

When Data Sharing is Required: I. What is this Requirement?¹

Virginia de Wolf²
Kerrville, TX

Joan E. Sieber³
California State University, Hayward

Philip Steel⁴
United States Census Bureau

Alvan Zarate⁵
National Center for Health Statistics

Openness is a norm of science that has been acknowledged for centuries, but until about 1990 the practice of sharing research data has varied widely and has rarely been subject to specific rules and procedures.⁶ However, recent changes in federal regulations have dramatically altered this picture. Now there are specific requirements, rules, and procedures.

As of October 1, 2003, the National Institutes of Health (NIH) requires specific plans for data sharing with protection of confidentiality to be addressed in larger funding proposals.⁷ This has important implications for Institutional Review Boards (IRBs) which oversee the way in which informed consent and confidentiality are handled in relation to data sharing, and for researchers who must prevent disclosure of identities and spend more time and effort in preparing data.

In addition, in accordance with provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Department of Health and Human Services has issued the Privacy Rule ("Standards for Privacy of Individually Identifiable Health Information").⁸ This Rule permits research use of individually identifiable health information without the patients' authorization when a waiver of authorization is approved by an IRB. Waiver criteria to be employed by IRBs include the treatment of identifiable data and restrictions as to their reuse and disclosure to others.

The purposes of this three-part series of articles are:

- To acquaint IRBs with what is meant by data sharing, why it is an important norm of science, how it relates to HIPAA, basic ways in which data sharing is conducted, some current funding requirements, and a summary of how data sharing should be addressed in research proposals and presumably in IRB protocols.

- To provide information on methods of preparing data for sharing that protect confidentiality.

Many investigators who conduct human research are not acquainted with these matters. These three articles will enable an IRB to provide some guidance to its investigators as they seek to comply with data sharing requirements. This knowledge will also enable an IRB to better educate its research administration about its responsibilities as it participates in data sharing agreements with funders and secondary data users.

In this first article, we review current data sharing practices and the requirements for sharing of data of selected funders of human subjects research. We then discuss why researchers share data and what is encompassed by the word “data.” We begin the second article by describing how data are prepared for sharing. We then proceed to the data sharing provisions of the HIPAA Privacy Rule and how consent statements might be adapted to permit data sharing. In that article, we include a brief discussion of why the removal of names and addresses is insufficient to de-identify data – and would, therefore, preclude data sharing under HIPAA and the NIH confidentiality provisions. We also provide an example of a re-identification technique and describe methods for evaluating the risk of disclosure. The third article provides a summary of the technical and administrative procedures for assuring confidentiality in data files. It concludes with an overview of the issues that IRBs and research administrators need to consider in order to meet the requirements for responsible and legal sharing.

Current Data Sharing Practices

Data sharing has been a well-established practice in some physical sciences (e.g., astronomy, oceanography) for many decades, and in meteorology for well over a century. Some human subjects funding agencies have long required or encouraged data sharing, e.g., the National Science Foundation (NSF) which incorporates as a default element of its research grants a general requirement that its grantees arrange to make their data available to other scientists,⁹ and the National Institute of Justice (NIJ) which incorporates in its research contracts, grants, and cooperative agreements the stipulation that resulting data be returned to NIJ which, in turn, prepares and archives those tapes in the National Archive of Criminal Justice Data at the Inter-university Consortium of Political and Social Research (ICPSR).¹⁰ In some of the biosciences, the sharing of reagents, cell lines, probes, clones, descriptions, quantitative data, and other items of data has been the basis for rapid scientific development. However, these sciences have also experienced a tension between following the ideal of sharing, and pressure on individual investigators to retain dominance in their field, protect their investment, or comply with university licensing requirements.¹¹ Most of this bioscience research has not involved human subjects. Where it does, confidentiality adds further to the tension of sharing.

Table 1 Glossary of Key Terms	
Term	Definition
Anonymous	Strictly speaking the removal of identifiers only, but used here synonymously with de-identification.
De-identification	The removal of direct identifiers together with the removal or modification of other information about study participants in order to minimize the risk of re-identification.
Identifier (also referred to as direct or explicit identifier)	Information that is uniquely or very closely associated with only one person, e.g., names, addresses, telephone numbers, Social Security numbers.
Microdata files (also referred to as "line listed" or "person" records)	Microdata files are computerized files that consist of individual records, each containing values of variables for a single person, business establishment, or other unit.
Public-use data products	Data that are released to anyone without restrictions on use or other conditions, except for the payment of fees to purchase publications or data files in electronic format. ¹² Typically, public-use data products are released either as tables or microdata files.
Re-identification	(a) The association of information about a study participant in a released file with the name of that participant; or (b) the identification of a study participant from information contained in a released file.

Some major surveys (e.g., labor, voting behavior, social attitudes, economics) – whose data have been rendered anonymous (“de-identified”) – have long been subject to sharing requirements and their data reside in archives readily available to scientists and the public.¹³ A major national archive, ICPSR, permits its subscribers to access its data sets via the internet.¹⁴ Students as well as scientists have routinely used these data for teaching and research purposes since 1962.

Federal statistical agencies have also released de-identified data from a wide variety of surveys as “public-use” microdata files.¹⁵ In addition, several of these agencies have developed special facilities, Research Data Centers, where approved researchers can analyze data that have not been de-identified, under carefully controlled conditions.¹⁶

However, the sharing of smaller human-subjects data sets from biomedical studies and social-behavioral research (SBR) has had a contentious history over such questions as: what constitutes data and sharing; who pays the cost of data preparation, archiving, and sharing; who owns data gathered under federal funding; who is required to share; what are the informed consent requirements; and especially how confidentiality of human subjects data can be protected.¹⁷ In 1989, an article appeared in this journal that informed IRBs of challenges they would face as more kinds of human subjects data began to be shared.¹⁸ That article outlined the reasons for data sharing, its ethical foundations, barriers to sharing, rights of the original

researcher, how IRBs may be involved, and how to solve problems surrounding data sharing. The article had a tentative tone. At that time, data sharing was still new to most small-scale biomedical and SBR. While NSF and NIH by then were fostering data sharing, the role of IRBs in this matter was still ambiguous. This is no longer the case.

With respect to sharing SBR and biomedical data, three things have changed: (1) Answers have emerged to most of the issues that have been debated. (2) Federal funding agencies are requiring, rather than urging, data sharing in cases where it can reasonably occur. (3) There is a clear role for IRBs in reviewing protocols when research funding is contingent on an agreement to share data.

Funders now require or urge that proposals contain evidence of a well-planned sharing arrangement that provides data in a useful form to other researchers without breaching promises of confidentiality. This raises immediate questions for IRBs and investigators:

- Does the informed consent permit the sharing of identifiable data? If so, how does the primary researcher obtain assurances of confidentiality and security from secondary users?
- If data are not to be shared in identifiable form, how does the primary researcher reduce the risk of identifiability to acceptable levels? By what criteria are “acceptable” levels of risk to be judged?
- What constitutes data? Does this include the initial data in their most raw form? Does it include the financial data (grant administration data) for the project?
- How soon must the data be shared? Must they be shared with anyone who asks? May the principal investigator (PI) charge for sharing? If so, how much? How long must the data be kept? Who owns the data: the institution, the PI, or ...? Who is responsible for sharing? Does the IRB have any role in these decisions? For example, can the IRB be called upon to decide on the legitimacy of certain requests for data (e.g., by those likely to use the data to harm the human subjects involved)?
- How final a data sharing plan can or must be developed before any data are gathered? What do funders expect? What should an IRB expect?
- What institutional resources may be needed to comply with sharing agreements?

Data Sharing Requirements of Selected Agencies

Not all public funders require data sharing. The most specific requirement is the most recent. NIH mandated, effective October 1, 2003, that all research grant proposals over \$500,000 include a detailed description of how the data will be prepared for public use or restricted use by others without breach of confidentiality – or explain why this cannot be done.¹⁹ Simple stripping of identifiers is not sufficient for this purpose. Given the many kinds of data bases available today, privately and on the web, data matching software can enable snoopers to

determine with a high degree of accuracy the identity of some research participants whose data are “anonymous.”²⁰

NIJ has pioneered methods of assuring confidentiality, since data collected under its sponsorship typically have been highly sensitive. NIJ’s long-standing and specific data sharing requirement – that those receiving grants, research contracts, or cooperative agreements deliver data to NIJ at the completion of the project – leaves to NIJ the responsibility to archive de-identified data with ICPSR, and to keep identifiable data within its own secure archives where they may be used on a restricted data basis. (We will discuss such archives in our third article).²¹

Other agencies, such as NSF, have less specific requirements. Over the past 25 years, NSF has strongly encouraged data sharing and sought to better understand how to maximize the usefulness of shared data.²² NSF has funded projects developing various technologies of data sharing (e.g., via the internet). Although NSF’s requirements of researchers are not as specific as NIH’s new rules, proposals to NSF that spell out specifically how data will be shared and what methods of nondisclosure will be employed have a distinct advantage over those that do not.

Why Share Data?

Apart from funder requirements, there are many other reasons to share. The norm of openness in science is embraced by most investigators, and urged through peer pressure upon reluctant colleagues. Scientific societies urge data sharing and many journal editors require it as a condition of publication. Shared data give some assurance of the validity of the research and provides a common ground upon which to work out controversy.

Data sharing provides an inexpensive and feasible way to make comparisons across populations; build upon one’s own data with additional related data; reanalyze data with different designs, methods, or hypotheses; promote interdisciplinary analyses; validate original results; conduct methodological and policy research; and create curriculum in research methodology or statistics using real data. Researchers who do secondary analysis of data, and those whose research draws on many disciplines, need shared data; sometimes such researchers engage in reciprocal sharing. Clinicians, too, often share data (i.e., medical record information) with researchers.

What is Meant by Data?

There are many forms of data that a secondary analyst might need. Some raise issues involving confidentiality, and others raise questions about costs and sharing procedures. Consider the following partial list:

- Quantitative data: graphs, handwritten data sheets, data tapes and diskettes;
- Qualitative data: field notes, video tapes, recorded verbal responses, case studies, ethnographies;

- Samples: biological specimens, irreplaceable archeological artifacts, Rorschach test stimuli and responses;
- Materials: surveys, observation systems, stimulus materials, access to a research site;
- Tools: cell lines, equipment required to do the research, software; and
- Research procedures: documentation of methods for coding, cleaning, and organizing the data; description of the methods of statistical analysis; pilot data; and training in the skills needed to conduct the research (“know how”).

This list may provoke puzzlement. How can some of these items be shared? Some items cannot be allowed out of the investigator’s laboratory and must be used there by the secondary analyst, perhaps under strict supervision. Who bears the costs of sharing? There is some cost involved in any of these kinds of sharing, and the secondary user may legitimately be charged the marginal amount that it costs to share, consult, or provide laboratory resources. Costs of preparing the data for sharing may be (and in NIH's case it should be) built into the research proposal.

Some debate about what must be shared has been raised in connection with concerns about various forms of research fraud. In that connection, institutions that administer many large grants have expressed concern that financial records of projects might need to be retained and made available over long periods of time. Such records are extensive, the personnel responsible for their maintenance and for the documented expenditures may move from one institution to another, and the costs of maintaining such records in a useful form for posterity are exorbitant. While funding agencies may choose to audit such data during or shortly after completion of the project, financial records are not under the rubric of data sharing.

Kinds of Sharing

Most human subjects participate in research to advance science and benefit society, not just to promote the career of one person. Many have some notion that science is open. Hence, most would probably be upset at the notion that their data were used solely by one scientist who refused to share with other trusted and qualified scientists so that they could test their own important scientific hypotheses. In fact, certain kinds of data sharing are normal and expected, as when an investigator shares data with a graduate student, collaborator, or close colleague. In these instances, the investigator also shares the trust the participants have placed in her/him, and respects that trust by sharing only with persons of integrity. Integrity is the basis of confidentiality in such sharing arrangements.

Other kinds of informal sharing have also occurred over the years. Given the norm of openness in science, most researchers have, at some time in their career, received a request from another scientist for data, materials, samples, tools, or some other items that would permit that other scientist to replicate or extend the work. When scientists have produced research that others have questioned, colleagues and boards of inquiry have asked for information bearing on every aspect of the research from its beginning to the final data analyses. Arguments over

statistical interpretation have sometimes required not just the initial data, but additional research and analysis. With that background in mind, consider some of the kinds of sharing arrangements that exist:

Collaborative reanalysis: Researcher A produces results that B questions. They decide to work together on a reanalysis using techniques that B suggests. A understands the data and teaches B, while B teaches A new kinds of analysis. The data and materials may never have to leave A's laboratory. The new analysis is probably published under joint authorship.

Reciprocal exchange of data: A and B use comparable techniques to study different populations, or are doing research with some other commonality; the reciprocal exchange of data enables them to make useful comparisons.

In these first two kinds of sharing, the initial researcher affirms to the colleague how confidentiality will be assured, and how a general standard of scientific integrity will be upheld. In many cases this affirmation is implicit, and a formal contract would be offensive, although the IRB might require it.

Unilateral sharing: A's data are made available to B. If the data are sensitive, identified, or include irreplaceable samples or artifacts, B must make a good case for needing the data (e.g., a proposal is presented). B might do the work in A's lab under supervision, or might receive the data to be used elsewhere. Conditions may be imposed that assure confidentiality (using techniques to be described in the third article) of identified data (e.g., video tapes) or to assure that re-identification does not occur.

Projects organized for sharing: An anthropologist organizes a research program to benefit as many stakeholders as possible. He trains indigenous people to gather and analyze data on members of their community; the data, software, computers, and other necessary items are given to the legitimate local government so that they may continue to gather and analyze data for their own policy purposes. The PI trains research personnel in appropriate confidentiality practices. Alternatively, an anthropological project involving various investigators may be organized for reciprocal sharing.

Public data archives: A's data are documented and altered to prevent re-identification of any of the subjects (using techniques to be described in the third article). The data are stored in an archive such as ICPSR, which then handles requests for the data and may also assist borrowers in understanding the data. Secondary users of A's data do not need IRB approval to use these data, as the data are de-identified, hence are not human subjects data in the regulatory sense (as defined under 45 CFR 46).

Restricted access archives and research data centers: Alternatively, the archive may store identified or identifiable data and employ mechanisms (described in the third article) to limit the possibility of a breach of confidentiality, through control of the access and uses by secondary users. For example, the Murray Center at Radcliffe College is an archive that holds data from longitudinal studies of human development.²³ Some of the data include video tapes.

Some users conduct their research under supervision at the Murray Center. Others are permitted to remove such data under highly restrictive conditions.

Similarly, several federal agencies, e.g., the National Center for Health Statistics (NCHS), the U.S. Census Bureau, and the Agency for Health Care Research and Quality (AHRQ), maintain facilities where researchers can analyze data not released to the general public.²⁴ In these data centers, researchers work under controlled conditions and their research output is subject to disclosure analysis to insure that what is taken away cannot lead to the re-identification of research participants. NCHS also permits remote analysis of (but not actual access to) confidential data.²⁵

Draft of manuscript to be submitted to *IRB: Ethics & Human Research*. Please do not cite or quote.

When Data Sharing is Required: II. HIPAA and Disclosure Risk Issues²⁶

Virginia de Wolf²⁷
Kerrville, TX

Joan E. Sieber²⁸
California State University, Hayward

Philip Steel²⁹
United States Census Bureau

Alvan Zarate³⁰
National Center for Health Statistics

This is the second in a series of three articles whose focus is to assist Institutional Review Boards (IRBs) in providing guidance to investigators and research administrators who need to comply with data sharing requirements. This article begins with a discussion of how data are prepared for sharing. We then proceed to the data sharing provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule and how consent statements might be adapted to permit data sharing. Next, we turn to the topic of de-identification and show why the removal of names and addresses is insufficient to de-identify data – and would, therefore, preclude data sharing under HIPAA and the National Institutes of Health (NIH) confidentiality provisions. An example of a re-identification technique is provided. The final section briefly looks at some of the elements of a risk evaluation for one type of data, microdata.

How are Data Prepared for Sharing?

When sharing is done informally among colleagues, little needs to be done as the initial researcher is on hand to explain the data, and prudent sharing with trusted colleagues avoids breach of confidentiality. However, it is easy for an investigator to forget, within a few months, just how the data were gathered, how each variable is defined, and so on. Hence, some documentation of the data is always appropriate.

When data are to be made available for public use, however, extensive documentation is needed. Most funders that encourage sharing, and archives that receive data for public use, specify how they want the data prepared. For example, the Inter-university Consortium of Political and Social Research (ICPSR) provides extensive instructions for persons who plan to deposit data with ICPSR.³¹ These excellent instructions include details on how to plan for sharing at the research stage, how to ensure confidentiality, and how to document the data. As its website indicates, ICPSR prefers that investigators take initial responsibility for removal of all identifiers and other kinds of information that could possibly lead to deductive identification (re-identification) of subjects. However, ICPSR's expert staff work with investigators to help resolve confidentiality issues. Hence, with respect to ICPSR and many other public archives, responsibility for ensuring confidentiality does not rest solely on the researcher or the IRB. Briefly, the main steps of preparation for sharing, in most cases, are as follows.

Data are documented. Documenting of data consists of describing the data so that secondary users can understand and use the data correctly. There are many kinds of problems that a secondary user is likely to encounter when trying to decipher a quantitative data set. The purpose of documentation is to make data as user-friendly as possible for the secondary user. However, some information supplied in documentation may add to disclosure risk. The most common problem is when the identity of geographic areas having small populations is revealed. From such information it may be easy to pinpoint, for example, which participant is the 15-year-old with a Ph.D., city mayor, or 105-year-old man, and from that inference to learn other kinds of sensitive information about the person from the data.

Direct identifiers are removed. These include names, addresses, employers' names or addressees, relatives' names or addresses, dates, telephone and fax numbers, email addresses, Social Security numbers, medical record numbers, account numbers, photos, and so on. These are the obvious steps toward rendering data anonymous.

Disclosure review is performed and data are adjusted. Once identifiers have been removed, the remaining data must be checked for possible risk of re-identification through matching with other records that do contain identifiers. Typically, checking data for risk of re-identification requires the services of an individual who has knowledge of these statistical methods to limit disclosure (such as a statistician, biostatistician, demographer, sociologist, epidemiologist, or health services researcher) and is not normally within the scope of an IRB's expertise. Rather, the consultation of such an expert by the researcher and the IRB ~~are~~ **is** usually required. Some institutions employ the services of a Disclosure Review Board, rather than a single individual, to fulfill this role. Where necessary, the researcher and the statistical expert(s) select and employ disclosure limitation techniques to prevent re-identification of the data.

How is the HIPAA Privacy Rule Involved with Data Sharing?

For purposes of research, the Privacy Rule permits a holder of identifiable (protected) health information to release that information with or without patients' authorization, but if it is to be released without authorization, certain conditions must be met. For example, the data holder (the "covered entity"³²) must receive documentation that an IRB or a Privacy Board has found that the proposed research involves no more than a minimal risk to patient privacy based on several criteria including assurances that identifiable information will not be improperly disclosed, i.e., to third parties not approved by the IRB in their waiver of authorization. Note that in granting a waiver of authorization, the IRB must decide whether assurances submitted by the researcher not to reuse or disclose protected health information are adequate.³³ Both the IRB and the researcher, then, need to know how data can be de-identified should the latter desire to further share data obtained under the Privacy Rule.

The Privacy Rule does not restrict the disclosure or sharing of information that has been properly de-identified. Under the Rule there are several ways of doing this. First, the covered entity may ensure that the data are limited geographically³⁴ and stripped of each of 18 "identifiers" listed in the rule, ranging from name, address, and Social Security number to electronic device and biometric identifiers.³⁵ In addition, the covered entity must be satisfied that



what remains in the body of data cannot be used to identify anyone. This involves a judgment as to what data might, by themselves or in combination, be used to attach a person's name to the data. Knowledge and expertise to assess disclosure risk may not be available to all covered entities. This approach should be used with caution.

The second way in which information can be de-identified by a covered entity is to have a person with certain skills and experience determine what modifications are needed so that it is acceptable to release the data. Just who can do this is indicated in the Rule not by identifying specific individuals or statistical organizations, but by describing the knowledge that must be employed – i.e., generally accepted statistical and scientific principles for assessing disclosure risk and implementing disclosure limitation.

Finally, protected data may be provided through the mechanism of a Limited Data Set.³⁶ In this case, the emphasis is not on de-identification, but on control of the circulation of the data. The release may contain geographic detail or variables central to the analysis, and known to be identifying of themselves or in combination, such as zip code, dates, and street address, but not personal identifiers of an individual.

Implications for Informed Consent

As HIPAA indicates, privacy does not require that subjects consent to the release (sharing) of sensitive data if it is de-identified. Moreover, the Common Rule does not regard de-identified data as human subjects data. Given the importance of keeping consent forms as brief and understandable as permissible, discussion of sharing and de-identification procedures in consent agreements may be a poor idea.

However, information useful for research is nearly always of a detailed nature and these very details which make analysis so fruitful are the same ones that make disclosure risk high. Therefore, it may be anticipated that sharing of information will, as a rule, involve potentially identifiable data. This is even more evident when sensitive, identified data such as videotapes of families or of dynamics within a certain group, are to be shared. In either case, informed consent is essential. Subjects should be told as much as possible about the parties with which identifiable data will be shared, whether the data will be archived, and how the archive will handle confidentiality. For example, if the data are archived at the Murray Center, any users of the data, their IRB, and other institutional officials must agree to many stipulations as spelled out on the Murray Center web site.³⁷ Such details might be usefully provided in the informed consent. If the researcher fails to plan ahead and does not include adequate detail about sharing identified data when consent is obtained, either study participants would need to be recontacted or the file de-identified – both procedures which are likely to result in a loss of data quality and usefulness.

What are the implications of sharing within one's institution or with trusted colleagues elsewhere? The answer to this question depends on such factors as the sensitivity and identifiability of the data, the degree of supervision of the recipient by the data donor, and, when the data are sensitive and identified, the kinds of agreements that the recipient and his/her IRB will enter into. In any event, the risks are far fewer than with public-use data which anyone might download from the internet.

Why are Ordinary Methods of Removing Identifiers Sometimes Insufficient?

It is often assumed that the removal of so-called “direct” identifiers such as name and street address is sufficient to render a data set “anonymous.” As the latter term implies, there would no longer be “names” associated with the data. It has been demonstrated repeatedly, however, that some research files contain more than enough information to lead an intruder back to name or address if he/she also has access to files containing direct identifiers whose information “overlaps” with that of the anonymized file. An example of how this process works is presented in the next section, but here we would like to stress that given such vulnerabilities, we need to consider the following two factors:

Motive. Data are a commodity and, for certain items, there may already be an active black market. Certainly, large public list searches are readily available.³⁸ Employers and insurance companies can save millions of dollars by knowing the health or genetic status of prospective employees or insurees. Re-identified data could show HIV+ status, bankruptcy, criminal behavior, or mental illness. While the simple desire by a hacker to embarrass a data holder is always a possibility, we are increasingly aware that there are other motives.

Opportunity. Record linkage technology continues to advance. The re-identification of data that ten years ago would have taken an unreasonable amount of time and resources can now be accomplished quickly and cheaply – the probability of re-identification can no longer be considered low because it would require too much work or money. Indeed, some data that were considered to be completely de-identified at that time now can be at least partially re-identified.

Given motive and opportunity on one hand, and ethical, legal, and scientific imperatives on the other, it is clear that today’s researcher needs to become aware of, if not relatively proficient in, the assessment of disclosure risk and methods to limit disclosure.

An Example of a Re-identification Technique

How can anonymous data be re-identified, and how does one prevent re-identification of one’s data? A simple example will illustrate why re-identification of sensitive data is of concern.

Let’s say we have a file, which we will label the “original research file,” and that we need to create a “de-identified” data set for sharing with other researchers or for release to the general public. Re-identification requires that one have independent knowledge of some details about persons in the research file and that this knowledge be associated with a name, address, or some other information that will lead to a name or address. There are many sources of such details. One is a data base, available on the internet, that contains records for 1.1 million marriages in the State of Kentucky from 1973 to 2001 (the Kentucky Marriage File). These records include name, age, race, residence, and number of prior marriages for both the bride and groom, as well as the date and county of the marriage and the marriage certificate number.

Suppose one has a data set from research done on residents of Kentucky. The names have been removed, but the data include gender, age, age of spouse, state of birth, and date of

marriage, as well as sensitive data about them, e.g., their health status, their race. The following example shows how re-identification of respondents might be done after the names have been removed.

Table 1: Re-identification Example	
File	File content
File with direct identifier (original research file)	Name ABCDEFGHIJKL
File with direct identifier removed (public research file)	ABCDEFGHIJKL

Table 2: Matching of Hypothetical Research File with External File Using Items Held in Common	
File	File content
Public research file (created in Table1)	ABC DEFGHIJKL
Material available from Kentucky Marriage File	ABCWXYZ Name

Table 3: Key to Variables Listed in Tables 1 and 2	
Alphabetic designation	Variable
A	Age, male
B	Age, female
C	Date of marriage
D	State of birth
E	Race
F	Race of spouse
G	HIV status of subject
H	HIV status of spouse

In Table 1, we removed the names and create a file for sharing (public research file). In Table 2, we show the two files, the public research file and the Kentucky Marriage File, and their variables.

Note that both files hold certain variables (bolded) in common. Under given circumstances, it may be concluded that an individual who has these characteristics in common is the same person in both data sets. And the name associated with these characteristics in the Kentucky Marriage File can be appended to the public research file. What's more, not only is a name associated with the information held in common, but the rest of the information in that file is now known to be associated with that person. In case the reader is wondering just how possible such a scenario might be, consider that when the people in the Kentucky data base are classified by age and date of marriage, fully *half* of these couples are unique. For those couples, the combination forms an identifier that matches to one, and only one, pair of names.

We are well accustomed to the idea that a name or address serves to identify a research participant; we need to be aware too, that certain *combinations of data* about them, can serve an enterprising intruder almost as well.

This is a simplified example of how the data in one file may be used, in conjunction with similar data in another file, to re-identify a research participant. Of course, actual situations are rarely this simple and a variety of factors must be considered as influencing any comparison or match between two files.

Evaluating Disclosure Risk of Microdata

Recall that microdata files are computerized files that consist of individual records, each containing values of variables for a single person, business establishment, or other unit. There are two distinct types of disclosure risk of microdata: (1) The data may be linkable to external data containing the identity of some respondents. (2) The data may include some highly visible and recognizable records (e.g. a 12 year old with a college degree). Guarding against these two problems is a critical step in data sharing.

To discover whether there are linkable external data that might be used to identify some respondents, one needs to review what other data have been released on these respondents, or what information is generally available to the public. E.g., if the research data are restricted to some sort of administrative unit, for instance state, county, school district, or hospital service area, data published for that area should be screened for variables that are also found in the research data. The variables in the research data likely to be found in external files are called key variables. These usually include demographic information like age and sex, but may also include more specific information like profession, educational degree, "is a registered voter," etc. In our hypothetical example in the preceding section, the key variables used to "link" the two files were age of wife, age of husband, and date of marriage.

The inclusion of *contextual* variables, i.e., those that describe some aspect of an area, can increase the risk of disclosure. Contextual variables are not part of the original data collection but are added to the research data set. For example, a Principal Investigator (PI) conducts a survey and is interested in the relationship of medical care to poverty. He/she knows the address for each of his/her respondents. Each address is then mapped to the census tract and the census tract to a poverty rate. To de-identify the file the PI deletes the respondents' names and addresses but includes two variables that pertain to their place of residence, i.e., 3-digit zip code and poverty rate. It is important to note that the "level of geography" may be more detailed than intended. The PI should do a careful analysis of disclosure risk to be sure that respondents cannot be identified.

Once one has constructed a list of key variables, how does one determine if they can be used for record linkage? This is the \$250 question. A useful back-of-the-envelope calculation to assess disclosure risk is based on the combination of key variables, i.e., the size of cross-classification of key variables must be more than 3 times the population. For instance, suppose a

study (of any size) involves physicians and the PI wants to include the key variables age, sex, marital status, specialty, and practice size in the research data file. First, the PI hypothesizes that the age range for practicing physicians is between 29 and 68 years of age -- this equals 40 categories. In the next step, he/she recodes two variables: the specialization is collapsed to 10 categories and the practice size is categorized to four (single, small, medium, and large). Computing the cross-classified key variables (age by sex by marital status by specialty by size) is $40 \times 2 \times 2 \times 10 \times 4 = 6,400$ cells. Therefore, in order for the average cell size to be 3, the state should have $6,400 \times 3$ or 19,200 physicians. Let's compare two states -- Nebraska and Texas. Nebraska has approximately 3,500 physicians. Consequently, a microdata file that includes these five key variables would more than likely uniquely categorize many Nebraska physicians and the data would be at risk of re-identification. On the other hand, Texas with 50,000 physicians should by and large be safe. With an average cell size of approximately 8 in the cross-classification of the key variables, most of the physicians in a Texas study cannot be distinguished from 3 or more other physicians that share those same characteristics. For Texas, most of the data are protected by force of numbers. There remain records with odd combinations of characteristics, but that is a separate problem. The back-of-the-envelope calculation indicates that the question of whether or not the key variables can be used for record linkage depends heavily on the size of the state's population of physicians and gives a preliminary determination of what size is adequate.

Researchers can locate highly visible and recognizable records by visually scanning the content of small data sets, or, for larger data sets, by cross tabulations of some or all key variables to discover unusual categories or combinations of categories of data. The point is to discover what cases are unique in the research data set and therefore potentially vulnerable to disclosure using an external source in which they are also unique. Some examples of highly visible records would be a billionaire in Nebraska, a paraplegic who has climbed Mount Everest, or a 70-year-old black female dentist (still practicing) with twin grandsons living with her. Other combinations are less noticeable but may be nonetheless distinguishable in external records. Some examples are the 13-year-old mother in a metropolitan area, the family with an income of \$500,000 or more in a small county, or a 21-year-old male who died from prostate cancer.

The evaluation of disclosure risk is generally beyond the purview of an IRB. Yet it is critical that IRB members be aware of such risks. As we illustrated in the preceding two examples, it is very important for researchers to inventory existing data sets in order to determine if their proposed data releases can be compromised and confidential data revealed. Proposals to share data should include plans for the publication of research results. Experts in assessing disclosure risk may need to be consulted.

Part 3 of this paper will describe two basic methods of protecting the confidentiality of data. IRBs need to be assured that appropriate steps have been taken to de-identify tabular or electronic data and in the next paper we suggest a three-pronged approach to achieving this goal. We also discuss data sharing agreements and the IRB's role in the challenges posed by data sharing.

Draft of manuscript to be submitted to IRB: Ethics & Human Research. Please do not cite or quote.

When Data Sharing is Required: III. Meeting the Challenge³⁹

Virginia de Wolf⁴⁰
Kerrville, TX

Joan E. Sieber⁴¹
California State University, Hayward

Philip Steel⁴²
United States Census Bureau

Alvan Zarate⁴³
National Center for Health Statistics

In this final article of our three-part series, we describe some approaches that can be used in meeting the challenges posed by the sharing of confidential data. First, we provide an overview of two basic methods of protecting the confidentiality of data. We then discuss data sharing agreements. The **concluding last** section summarizes issues to be considered by Institutional Review Boards (IRBs) and research administrators. A glossary of important technical terms concludes the paper. (**too many concludes**)

Basic Methods of Assuring Confidentiality

U.S. Federal statistical agencies have established practices and procedures that enable others to access and use the data that government agencies collect. Their goal is to provide useful statistical information to data users while assuring that the responses of individuals are protected. Over the past twelve years, several groups have examined the role of the Federal government as a “data steward” and summarized its contributions, including the Panel on Confidentiality and Data Access,⁴⁴ the Federal Committee on Statistical Methodology's Subcommittee on Disclosure Limitation,⁴⁵ and FCSM's Confidentiality and Data Access Committee.⁴⁶

The Panel on Confidentiality and Data Access provided generic labels for the two methods that U.S. Federal statistical agencies use to protect the confidentiality of data they collect. These are:

- Restricted data: Any of a variety of techniques may be used to restrict the *content* of the data prior to releasing it to the general public. These techniques are called *statistical methods to limit disclosure*.
- Restricted access: Techniques to restrict the *conditions of data access*; i.e., who can have access to the data, at what locations, with what supervision, and for what purposes.

In our first article we provided several alternative answers to the question "what is meant by data?" Many forms of data are shared, including instruments and instructions required to conduct the research. For the purposes of this section we will define *data* to mean the final research data which are usually in electronic form. Final research data means the recorded factual material that was the basis of the research publication and the documentation of those data. This is very similar to the definition of data used by NIH in its data sharing guidance.⁴⁷ Such data are frequently shared as a computerized file, called *microdata*, that consists of individual records, each containing values of variables for a single person, household, business establishment, or other unit.

It is beyond the scope of this article to discuss evaluation of disclosure risk of qualitative data such as video tapes, field notes, or case records. Typically, however, such data are not appropriate for public-use sharing, and need to be shared with qualified individuals under a restrictive data sharing agreement, as discussed later in this paper.

Restricted Data

Data may be restricted in many ways, yielding a wide range of data products that vary in usefulness to secondary users. There is always a trade-off between disclosure risk and data usefulness to secondary analysts.⁴⁸ Summarizing briefly, some of these methods are: deleting some variables, recoding categorical variables into larger categories, rounding off and truncating continuous variables, masking outliers, enlarging geographical areas, adding noise to the data, data swapping, rank swapping, and blurring (e.g., aggregating values across small sets of respondents and replacing a reported value with the mean of the group).

The type of data product to be released dictates the choice of methods used to limit disclosure.⁴⁹ Most quantitative data are released as either tables or microdata files. Non-quantitative data, such as ethnographies and case studies, also may contain data tables and descriptions that may allow re-identification. The following brief description of methods of reducing disclosure risk of tables is relevant to both quantitative and (primarily) qualitative data.

Restricting tabular data. The key issue in restricting the data in a table is to determine what is a "sensitive" cell. Two kinds of rules are used to identify sensitive cells:

- **Threshold rule:** A cell is sensitive if the number of respondents is less than some specified number (e.g., if a researcher or statistical consultant decides that all cells in a table must have at least five cases of HIV infection, then a cell with only two cases in a county would be "sensitive.").

Ginny, let's reorganize the above so that it makes its point more generically:

- **Threshold rule:** A cell is sensitive if the number of respondents is less than some specified number. Thus, if a researcher or statistical consultant decides that all cells in a table must have at least five cases (e.g., of HIV infection), then a cell with only two cases in a county would be "sensitive."

- **Dominance rule:** A cell is sensitive if a small number of respondents contribute a larger percentage of the total value of a cell (e.g., one establishment employs over 80% of the people in a certain industry in a given county).

Sensitive cells are protected via *suppression* – they are not published or shared as public-use data. Rather, the categories containing sensitive data are broadened to increase cell size, the geographic locations containing sensitive cells are eliminated (or in the case of qualitative data, carefully disguised), or the cells themselves are simply eliminated. Suppression of sensitive cells is called *primary suppression*.

In *secondary suppression*, each row and column must have at least two suppressed cells; otherwise an intruder could sum the cell values and subtract from the row or column total to obtain the value of the sensitive cell. Secondary suppressions are non-sensitive cells that are selected for suppression to assure that the respondent level data in sensitive cells cannot be estimated accurately.

Restricting microdata. Selection of the appropriate data restriction method depends in part on the nature of the data, and on the vulnerabilities of the respondents and the research institution. How much risk of re-identification can the respondents and the research institution tolerate? No restricted data method is entirely without risk. The only way to avoid disclosure risk is to gather no data, that is, to do no research!

For that matter, it is useful to recognize that long before the data are shared they are vulnerable to the risk that the investigator or research assistants will discuss their interesting cases with others in a way that leads to identification of the individuals in those cases. As in all other kinds of risk assessment, there are always ambient, everyday risks that we simply live with.

Knowledge of statistical methods to limit disclosure is not normally within the scope of an IRB's expertise. IRBs need to be assured that appropriate steps have been taken to de-identify tabular or electronic data. We suggest a three-pronged approach:

- First, IRBs should have a basic understanding of what is involved. To this end, we recommend that IRB members read a relatively short "primer" on the subject.⁵⁰
- In addition, the IRB should identify and consult experts in this area (e.g., a statistician, biostatistician, demographer, sociologist, epidemiologist, or health services researcher who is experienced in the use of data restriction techniques) to review what was proposed by researchers.
- As part of the review process, the IRB may wish to employ documentation in the form of a "disclosure limitation checklist" that would guide the researcher and inform IRB members as to steps taken for statistical protection and help in the determination of the effect of those procedures on data quality and whether additional expert review is needed.⁵¹

Normally, the concern for IRBs members should be whether the measures taken produce data that represents "minimal risk" to the study participants. Techniques are available for disclosure protection that can accomplish this objective. Most of them are quite straightforward and can be implemented by most data managers and/or researchers. If a researcher wants to provide unusual detail (e.g., day, month, and year of birth; age at death in days or months) for small areas or special populations, the data need special attention that only an expert can provide. Selection of the appropriate data restriction method depends in part on the nature of the data, the vulnerabilities of the respondents and the research institution, and on the effects of distortion of the data on the subsequent usefulness to secondary analysts.

Restricted Access

This refers to arrangements for controlling the purpose and location of data use, as well as the protection, re-disclosure, and final disposition of the data. These arrangements may include supervised use by a guest researcher at the principal researchers' institution, research data centers/enclaves, or a remote access system; alternatively, data are used under legally binding "licenses" at the secondary researchers' institution.⁵² A brief description of one of these restricted access methods, use via a research data center or data enclave, follows.

Data enclave or research data center. Sensitive quantitative data may be maintained in a secure archive (research data center or enclave). The enclave staff may perform requested analyses for the secondary user (for a fee). Alternatively, the secondary user may come to the archive and perform the desired analyses on a dedicated computer, with a disclosure review of all output, inspection of material to be removed from the site, and physical oversight by the resident staff. Typically, in either situation where the data remain in an enclave, the secondary user submits a research proposal and enters into a formal agreement about the work that is to be done and the kinds of output sought. The application is usually extensive, including personal identification and institutional affiliation, a current resume, dates of proposed use of the enclave, source of funding, a detailed account of why the data are need, and the parts of the data set that are needed. There may be limits on the types of analysis permitted and on any outside (linkable) data brought in by the secondary user. In some contexts, when the data are highly complex, the secondary user may stay at the site for a long period of time under a fellowship arrangement.

A data enclave may be established at the home institution of the investigator, or, in some cases, in an enclave under the auspices of the funding agency.

Review of output based on data shared in a data enclave. Shared data used under a restricted access agreement typically consists of an electronic file that will be analyzed by the researcher. Such "final research data" (microdata file) are used to generate statistical output (e.g., tables, or new statistical tests not performed by the original researcher). This output should be reviewed by those having expertise in statistical methods to limit disclosure. As we noted earlier, the research data centers funded by federal statistical agencies review analytic results before a researcher is allowed to take them off-site.⁵³

Data Sharing Agreements

Such agreements are used to impose appropriate limitations on users. For example, they would include the criteria for data access, stipulate confidentiality standards to ensure data security at the recipient site, and prohibit manipulation of data for purposes of identifying subjects. They typically also state that the recipient may not transfer the data to others, that the data are to be used only for research purposes, and that the proposed research will be reviewed by an IRB. Penalties are typically included for violating the terms of the agreement.

Many examples of data sharing agreements for specific data sets are available on the internet, e.g., the Agency for Healthcare Research and Quality's National Inpatient Sample,⁵⁴ the University of North Carolina's Russian Longitudinal Monitoring Survey,⁵⁵ and Center for Medicare and Medicaid Services' data that contains individual identifiers.⁵⁶

Joan and AI: Alternative for the preceding paragraph:

Ginny, Phil & AI, Let's use this second (in box) set of examples since the above examples are redundant to those cited by NIH in their web site. JS

Many examples of data sharing agreements for specific data sets are available on the internet, e.g., obtaining a restricted use data license with the National Center for Education Statistics⁵⁷ and using restricted access data files from the University of Michigan's Health and Retirement Study.⁵⁸

Data sharing agreements may be developed by individual researchers in consultation with their institutional research administration. Such agreements are also developed by archives which accept researchers' documented data and administer the sharing of it. For example, the Murray Research Center at Radcliffe College archives and administers the sharing of psychological data including developmental studies whose data include videotapes of individuals and families. The Murray Research Center has a process through which potential users apply for data.⁵⁹ Depending on the nature of the data sought, an appropriate data sharing agreement is reached with the borrower. Other useful models can be found on the web sites of the Research Data Centers sponsored by the federal statistical agencies.⁶⁰

Implications for IRBs and Research Administrators

Institutions that anticipate generating data that are to be shared need to consider how they will meet requirements for responsible and legal sharing. How is this done? The simple answer is that the major responsibility for creating an appropriate sharing arrangement falls to investigators. However, as every competent IRB chair or administrator knows, such activities work best when the IRB is knowledgeable and can offer useful guidance to naïve investigators. There are four areas in which guidance and shared understanding are important: costs of sharing, confidentiality measures, informed consent in relation to data sharing, and, in the case of medical data, compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule.

Costs. Researchers and research administrators, including IRBs, need to anticipate the costs of data sharing, which may include the cost of data documentation, preparation of restricted data, consultation with those having expertise to assess the risk of re-identification, and appropriate

preparation of items to be shared. There is little point in agreeing to share data if one does not budget for the cost of doing so properly. If data are to be archived at the investigator's home institution, appropriate space and staffing requirements must be met.

Confidentiality. The IRB needs to know about re-identification issues and whether the proposed restricted data or restricted access arrangements will reduce disclosure risk to an acceptable level. If this determination is beyond the expertise of IRB members, the IRB should consult with someone who is qualified to advise in this matter, such as a statistician who is knowledgeable of statistical disclosure limitation methods.

Informed consent. If data sharing plans have been developed responsibly to minimize the possibility that any subjects could be identified or re-identified, the data are not human subjects data and subjects' permission need not be obtained for sharing. However, if sharing of identifiable data, e.g., video tapes, is anticipated, it is prudent to inform subjects of sharing plans even if restricted access arrangements would preclude irresponsible use or identification of individuals by secondary users.

HIPAA. The IRB will want to understand quite precisely the implications of the HIPAA Privacy Rule for researchers.⁶¹ E.g., the Privacy Rule does not restrict the disclosure or sharing of medical data that has been properly de-identified.

While primary responsibility rests with investigators, who, in turn can obtain much relevant assistance from their funding agency via their program officer, the IRB and others in the institution's research administration still have considerable independent responsibility. Most funding agencies will be sensitive to the adequacy of the proposed data sharing plan. However, the institution's research administration nevertheless has a role in ensuring that proposals for projects that will involve data sharing provide a concise description of how the data will be documented and prepared for sharing, how risks of disclosure will be identified, and what means will be employed to restrict data or access, as appropriate.

Depending on the nature of the research and of the anticipated data, the identification of disclosure risks and selection of appropriate restricted data or restricted access methods typically requires special expertise. Those most likely to have such expertise or to be able to develop that expertise efficiently are statisticians or persons with considerable knowledge of statistics and research design (including some epidemiologists, sociologists, and health services researchers). For expertise in data linkage techniques that can be used in re-identification, an information or computer scientist^s might need to be contacted. The IRB and investigators, collaboratively, may need to develop a request of their research administration to provide these needed resources. Some institutions may wish to provide the necessary training to a small group of statisticians within the institution, and possibly others who have considerable competency in the area of quantitative research design. This group of experts could then comprise a Disclosure Review Board and, for example, could adapt the Confidentiality and Data Access Committee's "Checklist on Disclosure Potential of Proposed Data Releases" for use.⁶²

How can investigators, IRBs, and other research administrators gain the needed knowledge and expertise required to understand and oversee the work of a Disclosure Review Board, or an

expert who consults with the IRB on matters of preventing disclosure, without outside help? Fortunately, relevant information exists on the web. The American Statistical Association's Committee on Privacy and Confidentiality has created a comprehensive web resource which offers most of the information that one would need to develop and maintain expertise in methods of assuring the confidentiality of human research data that are to be shared.⁶³ In addition, the Confidentiality and Data Access Committee offers special training courses in these subjects, some of which are available on video.⁶⁴

Glossary

Anonymized data – Data that has had direct identifiers removed but may still be identifiable.

Confidentiality – agreements about who will have access to identifiable data. Such agreements may be implicit or spelled out in the informed consent.

Contextual variables – variables are those that describe some aspect of an area, such as a State, county, census tract, or block group; percent or frequency of the area's population employed, foreign born, receiving public assistance; number of health facilities; number and specialty of physicians; local government expenditures; measures of air quality; etc. (Also called ecologic variables.)

Data archive – a data “library” which stores the data and makes it available to others, typically along with the data documentation. There may be costs or conditions under which the data may be obtained. The data may be stored in various forms; currently most quantitative data are stored and transmitted electronically.

Data documentation – a description of the data that instructs secondary users how the data were obtained, how the variables are defined, how the data were cleaned and analyzed, and any other details that would be needed to make accurate use of the data.

Data enclave – a place where sensitive or irreplaceable data or artifacts may be used under conditions of supervision and limited access.

Deductive identification – using combinations of information to “snoop” and identify data.

Microdata – a computerized file that consists of individual records, each containing values of variables for a single person, business establishment, or other unit.

Non-disclosure methods – restriction of data or access to data to minimize risk of identification.

Re-identification – (a) The association of information about a study participant in a released file with the name of that participant; or (b) the identification of a study participant from information contained in a released file. Also see “deductive identification.”

Restricted data – Any of a variety of techniques may be used to restrict the *content* of the data prior to releasing it to the general public.

Restricted access – Techniques to restrict the *conditions of data access*; i.e., who can have access to the data, at what locations, with what supervision, and for what purposes.

Statistical disclosure limitation – the application of statistical techniques to transform data to limit the risk of disclosure. Some techniques are designed for data in tabular form, some for microdata. (Also called statistical disclosure avoidance or statistical disclosure control.)

Tabular data – grouped data in matrix form (data tables).

¹ The order of authorship is alphabetical.

² Virginia de Wolf, Consultant, may be reached at dewolf@ktc.com

³ Joan Sieber, Professor Emerita of Psychology, may be reached at jsieber@bay.csu Hayward.edu

⁴ Philip Steel, Statistician, and Disclosure Avoidance Staff Member, may be reached at Hphilip.m.steel@census.gov. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

⁵ Alvan Zarate, Confidentiality Officer, may be reached at aoz1@cdc.gov

⁶ E.g., see (1) Fienberg, SE. Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions. *American Review of Public Health* 1994;15:1-18; (2) Mackie, C, Bradburn, N, eds. *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. Washington, DC: National Academy Press, 2000. ([Hhttp://www.nap.edu/books/0309071801/html/H](http://www.nap.edu/books/0309071801/html/H); last accessed March 10, 2004); and (3) Sieber, JE, ed. *Sharing Social Science Data: Advantages and Challenges*. Newbury Park, CA: Sage Publications, Inc., 1991.

⁷ Documents pertaining to NIH's data sharing policy can be found at [Hhttp://grants2.nih.gov/grants/policy/data_sharing](http://grants2.nih.gov/grants/policy/data_sharing) (last accessed March 9, 2004).

⁸ U.S. Department of Health and Human Services. Office of Civil Rights. (August 2003). 45 CFR Parts 160 and 164: Standards for Privacy of Individually Identifiable Health Information; Regulation Text: (Unofficial Version. December 28, 2000 as amended: May 31, 2002, August 14, 2002, February 20, 2003, and April 17, 2003.). ([Hhttp://www.hhs.gov/ocr/combinedregtext.pdf](http://www.hhs.gov/ocr/combinedregtext.pdf); last accessed March 9, 2004).

⁹ For an example see the archiving policy of NSF's Social, Behavioral, and Economics Division ([Hhttp://www.nsf.gov/sbe/bcs/common/archive.htm](http://www.nsf.gov/sbe/bcs/common/archive.htm); last accessed March 9, 2004).

¹⁰ [Hhttp://www.icpsr.umich.edu/NACJDH](http://www.icpsr.umich.edu/NACJDH); last accessed July 3, 2004.

¹¹ Hamilton, DP. Data sharing: A Declining Ethic? *Science* 1990;248:952-958.

¹² Definition from page 3, U.S. Office of Management and Budget. *Report on Statistical Disclosure Limitation Methodology*. (Statistical Policy Working Paper 22). Washington, DC: May 1994. ([Hhttp://www.fcs.gov/working-papers/spwp22.html](http://www.fcs.gov/working-papers/spwp22.html); last accessed July 4, 2004)

¹³ For an extensive list of social science archives worldwide see [Hhttp://www.socsciresearch.com/r6.html](http://www.socsciresearch.com/r6.html) (last accessed March 9, 2004).

¹⁴ [Hhttp://www.icpsr.umich.edu/H](http://www.icpsr.umich.edu/H); last accessed March 9, 2004.

¹⁵ E.g., [Hhttp://www.census.gov/main/www/pums.html](http://www.census.gov/main/www/pums.html); last accessed March 9, 2004.

¹⁶ E.g., the National Center for Health Statistics ([Hhttp://www.cdc.gov/nchs/r&d/rdc.htm](http://www.cdc.gov/nchs/r&d/rdc.htm); last accessed March 9, 2004) and the U.S. Census Bureau ([Hhttp://148.129.75.160/ces.php/rdc](http://148.129.75.160/ces.php/rdc); last accessed July 3, 2004).

¹⁷ See Feinberg, SE, Martin, ME, Straf, ML, eds. *Sharing Research Data*. Washington, DC: National Academy Press, 1985.

¹⁸ Sieber, J. E., Sharing scientific data I: New problems for IRBs to solve, *IRB: A Review of Human Subjects Research* 1989;11(6):4-7.

-
- ¹⁹ [Hhttp://grants2.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html](http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html)H; last accessed March 9, 2004.
- ²⁰ Another NIH website, "NIH Data Sharing Policy and Implementation Guidance" provides a wealth of detailed information at [Hhttp://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm](http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm)H (last accessed March 9, 2004). This website is useful to any investigator, NIH applicant or not, who plans to share data.
- ²¹ See ref. 10.
- ²² See ref. 9.
- ²³ [Hhttp://www.radcliffe.harvard.edu/murray/H](http://www.radcliffe.harvard.edu/murray/H); last accessed March 9, 2004.
- ²⁴ For the websites of the NCHS Research Data Center and the Census Bureau's Research Data Center Program see ref. 16. For the AHRQ's CFact Data Center see [Hhttp://www.meps.ahrq.gov/datacenter.htm](http://www.meps.ahrq.gov/datacenter.htm)H (last accessed May 20, 2004).
- ²⁵ A description of remote access, as well as other types of restricted access used by federal statistical agencies, is available in the Confidentiality and Data Access Committee's publication entitled "Restricted Access Procedures" ([Hhttp://www.fcsm.gov/committees/cdac/cdacra9.pdf](http://www.fcsm.gov/committees/cdac/cdacra9.pdf)H; last accessed July 19, 2004).
- ²⁶ The order of authorship is alphabetical.
- ²⁷ Virginia de Wolf, Consultant, may be reached at dewolf@ktc.com
- ²⁸ Joan Sieber, Professor Emerita of Psychology, may be reached at jsieber@bay.csuhayward.edu
- ²⁹ Philip Steel, Statistician, and Disclosure Avoidance Staff Member, may be reached at [Hphilip.m.steel@census.gov](mailto:philip.m.steel@census.gov)H. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.
- ³⁰ Alvan Zarate, Confidentiality Officer, may be reached at aoz1@cdc.gov
- ³¹ [Hhttp://www.icpsr.umich.edu/access/deposit/index.html](http://www.icpsr.umich.edu/access/deposit/index.html)H; last accessed March 9, 2004.
- ³² See section 160.103 of HIPAA Privacy Rule: U.S. Department of Health and Human Services. Office of Civil Rights. (August 2003). 45 CFR Parts 160 and 164: Standards for Privacy of Individually Identifiable Health Information; Regulation Text: (Unofficial Version. December 28, 2000 as amended: May 31, 2002, August 14, 2002, February 20, 2003, and April 17, 2003.). ([Hhttp://www.hhs.gov/ocr/combinedregtext.pdf](http://www.hhs.gov/ocr/combinedregtext.pdf)H; last accessed March 9, 2004).
- ³³ It is ultimately the covered entity's decision whether to allow a researcher to use or disclose such information under a waiver of authorization. The waiver of authorization does not compel a covered entity to use or disclose protected health information.
- ³⁴ See section 164.514(b)(2)(B) of the Privacy Rule listed above in ref. 7.
- ³⁵ See section 164.514(b) of the Privacy Rule listed above in ref. 7.
- ³⁶ See section 164.514(e) of the Privacy Rule listed above in ref. 7.
- ³⁷ [Hhttp://www.radcliffe.harvard.edu/murray/H](http://www.radcliffe.harvard.edu/murray/H); last accessed March 9, 2004.
- ³⁸ For an example of kinds of public list searches one can buy, see [Hhttp://www.accurant.com/price2.html](http://www.accurant.com/price2.html)H (last accessed March 9, 2004).
- ³⁹ The order of authorship is alphabetical.
- ⁴⁰ Virginia de Wolf, Consultant, may be reached at dewolf@ktc.com
- ⁴¹ Joan Sieber, Professor Emerita of Psychology, may be reached at jsieber@bay.csuhayward.edu
- ⁴² Philip Steel, Statistician, and Disclosure Avoidance Staff Member, may be reached at [Hphilip.m.steel@census.gov](mailto:philip.m.steel@census.gov)H. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.
- ⁴³ Alvan Zarate, Confidentiality Officer, may be reached at aoz1@cdc.gov
- ⁴⁴ The Panel on Confidentiality and Data Access: (1) Duncan, GT, Jabine, TB, de Wolf, VA, eds. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Committee on National Statistics and the Social Science Research Council. Washington, DC: National Academy Press, 1993. ([Hhttp://books.nap.edu/catalog/2122.html](http://books.nap.edu/catalog/2122.html)H; last accessed March 10, 2004); (2) Jabine, TB. Procedures for restricted data access. *Journal of Official Statistics* 1993;9:2:537-589 ([Hhttp://www.jos.nu/Contents/issue.asp?vol=9&no=2](http://www.jos.nu/Contents/issue.asp?vol=9&no=2)H; last accessed March 10, 2004); and (3) Jabine, TB. Statistical disclosure limitation practices of United States statistical agencies. *Journal of Official Statistics* 1993;9:2:427-454 ([Hhttp://www.jos.nu/Contents/issue.asp?vol=9&no=2](http://www.jos.nu/Contents/issue.asp?vol=9&no=2)H; last accessed March 10, 2004).

-
- ⁴⁵ Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper #22. Washington, DC: Office of Management and Budget, May 1994 ([Hhttp://www.fcsm.gov/working-papers/spwp22.html](http://www.fcsm.gov/working-papers/spwp22.html)H; last accessed July 7, 2004)
- ⁴⁶ [Hhttp://www.fcsm.gov/committees/cdac/cdac.html](http://www.fcsm.gov/committees/cdac/cdac.html)H, last accessed March 9, 2004..
- ⁴⁷ [Hhttp://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm](http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm)H, last accessed July 5, 2004.
- ⁴⁸ E.g., Duncan, G.T. (2003). Confidentiality and Data Access Issues for Institutional Review Boards. Appendix E in Citro, CF, Ilgen, DR, & Marrett, CB. eds. *Protecting Participants and Facilitating Social and Behavioral Sciences Research*. Committee on National Statistics. Washington, DC: National Academy Press, pp. 235-252. ([Hhttp://www.nap.edu/catalog/10638.html](http://www.nap.edu/catalog/10638.html)H; last accessed July 31, 2004)
- ⁴⁹ E.g., see references 7 and 8.
- ⁵⁰ E.g., see Chapter 2, "Statistical Disclosure Limitation," of Statistical Policy Working Paper # 22 listed above in ref. 7.
- ⁵¹ The Confidentiality and Data Access Committee's "Checklist on Disclosure Potential of Proposed Data Releases" and its Checklist Overview" can be adapted for this use ([Hhttp://www.fcsm.gov/committees/cdac/checklist_799.doc](http://www.fcsm.gov/committees/cdac/checklist_799.doc);H [Hhttp://www.fcsm.gov/committees/cdac/overview.html](http://www.fcsm.gov/committees/cdac/overview.html)H, respectively; last accessed July 19, 2004).
- ⁵² A description of other types of restricted access used by federal statistical agencies is available in the Confidentiality and Data Access Committee's publication entitled "Restricted Access Procedures" ([Hhttp://www.fcsm.gov/committees/cdac/cdacra9.pdf](http://www.fcsm.gov/committees/cdac/cdacra9.pdf)H; last accessed July 19, 2004).
- ⁵³ E.g., see the U.S. Census Bureau's research proposal guidelines ([Hhttp://148.129.75.160/ces.php/guidelines](http://148.129.75.160/ces.php/guidelines)H; last access July 11, 2004). In the subsection "Proposal Review Process," the paragraph on "Risk of disclosure" begins by stating that "Output from all research projects must undergo and pass disclosure review."
- ⁵⁴ [Hhttp://www.ahcpr.gov/data/hcup/datause.htm](http://www.ahcpr.gov/data/hcup/datause.htm)H; last accessed July 7, 2004.
- ⁵⁵ [Hhttp://www.cpc.unc.edu/projects/rfms/data/commform.html](http://www.cpc.unc.edu/projects/rfms/data/commform.html)H; last accessed July 7, 2004.
- ⁵⁶ [Hhttp://www.cms.hhs.gov/data/requests/cmsdua.pdf](http://www.cms.hhs.gov/data/requests/cmsdua.pdf)H; last accessed July 7, 2004.
- ⁵⁷ [Hhttp://nces.ed.gov/statprog/confid5.asp](http://nces.ed.gov/statprog/confid5.asp)H, last accessed August 3, 2004.
- ⁵⁸ [Hhttp://hrsonline.isr.umich.edu/rda/index.html](http://hrsonline.isr.umich.edu/rda/index.html)H, last accessed August 3, 2004.
- ⁵⁹ [Hhttp://www.radcliffe.edu/murray/data/applicat.htm](http://www.radcliffe.edu/murray/data/applicat.htm)H, last accessed August 3, 2004.
- ⁶⁰ E.g., National Center for Health Statistics ([Hhttp://www.cdc.gov/nchs/r&d/rdc.htm](http://www.cdc.gov/nchs/r&d/rdc.htm)H ; last accessed July 7, 2004) and the U.S. Census Bureau ([Hhttp://148.129.75.160/ces.php/guidelines](http://148.129.75.160/ces.php/guidelines)H ;last accessed July 7, 2004)
- ⁶¹ A useful resource is: Muhlbaier, LH. *HIPAA Training Handbook for Researchers: HIPAA and Clinical Trials*. Marblehead, MA: HCPro, 2003. ([Hhttp://www.medstarresearch.org/departments/ora/HIPAA/documents/hipaahandbook.pdf](http://www.medstarresearch.org/departments/ora/HIPAA/documents/hipaahandbook.pdf)H ; last accessed July 7, 2004)
- ⁶² See ref. 12.
- ⁶³ See its *Privacy, Confidentiality, and Data Security* website ([Hhttp://www.amstat.org/comm/cmtepc/index.cfm](http://www.amstat.org/comm/cmtepc/index.cfm)H ; last accessed July 7, 2004).
- ⁶⁴ E.g., [Hhttp://www.fcsm.gov/committees/cdac/tutorials-courses.html](http://www.fcsm.gov/committees/cdac/tutorials-courses.html)H ; last accessed July 20, 2004.