaluation and Methods Research**—These reports are studies of new statistical methods and include analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. These studies also include experimental tests of new survey methods and comparisons of U.S. methodology with those of other countries.

SERIES 3. **Analytical and Epidemiological Studies**—These reports present analytical or interpretive studies based on vital and health statistics. These reports carry the analyses further than the expository types of reports in the other series.

SERIES 4. **Documents and Committee Reports**—These are final reports of major committees concerned with vital and health statistics and documents such as recommended model vital registration laws and revised birth and death certificates.

SERIES 5. **International Vital and Health Statistics Reports**—These reports are analytical or descriptive reports that compare U.S. vital and health statistics with those of other countries or present other international data of relevance to the health statistics system of the United States.

SERIES 6. **Cognition and Survey Measurement**—These reports are from the National Laboratory for Collaborative Research in Cognition and Survey Measurement. They use methods of cognitive science to design, evaluate, and test survey instruments.

SERIES 10. **Data From the National Health Interview Survey**—These reports contain statistics on illness; unintentional injuries; disability; use of hospital, medical, and other health services; and a wide range of special current health topics covering many aspects of health behaviors, health status, and health care utilization. They are based on data collected in a continuing national household interview survey.

SERIES 11. **Data From the National Health Examination Survey, the National Health and Nutrition Examination Surveys, and the Hispanic Health and Nutrition Examination Survey**—Data from direct examination, testing, and measurement on representative samples of the civilian noninstitutionalized population provide the basis for (1) medically defined total prevalence of specific diseases or conditions in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics, and (2) analyses of trends and relationships among various measurements and between survey periods.

SERIES 12. **Data From the Institutionalized Population Surveys**—Discontinued in 1975. Reports from these surveys are included in Series 13.

SERIES 13. **Data From the National Health Care Survey**—These reports contain statistics on health resources and the public's use of health care resources including ambulatory, hospital, and long-term care services based on data collected directly from health care providers and provider records.

SERIES 14. **Data on Health Resources: Manpower and Facilities**—Discontinued in 1990. Reports on the numbers, geographic distribution, and characteristics of health resources are now included in Series 13.

SERIES 15. **Data From Special Surveys**—These reports contain statistics on health and health-related topics collected in special surveys that are not part of the continuing data systems of the National Center for Health Statistics.

SERIES 16. **Compilations of Advance Data From Vital and Health Statistics**—Advance Data Reports provide early release of information from the National Center for Health Statistics' health and demographic surveys. They are compiled in the order in which they are published. Some of these releases may be followed by detailed reports in Series 10–13.

SERIES 20. **Data on Mortality**—These reports contain statistics on mortality that are not included in regular, annual, or monthly reports. Special analyses by cause of death, age, other demographic variables, and geographic and trend analyses are included.

SERIES 21. **Data on Natality, Marriage, and Divorce**—These reports contain statistics on natality, marriage, and divorce that are not included in regular, annual, or monthly reports. Special analyses by health and demographic variables and geographic and trend analyses are included.

SERIES 22. **Data From the National Mortality and Natality Surveys**—Discontinued in 1975. Reports from these sample surveys, based on vital records, are now published in Series 20 or 21.

SERIES 23. **Data From the National Survey of Family Growth**—These reports contain statistics on factors that affect birth rates, including contraception, infertility, cohabitation, marriage, divorce, and remarriage; adoption; use of medical care for family planning and infertility; and related maternal and infant health topics. These statistics are based on national surveys of women of childbearing age.

SERIES 24. **Compilations of Data on Natality, Mortality, Marriage, and Divorce**—These include advance reports of births, deaths, marriages, and divorces based on final data from the National Vital Statistics System that were published as *National Vital Statistics Reports* (NVSR), formerly *Monthly Vital Statistics Report*. These reports provide highlights and summaries of detailed data subsequently published in *Vital Statistics of the United States*. Other special reports published here provide selected findings based on final data from the National Vital Statistics System and may be followed by detailed reports in Series 20 or 21.

For answers to questions about this report or for a list of reports published in these series, contact:

Information Dissemination Staff
National Center for Health Statistics
Centers for Disease Control and Prevention
3311 Toledo Road, Room 5412
Hyattsville, MD 20782

1-866-441-NCHS (6247)
E-mail: nchsquery@cdc.gov
Internet: www.cdc.gov/nchs

**U.S. DEPARTMENT OF**
**HEALTH & HUMAN SERVICES**

Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

MEDIA MAIL
POSTAGE & FEES PAID
CDC/NCHS
PERMIT NO. G-284

07-115596
T30292 (01/2008)
DHHS Publication No. (PHS) 2008-1344, Series 2, No. 144