Federal Data Sharing Requirements and Issues: Contributions to be Made by Statistics, Survey Research, and Related Disciplines

Virginia de Wolf
522 Oakland Hills Lane
Kerrville, TX 78028
dewolf@ktc.com

Key Words: NIH; NSF; HIPAA; Statistical Disclosure Limitation Methods; Accessing Confidential Data; Data Protection

As noted in a National Science Foundation (NSF) policy document, the sharing of data among researchers advances science and "... strengthens our collective capacity to meet scientific standards of openness by providing opportunities for further analysis, replication, verification and refinement of research findings" (NSF, no date). And, yet, when a researcher shares data collected under a pledge of confidentiality, it is of paramount importance that this pledge be honored. How can confidentiality be maintained and yet sharing safely be accomplished? How can researchers ensure that the data to be shared are used appropriately? How can investigators and grantees gain the necessary knowledge to accomplish this?

This paper's focus is the sharing of data collected under a pledge of confidentiality. The first section briefly outlines the data sharing requirements and issues that affect researchers obtaining grants from the National Institutes of Health (NIH) and NSF. It also describes the data sharing provisions contained in the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Researchers applying for NIH and NSF grants, as well as those requesting data under HIPAA, need to know about methods used to protect confidential data (termed "restricted access" and "restricted data"). The second section provides an overview of such methods. The last section contains the "meat" of this paper where I strongly encourage faculty with knowledge of these methods to pool their expertise and serve as resources for their colleagues. Ideas for collaborations are given.

This material in this paper pertains to the sharing of confidential data collected from individuals and organizations. By limiting it, I do not mean to imply that the sharing of other types of data is not important. This restriction merely reflects the knowledge base of the author.

Brief Review of NIH, NSF, and HIPAA Data Sharing Policies

Beginning in Fall 2003, the NIH required that all funding proposals requesting over $500,000 in direct costs in a year include specific plans for data sharing (Department of Health and Human Services [DHHS], National Institutes of Health, 2003). These plans must safeguard the privacy of participants and protect confidential and proprietary data.

NSF, another major federal source of research funding, has a policy that strongly encourages the sharing of data collected by its grantees (NSF, no date). The inclusion of data sharing provisions in NSF grant applications is viewed very positively by the agency's reviewers.[1]

HIPAA's Privacy Rule, "Standards for Privacy of Individually Identifiable Health Information," was issued in August 2003 (DHHS, Office of Civil Rights, 2003). This Privacy Rule permits the sharing of "protected health information" for use in research with or without patients' authorization when a waiver of authorization is approved by an Institutional Review Board. Waiver criteria include the treatment of identifiable data and restrictions as to their re-use and disclosure to others.

Methods for Protecting Confidential Data

Federal statistical agencies collect data under a pledge of confidentiality to inform public policy. These statistical agencies also have an obligation to disseminate as much data as possible while protecting the confidentiality of the information provided by respondents. The Panel on Confidentiality and Data Access characterized the two main options that these agencies have for protecting the confidentiality of

[1] de Wolf, V.A., Sieber, J.E., Steel, P., & Zarate, A.O. "When Data Sharing is Required: I. What is This Requirement?" Paper submitted for publication.

released data (Duncan et al., 1993; Jabine, 1993a; 1993b):

- Provide restricted data by restricting the *content* of the data prior to releasing it to the general public. A variety of techniques are used and are typically called *statistical methods to limit disclosure*. For an overview of restricted data methods, see the 1994 report of the Federal Committee on Statistical Methodology (FCSM). The European-based web site, Computational Aspects of Statistical Confidentiality (CASC),[2] is another valuable resource.

- Provide restricted access by restricting the *conditions of data access*, i.e., who can have access to the data, at what locations, for what purposes, etc. For example, research data centers have been established by several agencies (including the U.S. Census Bureau, National Center for Health Statistics, and Agency for Health Care Research and Quality). Researchers submit proposals to analyze data that are not released to the public. If the agency approves the project, then analyses are done at the agency's research data center. A description of the restricted access methods used by federal statistical agencies is available in the Confidentiality and Data Access Committee's (CDAC) web-based publication entitled "Restricted Access Procedures."[3]

Federal statistical agencies fund a lot of the research to develop these methods. Major contributions have been made by statisticians and survey researchers.

Impetus for this Paper

The idea for this paper came from remarks made by academic colleagues who were very interested in learning about statistical and administrative methods to limit disclosure in confidential data. One coworker assumed that all statisticians knew about statistical methods to limit disclosure -- I had to clarify and explain that this was not true. I referred another academic to the tutorial contained in Chapter 2 of the FCSM's 1994 *Report on Statistical Disclosure Limitation Methodology*. In a subsequent conversation, I was told that this material was "dense" and not as straightforward as I had implied. This past spring, at a meeting where restricted data and restricted access methods were discussed, a medical researcher objected to the fact that agencies would not release exact dates.

Given comments like these, I became aware that there was a problem -- we need to translate restricted access and restricted data methods into terms, examples, etc., that other disciplines can readily understand. Essentially, we need to speak their language to get our points across.

Sharing the Knowledge

Statisticians are not the sole group/discipline that has knowledge of restricted access/data approaches. While statisticians have been in the forefront of developing these methods, not all statisticians know about disclosure limitation methods. Researchers who use such methods can be found in disciplines that collect confidential data and include statistics, survey research, sociology, psychology, biostatistics, and health services research.

The main thrust of this paper is to strongly encourage individuals with this expertise to pool their knowledge. I envision such a collaborate effort would lead to the formation of interdisciplinary teams that would develop web sites about "data sharing methods/approaches." Each group would use available resources to create its web site. Such an interdisciplinary group could offer its services through a campus-based "data access committee" or "data sharing methods group." It would be important to involve graduate students.

Of course, there is no need to restrict such interdisciplinary committees to one campus. Such a committee could involve individuals at different campuses across the country. In fact, it might be "overkill" if every university developed such an interdisciplinary team. In this day-and-age, such a collaborative effort would be relatively easy via the internet.

Suggestions for such a web site follow:

- Clear exposition: Material should be written in "plain English" and avoid jargon. Speak in terms that other disciplines can understand.

[2] http://neon.vb.cbs.nl/casc; last accessed October 27, 2004.
[3] http://www.fcsm.gov/committees/cdac/cdacra9.pdf; last accessed October 27, 2004.

- Define important terms: This includes public-use files, secondary use of data, re-identification, and different types of disclosure (identity, inferential, attribute, etc.).

- Describe key ideas and concepts: Below are four suggestions:

  Example 1 -- There is no such thing as a zero-risk of disclosure: Zero-risk is an impossibly high standard. The only way to have a zero-risk is not to collect data. The goal is to make the risk of re-identification as low as possible. Elaborate on this.

  Example 2 -- Discuss "snoopers" and how to minimize the risk of re-identification: For instance, federal statistical agencies do not describe the exact procedures used to protect confidentiality nor do they release information about the risk of re-identifying respondents. They intentionally withhold this information from publication. But this has not been made clear to the academic community. For instance, in the Fall 2003 issue of the *ICPSR Bulletin* (O'Rourke, 2003) the author faulted CDAC's *Checklist on Disclosure Potential of Proposed Data Releases*[4] for not providing a mechanism for measuring disclosure risk. But such an omission was intentional since giving the details of the procedure(s) used by an agency to measure the risk of re-identification in its public-use data products could be used by a snooper to "unravel" the protections given to the confidential data. It is important to explain why this information is not published.

  Example 3 -- Include material that explicitly deals with confidential data collected from organizations: Show how such data differ from confidential data collected from individuals. The federal statistical agencies, almost without exception, do not release public-use files containing establishment data because of the high risk of re-identification -- explain why and give an example. For those who want to access microdata that contain data from organizations suggest restricted access approaches that could be used.

  Example 4 -- Describe how a Disclosure Review Board operates: Some federal statistical agencies have established special panels, called Disclosure Review Boards, that review data releases before they are made public. These Boards review microdata files and tables to determine if releasing the information to the public would conflict with the agency's confidentiality policy. It would be useful to provide examples of such Boards and suggest adaptations for the academic setting.

- Give examples of "high risk" variables: Tell researchers why certain variables (such as geography and dates) are "high risk" and could be used to re-identify respondents. Duncan's 2003 article contains a list of such variables. Use examples from several disciplines and describe the risks. Present alternatives. E.g., for certain types of analyses, such as time series, exact dates are needed. Describe how federal agencies provide access to such data though an agency-sponsored research data center.

- Describe important considerations when sharing data from small-scale research projects: What are some of the issues involved in sharing such data? Describe these issues.

- Include discipline-specific examples: Have examples that are specific to particular disciplines. For ex., anthropologists obtain NIH grants. In discussing the sharing of data collected under a NIH grant, it would be important to include data that anthropologists collect and in formats that are relevant. My suspicion is that their focus would be presenting data in tabular format and not in creating public-use microdata files. If this is true then I would recommend providing only tabular examples for the section that pertains to anthropologists. (Of course, the research data to be shared would be in electronic format so some attention needs to be given to this.)

- Use extant resources: Take materials that are already available and translate into terms, examples, illustrations, etc., that academics in other disciplines find useful. For example, consider these sources:

[4] http://www.fcsm.gov/committees/cdac/checklist_799.doc ; last accessed October 27, 2004.

1. *Committees*: Utilize the materials available from the ASA's Committee on Privacy and Confidentiality[5] and from CDAC.[6]

2. *Academe*: Give examples from non-federal sources, e.g., Inter-university Consortium for Political and Social Research,[7] University of Michigan's Health and Retirement Survey,[8] and the Murray Research Center at Radcliffe.[9]

3. *Publications*: Include relevant references, such as the recently published article in *IRB: Ethics & Human Research* entitled "Confidentiality: More than a Linkage File and a Locked Drawer" (Easter et al., 2004).

- Develop tutorials: Include modules for both tabular data and public-use microdata.

- Include links to relevant software: The European-based CASC web site contains information on two pieces of software: *mu-Argus*, a software program designed to create safe microdata files, and *tau-Argus*, a software program designed to protect statistical tables.[10] CDAC's web site contains its Disclosure Auditing Software that can be used on tabular output.[11]

- Give pertinent examples of what could go "wrong" when sharing data -- focus on re-identification: E.g., one example could illustrate how published tables from two researchers who used the same dataset could be combined and inadvertently reveal confidential information. Let's say Researcher A, the original grantee, shares data with Researcher B. Researcher B does an analysis and in his/her paper presents tables that are not reviewed by Researcher A. Show how confidentiality could be compromised and provide alternative(s) to avoid this.

- Suggest a set of "best practices": For instance, Researcher A shares data collected under a pledge of confidentiality with Researcher B. One best practice would be for Researcher A to include a provision in the data sharing agreement with Researcher B that Researcher A reviews all tables produced by Researcher B prior to publication. The purpose of this review would be to determine whether or not tables that were "related" (i.e., contained similar information), when viewed together, could lead to inferential disclosure of a respondent. Another potential best practice would be that Researcher A include a provision in the sharing agreement that he/she posts all analytic results on his/her web site.

- Frequently Asked Questions: Perhaps a useful component would be a series of Frequently Asked Questions (FAQs). One potential model for a FAQ web site is the NSF's web page "Interpreting the Common Rule for the Protection of Human Subjects for Behavioral and Social Science Research."[12]

- Use focus groups while developing the web site: Focus groups should be used at the start of the project to obtain feedback from academic colleagues about the existing documents. Learn what is not clear, what needs to be improved, etc. Once the web site is developed it should undergo extensive pretesting with academics from a wide range of disciplines (i.e., anthropologists, medical researchers, etc.) One potential source for participants in such focus groups would be members of NIH's advisory committees.

What about funding for such a web-based tutorial? While I am not conversant with the range of funding sources for academic research, several ideas come to mind. First, perhaps the federal agencies that encourage and foster such sharing would be willing to fund the development of such a web site or web sites. Alternatively, the major professional associations whose members would be likely to apply for grants from either NIH or NSF, or request data

[5] http://www.amstat.org/comm/pc/ASA-P&C-Committee-Home.htm; last accessed October 27, 2004.
[6] http://www.fcsm.gov/committees/cdac/cdac.html; last accessed October 27, 2004.
[7] http://www.icpsr.umich.edu/; last accessed October 27, 2004.
[8] http://hrsonline.isr.umich.edu; last accessed October 27, 2004.
[9] http://www.radcliffe.harvard.edu/murray/; last accessed October 27, 2004.
[10] See footnote 2.
[11] http://www.fcsm.gov/committees/cdac/DAS.html; last accessed October 27, 2004.
[12] http://www.nsf.gov/bfa/dias/policy/hsfaqs.htm; last accessed October 27, 2004.

under HIPAA, could pool resources and provide funding. Or, perhaps one of the IRB accrediting organizations, such as the Association for the Accreditation of Human Research Protections Programs,[13] might be a potential funder.

Concluding Remarks

The main intention in writing this paper was to start a discussion about how those attending this conference, statisticians and survey researchers, could aid colleagues in their efforts to share data. The web site(s) that I envision is(are) not intended to be a venue to present "cutting edge" work. While the world needs researchers who do such cutting edge work, it also needs "translators" -- which is my characterization of the interdisciplinary team that would create the proposed web site(s). I think of this work as a service to the broader academic community. I hope that these ideas in this paper are useful and start such a conversation.

The ASA's Committee on Privacy and Confidentiality has begun developing a web-based training module. Perhaps the ideas that I presented in this paper could be useful for this Committee as it proceeds.

References

Duncan, G.T. (2003). Confidentiality and Data Access Issues for Institutional Review Boards. Appendix E in Citro, C.F., Ilgen, D.R., & Marrett, C.B. eds. *Protecting Participants and Facilitating Social and Behavioral Sciences Research*. Washington, DC: National Academy Press, pp. 235-252. < http://www.nap.edu/catalog/10638.html; last accessed October 27, 2004>

Duncan, G.T., Jabine, T.B., & de Wolf, V.A., eds. (1993). *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. <http://books.nap.edu/catalog/2122.html; last accessed October 27, 2004>

Easter, M.M., Davis, A.M., & Henderson, G.E. (2004). Confidentiality: More than a Linkage File and a Locked Drawer. *IRB: Ethics & Human Research*, 26:2, 13-17. <http://www.thehastingscenter.org/pdf/publications/irb_sample_mar_apr_2004_confidentiality.pdf; last accessed October 27, 2004>

Federal Committee on Statistical Methodology. (May 1994). *Report on Statistical Disclosure Limitation Methodology*. (Statistical Policy Working Paper 22.) Washington, DC: Office of Management and Budget. <http://www.fcsm.gov/working-papers/spwp22.html; lasted accessed October 27, 2004>

Jabine, T.B. (1993a). Procedures for Restricted Data Access. *Journal of Official Statistics*, 9:2, 537-589. <http://www.jos.nu/Contents/issue.asp?vol=9&no=2; last accessed October 27, 2004>

Jabine, T.B. (1993b). Statistical Disclosure Limitation Practices of United States Statistical Agencies. *Journal of Official Statistics*, 9:2, 427-454. <http://www.jos.nu/Contents/issue.asp?vol=9&no=2; last accessed October 27, 2004>

National Science Foundation, Social, Behavioral, and Economics Division. (no date). Data Archiving Policy. <http://www.nsf.gov/sbe/bcs/common/archive.htm; last accessed October 27, 2004>

O'Rourke, J.M. (Fall 2003). Disclosure Analysis at ICPSR. *ICPSR Bulletin*, 24(1), 3-8.

U.S. Department of Health and Human Services, National Institutes of Health, Office of Extramural Research. (February 26, 2003). Final NIH Statement on Sharing Research Data. <http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>; last accessed October 27, 2004>

US Department of Health and Human Services, Office of Civil Rights. (August 2003). Standards for Privacy of Individually Identifiable Health Information. *Regulation Text*. (Unofficial Version.) (45 CFR Parts 160 and 164.). December 28, 2000 as amended: May 31, 2002, August 14, 2002, February 20, 2003, and April 17, 2003. <http://www.hhs.gov/ocr/combinedregtext.pdf; last accessed October 27, 2004>

[13] http://www.aahrpp.org ; last accessed July 27, 2004.