Barriers and Opportunities for Cross-Temporal Research with U.S. Health Survey Data: The Integrated Health Interview Series (IHIS) as a Case Study

By Miriam L. King, Pamela Jo Johnson, and William Block

In 2007, the U.S. National Center for Health Statistics celebrated the fiftieth anniversary of the National Health Interview Survey (NHIS), an annual, nationally-representative household survey covering a broad range of topics, including health status and disability, health care access and utilization, insurance coverage, health behaviors (including tobacco and alcohol use, exercise, and nutrition), cancer screening and risk factors, mental health, AIDS knowledge and testing, and socio-demographic characteristics (such as education, occupation, income, family structure, race and ethnicity, and immigrant status). Their chronological coverage and scope would seem to make the public use files of the NHIS (available for 1969 to the present) uniquely suited to studying changes in health status, health behaviors, and health differentials over time. In practice, however, studies that span multiple years of these survey data have been rare. A systematic sample from the PubMed database of abstracts analyzing NHIS showed that 70 percent used data from one survey year; 10 percent used two survey years; only 20 percent used three or more survey years; and only 4 percent were based on data before and after 1997.

What have been the barriers to time series analysis of these rich survey data? First, the complexity of the original file structures is daunting, with over 400 separate hierarchical data files, containing data at the household-, family-, person-, and episode-level, and with merging possibilities that include one-to-one exact matches, one-to-one subset matches, and many-to-one matches. Second, researchers who wish to analyze more than one year of NHIS data must work

with multiple codebooks, annual detailed methodology reports, many survey instruments, and several data files in which variable locations, coding schemes, and categories change over time for the same concept (with the documentation for single years running to thousands of pages covering thousands of variables). Third, keeping track of which topics are covered in the full range of core and supplemental surveys is very challenging. For example, researchers who wish to examine cross-temporal data on a given topic (e.g., Pap smears) have to search across multiple file types and supplements (e.g., core person, sample adult, Cancer Control, Health Promotion and Disease Prevention, Year 2000 Objectives) and must stay abreast of changes in question wording, the universe of respondents, the structure of the survey, and post-survey processing. Fourth, major changes in the survey methods--including several sample designs, use of subsampling to gather more detail, and changes to the weighting methodology--impact analyses. Due to the complex sample design, combining multiple years of data must be done with care.

With support from the National Institutes for Child Health and Human Development, researchers at the University of Minnesota associated with the Integrated Health Interview Series (IHIS) project are developing and disseminating an integrated and well-documented cross-sectional time series of health data based on the NHIS. Making these data freely available through a user-friendly web-based data dissemination system facilitates informed analysis of this invaluable source of information about the nation's health. The critical challenges to using multiple years of NHIS data include managing a large number of distinct files, making key documentation easily accessible for each survey year, and harmonizing varying file structures, weighting schemes, and variables. The IHIS builds on the model of the Integrated Public Use Microdata Series (IPUMS), which is a harmonized set of U.S. Census data from 1850 to 2000.

There are three key components to the IHIS data integration project: 1) harmonization, 2) documentation, and 3) dissemination.

Harmonization: Harmonization is the process of taking the original NHIS variables with different coding schemes and creating a new IHIS variable that is comparable over time, that retains the detail provided in the original data, and that is easy to use. In some cases, the task is as simple as always assigning the same numeric code (e.g., "Widowed" equals 5) for variables (e.g., Marital status) that are otherwise consistent over time. In other cases, when some years provide far more detail than others, we adopt a composite coding scheme, with the first digit providing information available across all years and subsequent digits providing additional information available only in some years (e.g., for race variables).

The IHIS harmonization process also allows us to address sample design discontinuities. We constructed the IHIS survey design variables (strata and primary sampling unit (PSU)) so they can be used when examining data from one year or from many years. To do this, we employed the concatenated design period pooling approach suggested by Korn and Graubard (1999, pg. 280) for pooling data from one survey over multiple years and sample designs. To implement this, we concatenated the strata variable with a design period indicator during harmonization. Strata and PSU variables from the same design period are now coded comparably, so researchers need do no additional recoding of these variables, regardless of which years of IHIS data are analyzed. As of this date, the IHIS database includes over integrated 800 variables, and that number is expected to more than double by the end of 2008.

Documentation: The IHI*S* provides a tightly integrated set of documentation that is designed to enhance researchers' ability to work with the data over time. From the home page of IHIS, users can immediately view a list of available variables grouped into broad categories and see in

which years each variable is available. Each IHIS variable name is hyperlinked to a variable description that: a) combines relevant information from multiple years and sources into a single narrative; b) highlights change in the variable universe; c) provides a detailed discussion of comparability problems and their possible solutions; and d) is hyperlinked to a table displaying the codes and frequencies of each value for that variable in each year. The IHIS also includes more general or summary documentation, such as user notes about the original NHIS source data, sample design and sampling weights, and guidance on analysis and variance estimation.

Dissemination: Rapid and user-friendly data dissemination is an integral component of the IHIS. We distribute IHIS data and documentation through a web-based data access system that is available free of charge. For each data extract, the researcher specifies the file type (hierarchical or rectangular), the data format (SAS, Stata, SPSS), and the years to be included, and the variables needed for analysis. For each data extract submitted, the researcher can download a compressed ASCII data file, an extract-specific codebook, and a SAS, Stata, or SPSS command file with syntax to convert the ASCII data to the preferred file format selected by the *IHIS* user.

Conclusion: Multi-year demographic, economic, and public health surveys conducted by the U.S. government and made available to researchers via anonymized public use files offer the potential for rich time series analysis. Often, however, their potential for providing insights into the causes and nature of social change goes unrealized because individual researchers lack the time, the resources, and the patience to tackle the daunting process of harmonizing annual survey data across years. Data harmonization projects like the IHIS (and its predecessor, IPUMS) unlock the potential of these data, making it not only hypothetically possible but also feasible for researchers to study changes in American population structure, behavior, and public health across multiple decades.