Data Archiving and the Implementation of Web-Based Statistical Analysis

Mark A. Carrozza, MA
Institute for Policy Research
University of Cincinnati

Ronald W. Freyberg, MS; Jonathan E. Kopke, BA
Institute for Health Policy and Health Services Research
University of Cincinnati, Cincinnati, Ohio

Shiloh Turner, MPA
Health Foundation of Greater Cincinati, Cincinnati, Ohio

Introduction

Today, more than ever, there are growing concerns and calls for protecting the privacy of individuals and the confidentiality of research data. "HIPAA Compliance" will be the watchword for years to come as data providers struggle to meet new federal guidelines for security and confidentiality. The traditional mechanisms of providing access to data are being called into question, and data producers are beginning to refuse requests for data that one or two years ago would have been considered routine. Data producers, data archivists, and researchers are struggling to find innovative ways to meet state and federal laws, and still allow researchers to conduct well-founded, methodologically valid research.

In response to this need, we have developed OASIS, the Online Analysis and Statistical Information System. OASIS is an innovative web-based program that delivers powerful data analysis capabilities without the need for statistical software on local computers, giving users unparalleled access to data and statistics.

Statement of Problem

The Health Foundation of Greater Cincinnati (HFGC) is an independent 501(c)(4) social welfare organization dedicated to improving community health. The HFGC awards grants to non-profit and governmental organizations for programs and activities that improve health in Greater Cincinnati.

In 2000, the HFGC made a major commitment to data sharing in the region, including the creation of HFGC Health Data Improvement Program and the adoption of significant policy changes in their funding guidelines. The major policy change was to require all grantees, whose grant applications include primary collection of original health data, to submit their data for possible inclusion in a public-release data archive maintained at the University of Cincinnati's Institute for Policy Research (IPR). The HFGC Data Archive policy was modeled after the data archiving and sharing policies in place at both the National Science Foundation and the National Institutes of Health.

To extend their policy, the Health Foundation also funded the development of an online data analysis system that would allow users, particularly smaller agencies and safety-net providers, to access data and statistical analysis tools without the investment in expensive hardware, software, and personnel.

Creating the Health Foundation Data Archive

The IPR created the Health Foundation of Greater Cincinnati Data Archive for project-related data files generated by foundation grantees. Specifically, the IPR develops archiving policies for use by the Foundation, evaluates data and documentation of files associated with each archived project, examines the study files for possible confidentiality and data disclosure concerns, disseminates the data through an Archive Web page, and provides high-end technical support for users of the data.

The IPR worked with the HFGC to develop and publish a set of Archive Guidelines for the Foundation and its grantees. The major components of these Guidelines are as follows:

Outline of Data Archiving Procedures. The Outline acts as a brief guide, for the Foundation and the grantee, of the steps necessary to prepare project data for archiving and public release. The Outline makes explicit reference to the Guide to Data Preparation, the Data Deposit Form, and the Checklist on Disclosure Potential for Proposed Data Releases.

Guide to Data Preparation. The Guide to Data Preparation describes to the grantee what are considered to be “best practices” of data collection and data management for archival purposes. The Guide includes such topics as consistency and wild-code checks, standardization of missing data within and across project files, study documentation, and other issues such as creation of portable project files (files that are operating system and software independent).

Data Deposit Form. The Data Deposit Form collects Study information such as principal investigator, data collection organization, type of data collection, and other study-specific topics such as population, sample design, geographic area, and primary publications resulting from the data collection. This information is used to catalog and archive the studies.

Checklist on Disclosure Potential of Proposed Data Releases. The Checklist on Disclosure Potential of Proposed Data Releases has been developed by the US Bureau of the Census and the National Center for Health Statistics to guide their data producers in the evaluation of the disclosure potential of micro-level data. The Checklist does not guarantee that a dataset will be completely devoid of disclosure risk; the Checklist simply allows for the systematic, detailed evaluation of such risk.

The Development of OASIS

While the HFGC Archive was successful in regional health data archiving and distribution, it was simply a raw data distribution system. In other words, to use the data in the HFGC Archive it was necessary to download the data, write SAS or SPSS code to read the data, and then issue appropriate commands to obtain the statistics of interest. HFGC Archive users fall into two general groups. The first is academic or government agency researchers with the core skills necessary to manage data and conduct original data analysis. The second is users unfamiliar with statistical packages such as SAS or SPSS and usually only in need of a few pieces of specific information (e.g., percent uninsured in a particular county, or the proportion of the population 6 to 12 years of age with asthma). While this information was available in data stored in the HFGC Archive, access to the data for limited pieces of information was cumbersome.

The IPR proposed to the HFGC, and received funding for, the creation of OASIS: the Online Analysis and Statistical Information System. OASIS was created using a base set of state-of-the-art applications including ColdFusion Server and ColdFusion Studio, Microsoft SQL Server, Base SAS, SAS/Stat, and SAS/IntrNet.

ColdFusion is an Internet "middleware" application that acts as a bridge between a user's browser, the HTML code that is the language of the Internet, and the databases available to Web servers. ColdFusion facilitates the development of Web-based applications that can submit standard SQL (Structured Query Language) statements to select, insert, modify, or delete records in an underlying database. Microsoft SQL Server is the enterprise-level relational database management system developed by Microsoft, and is capable of storing millions of records in a complex database structure. SAS/IntrNet provides the technology for building Web applications that allow users to access and execute remote SAS programs and to perform sophisticated analysis through the Web.

OASIS Metadata and Guided Analysis

While Internet applications such as ColdFusion and SAS/IntrNet do the work of OASIS, the system is centered on detailed metadata. Metadata, or data about data, is the information that is typically found in print documentation, except that metadata is stored in a database that can be used for subsequent programming.

The metadata currently used by OASIS contains information about every variable in each of the OASIS data files including variable name, descriptive labels, level of measurement (nominal, ordinal, interval, dichotomy), whether the variable is a geographic indicator, whether the variable is a statistical weight variable, and other details. These details allow the OASIS programmers to customize OASIS based on the characteristics of the data. For example, if there are weight variables in the data file, separate instructions appear on the screen allowing weighted analysis of the data. The OASIS metadata drives nearly all aspects of OASIS screens and navigation.

Especially important to OASIS is the level of measurement for each variable in the data file. Each variable is categorized as nominal (categorical data that have no inherent

order), ordinal (categorical data that can be placed in a logical ascending or descending order), interval (continuous data that have an inherent order with equal occurring intervals), or dichotomous (nominal data with only two categories). The "Guided Analysis" portion of OASIS is driven by the metadata describing the level of measurement.

In Guided Analysis, users select the outcome measures and predictors, and are then guided towards appropriate statistical analysis procedures by the OASIS interface. For example, if users select an ordinal outcome variable and a nominal predictor, OASIS performs appropriate analysis such as contingency tables, chi-square tests, or Kruskal-Wallis tests.

Future Directions

Our intention is to develop a research infrastructure that will enable authorized researchers to conduct sophisticated statistical analyses on shared confidential and restricted microdata without breaching confidentiality or disclosing the identity of the individuals associated with health data records. By providing such analytic-level access to data, we will allow researchers to design and complete health research agendas that earlier would have been either impossible or extremely prohibitive due to the myriad of organizational, bureaucratic, and administrative barriers to accessing confidential microdata.

Sharing research data, whether public-use data or confidential and typically restricted data, has clear and widely accepted benefits. Such sharing often:

- Reinforces open inquiry;
- Encourages diversity of analysis and conclusions;
- Promotes new independent research of alternative theories;
- Encourages use of empirical studies in public policy formation and evaluation;
- Allows for multi-disciplinary analysis of data;
- Serves as a protection against faulty (inadvertently distorted or willfully fabricated) data.

In fact, providing access to restricted microdata was the impetus behind the creation of the CDC/National Center for Health Statistics Research Data Center program as well as the similar Center for Economic Studies/Research Data Center program developed by the US Bureau of the Census.

Data analysis will be restricted, through the Web interface, to a limited subset of Statistical Analysis System (SAS) procedures. SAS procedures that can disclose confidential microdata (e.g., PROC PRINT, PROC IML) will not be provided in the Web interface, adding an additional layer of data security.