Data Security

ROCKVILLE, MD
AUGUST 8, 2003

Workshop Summary

WORKSHOP ORGANZING SPONSOR:

National Institute of Child Health and Human Development, National Institutes of Health
U.S. Department of Health and Human Services

The statements, conclusions, and recommendations contained in this document reflect both individual and collective opinions of the symposium participants and are not intended to represent the official position of the U.S. Department of Health and Human Services, or the National Institutes of Health.

Demographic and Behavioral Sciences Branch (DBSB)

Center for Population Research (CPR)

National Institute of Child Health and Human Development (NICHD)

National Institutes of Health (NIH)

Report of a

Data Security Workshop

August 8, 2003

Anita H. Yuan and V. Jeffery Evans

Background

Principal Investigators (PIs) of data-collection projects must devise data security procedures that protect respondent confidentiality, while maximizing access to the data by the scientific community. The potential risk of identifying a respondent varies with each data set, as does the potential harm caused by respondent identification. Thus, data security plans must be tailored to the unique needs and concerns of each data set: a “one-security-plan-fits-all” approach is neither feasible nor desirable. Nevertheless, population researchers responsible for assuring the security of their data can learn much from each other. Sharing experiences and approaches can help to create a consensus of minimum standards or practices necessary for any data-collection project and can provide examples of successful security practices available for more challenging circumstances. This document summarizes the results of the second of two workshops convened by the DBSB of the NICHD, to facilitate discussion of data sharing and data security practices among population researchers.

In the first workshop, held in the fall of 2001, the DBSB assembled a panel of “expert” data collectors, archivists, marketers, and users to discuss issues of data sharing and archiving (Melichar, Evans, and Bachrach, 2002). Four of the six recommendations that resulted from the workshop specifically encouraged more data sharing and increased financial support of such efforts:

- All reasonable steps should be taken to ensure that data collected under NIH assistance mechanisms are made accessible to all who wish to use them for scientific analyses.
- The DBSB should increase its support of data sharing by investing new funding in improving the data-sharing infrastructure of the population research community.
- PIs should be encouraged to fully articulate plans for data sharing and budget funds to pay for the costs of cleaning, documentation, storage, archiving, and distributing their data.
- The DBSB should use assistance awards as part of its commitment to avoiding cuts to/discontinuities in funding used for data sharing.

Workshop participants were particularly concerned about ensuring access to inexperienced and seasoned researchers alike, and about lowering financial costs to users to provide better access.

Participants also expressed significant concern about data security issues, in particular, risks posed by access to sensitive information, such as geographical identifiers, and the potential for deductive disclosure (i.e. identification of respondents through indirect means). Since the 2001 data sharing workshop, the NICHD/DBSB has sensed some urgency developing in the field of population research regarding concerns over best practices for data security. A central theme of these concerns was the tradeoff between access and security: was the NICHD/DBSB imposing standards that excessively and perhaps unnecessarily increased costs to the data user community?

In August 2003, DBSB convened a second workshop to explore these issues. Participants included PIs of social science data sets with challenging security needs, researchers from several federal data-collection projects, representatives of the user community, and a specialist in data security. Participants addressed two central questions: (1) what are effective and efficient existing practices that balance the risk of disclosure against accessibility to data; and (2) is there a consensus regarding data security standards?

This document summarizes major points from the presentations and discussions at the workshop and can be used to guide and assist researchers when thinking about issues of data security.

Designing a Data Security Plan

Data security plans must be designed to match the requirements of each unique project. Not all projects may require a complex security plan. The first step for any researcher who is collecting data is to conduct an assessment of data security needs, which includes determining confidentiality risks, potential harm from breaches of confidentiality and the likely demand for use of the data. Each of the following characteristics poses a risk to respondent disclosure: a high number of people who know that someone participated in a survey, a highly clustered sample design, collection of sensitive information and/or behaviors, and availability of fine-grained geographic identifiers. When determining the most appropriate and reasonable data security plan, researchers must consider both harm to respondents due to disclosure, and the actual risk of respondent disclosure. They must also consider potential demand for the data from outside users. It makes little sense to invest heavily in data-sharing security if few researchers are likely to want the data for secondary analysis.

Public Use Data Sets

Many researchers who collect demographic data sets are able to share data simply by releasing a public-use file after removing identifiers and other information that could lead to disclosure of respondent identities. Public use data sets should entail minimal risk of deductive disclosure and should have minimal legal and financial requirements so that all researchers can have easy access to the data. They should also contain enough information to be of substantial scientific value. If possible, public use data sets should contain sufficient information to account for complicated survey designs so that researchers who use the data are not disadvantaged in the publication process.

Most studies, but not all, require a pledge of confidentiality before access to public-use data is granted. This pledge includes, at minimum, user agreement to: make no attempt to identify respondents or sampling units in the study, not share data with other researchers, destroy data files if requested to do so by study staff, and report disclosure of participant or study unit identity as well as errors to study staff. This pledge can be obtained by electronic signature as well as by printed and signed document. Participants felt that having a pledge of confidentiality alerted users to concerns over data security.

Tiered-Access Security Plans

When a data set poses too great a risk for disclosure of respondent identities or too great a potential harm from disclosure to permit release of all useful data in a public-use file, many researchers adopt a tiered-access security plan. In a tiered-access system, different levels of security are employed for different versions of a data set. The level of security matches the level of risk for disclosure and harm of the data set to be shared. The “tiers” in such plans may include a public use data set, restricted data-use contracts, and use restricted to a data enclave or cold room.

As an illustration, Table 1 summarizes the tiered access system adopted by the Los Angeles Family and Neighborhood Study (LAFANS). Only an on-line user agreement is necessary to access the low-risk public-use file. A licensing agreement and other security measures are required to access two moderate-risk versions of the data. Data posing the highest risk -- including precise geographic coordinates of residences and other locations -- require the same security measures as the moderate-risk data and in addition may only be accessed in a data enclave.

Table 1. Tiered access to LA FANS data

	Public Use	Restricted V.1	Restricted V.2	Restricted V.3
On-line user agreement	Yes	-	-	-
Brief research proposal	-	Yes	Yes	Yes
Data protection plan	-	Yes	Yes	Yes
Licensing agreement	-	Yes	Yes	Yes
Institutional Review Board (IRB) approval	-	Yes	Yes	Yes
Processing fee	-	Yes	Yes	Yes
Secure data enclave	-	-	-	Yes

Creating a public use data set for a complex study that entails risk of disclosure or harm requires creative redaction of the data, in addition to removing personal and geographic identifiers. PIs of existing studies have used a variety of strategies to minimize disclosure risk to the point where data can be safely released in a public-use file. Such strategies include randomly sampling a subset of survey respondents for inclusion in a public-use file, excluding over-sampled populations, removing geographic identifiers at all levels of clustering, and excluding responses to sensitive questions.

Restricted Data-Use Contracts

Restricted data-use contracts are typically used by PIs to protect respondent confidentiality in studies containing information that poses a moderate level of risk for deductive disclosure or harm to respondents. Contracts serve two purposes. First, they are legal agreements between the investigator and the receiving institution to honor scientific integrity, to use the data for statistical reporting and analysis only, and to adhere to the study's data security plan. Contracts typically require users and institutions to actively protect respondent identity through measures such as copying the data only once, keeping restricted data sets in locked cabinets, and password-protecting individual computers. Second, contracts allow users access to data on sensitive behaviors and contextual data, which are sometimes of greatest scientific value. Advancement in population research would be limited if access to such data could not be made to researchers on an appropriate scale.

Workshop participants raised several concerns about these contracts, including financial burden, the possible conflict between researcher and institution from institutional commitment required by contracts, complexity of contract applications, and difficulty fulfilling technological requirements for data security. The latter concern can seriously hinder smaller institutions from gaining access to restricted data due to the lack of resources, such as Information Technology (IT) help-desk personnel to help set up firewalls and hacker-proof network systems at the receiving institution. This situation also highlights the need for user-support from institutions granting access to restricted data. Data users commented that examples of successful contracts and communication to investigators about the importance of security measures were helpful in completing contract applications.

Not all studies can or should release data via restricted-use contracts. In some studies, the data that can be provided at a moderate level of risk may be of insufficient value to warrant the investment required to establish contractual arrangements.

Enclaves

Data that present greater-than-moderate risks of disclosure or harm require the highest levels of security. Examples include ethnographic or videotaped data, and data sets that contain geographic coordinates or geographic positioning system (GPS) readings.

At present, data enclaves or cold rooms provide the level of security needed to ensure respondent confidentiality for such data. A cold room is a supervised data-use facility. For electronic data, it may be a locked or password-secured room containing a dedicated computer(s) (and printer, if needed) that is not connected to any network. Users are closely monitored to prevent unauthorized use or copying of the data. Workshop participants discussed using technology to create virtual enclaves so that researchers and/or institutions with limited resources would have access to data. It was noted that secure networks are currently in use by the Department of Defense and by banks of all scales to share highly sensitive information; so the technology is available to create a system of secure and privileged access to restricted data sets.

Access to restricted data sets or cold room data sets can be problematic when researchers, such as faculty or graduate students, change institutions and need to make revisions prior to publication.

Keeping Data Secure

There was a general consensus that monitoring data use and enforcing security agreements are weak links in the process. Although PIs present at the workshop knew of no known cases of deductive disclosure, they agreed that adequate monitoring and enforcement were key elements of any data security plan. The use of certificates of confidentiality is an important shield against disclosure of identifiers through a legal proceeding. The use of certificates was highly recommended for those collecting sensitive information.

Many PIs felt that these aspects of data security were beyond the capability of the PI, and that an outside organization would be better equipped to detect and correct data-security infringements. However, participants agreed that individual PIs and study institutions could not surrender their responsibility for respondent confidentiality to another organization. Even if another organization is charged with monitoring and enforcing security agreements, the ultimate responsibility remains with the PI and the institution that conducted the study.

Monitoring

Some PIs use the fee-charged-for-access to the data set to conduct unannounced site visits. Other PIs, especially those with smaller scale studies, do not have the financial or human resources to conduct such oversight. Monitoring procedures and/or activities for different data sets may be similar if not identical, and site visits to large research centers could cover use of multiple data sets. Thus, room for collective action exists. The question of what are effective means of assessing and monitoring disclosure risk remains. There are no known studies that systematically examine the risk of deductive disclosure.

Enforcement

Contract users acknowledge and agree to the consequences of a breach of contract, which include loss of access to the restricted data, sanctions imposed due to violation of research ethics code, and even criminal prosecution. There was one suggestion that collective action among data producers might be applied to deny violators access to a large number of data sets. Many workshop attendees were opposed to such measures. Most preferred a system in which there was better monitoring of contractual agreements through collective action among or on behalf of data producers, and better risk assessment *ex ante* of deductive disclosure. This process would result in more user-friendly agreements and a more secure system for protecting population research.

Future Directions

The workshop raised many unanswered questions about designing and implementing effective data security plans. The DBSB is committed to helping to find answers to these questions. In light of these needs, the NICHD will fund a program project headed by researchers at the Inter-University Consortium for Political and Social Research and the Survey Research Center at the University of Michigan to address the protection of human subjects through disclosure risk analysis and disclosure limitation. These projects specifically address the following topics:

- Informed consent and perceptions of risk and harm in survey participation
- Estimation of disclosure risk and statistical methods for disclosure limitation
- Statistical disclosure control: best practices and tools for the social sciences
- Resources for the secure dissemination of human subjects data

This program project will assess risk of disclosure and assess the balance between protection and information loss in disclosure-limitation methods. Presently, it is difficult to assess when the appropriate balance has been achieved between providing access to data and protecting research participants against risks associated with data sharing. This fact underscores the importance of developing an informed consensus of what major risks to identity disclosure are and how to secure data against such risks. Hopefully, this research will solve some problems that derive from the current uncertainty regarding deductive disclosure.

At the conclusion of the workshop, a consensus formed on the following points:

- All users of shared data should, at a minimum, provide a pledge of confidentiality, and data should routinely be made available in public-use data sets that contain enough information to create publishable analyses.
- For data that cannot be shared adequately in a public use file, PIs should consider tiered systems of access to data, designed to provide protection appropriate to the potential for disclosure and harm in the data.
- Cooperative systems to monitor and enforce security agreement are needed to better address a recognized weakness in current security systems.
- PIs should remain involved in the data security process and retain ultimate responsibility for the protection of respondent confidentiality.
- There is a need to codify best practices related to the development and implementation of data security plans.