October 7, 2011

Appendix – SBS White Paper on **A**NPRM Re 45CFR46

Protecting Research Participants and Facilitating Responsible Data Access—Recommendations from the National Academies*

The social and behavioral sciences rely on data collected from people (human subjects in the language of the Common Rule) for much of their work. Detailed information about individuals (microdata) is critically important for research and essential in studies based on surveys. Social and behavioral sciences (SBS) modeling aimed at understanding relationships among complex phenomena is hampered without microdata because of the loss of information that occurs in aggregation. There is, however, an inescapable tension between providing access to microdata and the ethical obligation to protect the confidentiality of those who provided the data. This tension is not new and continues to evolve. On one hand, there is increased opportunity to develop greater understanding of the human condition and evidence-based solutions to some of the nation's most pressing problems by exploiting in more detail and in new ways the fruits of the nation's highly decentralized apparatus of data collection and research activities, which include data collected by federal statistical, research, and operating agencies, academic institutions, state and local governments, and the private sector, including social media. On the other hand the challenge of protecting data confidentiality has grown, principally because the proliferation of available data coupled with software developments and high speed computing have substantially increased the potential for re-identifying individuals in a data set even after direct or obvious identifiers have been removed (*Expanding Access to Research Data*, 2006:1-2).

For this reason government agencies and others involved with the SBS research communities have on several occasions turned to the National Academy of Sciences and its operating arm, the National Research Council (NRC), and to the Institute of Medicine (IOM), for guidance. This brief note summarizes relevant recommendations of expert study panels and committees commissioned by the NRC and IOM to assess and make recommendations on data access and confidentiality protection (a complete bibliography of relevant NRC/IOM reports is attached). These reports capture state of the art thinking about ethical and technical issues at the time of their writing. The burden of these recommendations is that there is no "one size fits all" regime for accomplishing the desired protection of confidential data while facilitating research access; that the issues in this area will continue to evolve; and that responsible organizations, including federal statistical agencies and major social science data archives, have been and should remain at the forefront in determining appropriate guidance, protocols, and methods for data access and confidentiality protection. The efforts of such agencies and archives provide models for protecting research data sets funded by the federal government. Hence, a theme running through the reports is that federal rules for data protection should not just allow for but encourage continual governmental and nongovernmental review and modification of existing standards in order to promote simultaneous progress toward both easier data access and greater subject protection. In this connection, a 2009 Institute of Medicine report makes clear that the HIPAA Privacy Rule (final 2002 version) proposed for possible use in a revision of the Common Rule (45 *CFR* 46, Part A) is not well suited for protecting data confidentiality and at the same time is a barrier to responsible research.[1]

[1]See "Advance Notice of Proposed Rule-Making for Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Researchers," *Federal Register*, 76 (43), Tuesday, July 26, 2011.

SHARING RESEARCH DATA (1985)

This early, pathbreaking report of the Committee on National Statistics made a detailed case for responsible sharing of research data at a time when such practices were not widespread. The report noted (p. 3) that: "Data are the building blocks of empirical research, whether in the behavioral, social, biological, or physical sciences. To understand fully and extend the work of others, researchers often require access to the data on which that work is based. Yet many members of the scientific community are reluctant or unwilling to share their data even after publication of analyses of them." To correct this situation the report included the following recommendations:

> Recommendation 1: Sharing data should be a regular practice.
>
> The advantages of data sharing [reinforcement of open scientific inquiry; verification, refutation, or refinement of original results; promotion of new research through existing data; encouraging more appropriate use of empirical data in policy formulation and evaluation; improvements of measurement and data collection methods; development of theoretical knowledge and knowledge of analytic technique; encouragement of multiple perspectives; provision of resources for training in research; protection against faulty data] are sufficient to warrant considerable attention to ways to share data without imperiling privacy or breaching the confidentiality promised to data providers.
>
> Recommendation 2: Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances.
>
> Recommendation 3: Data relevant to public policy should be shared as quickly and widely as possible.
>
> Recommendation 4: Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.
>
> Recommendation 8: Funding organizations should encourage data sharing by careful consideration and review of plans to do so in applications for research funds.
>
> Recommendation 9: Organizations funding large-scale, general-purpose data sets should be alert to the need for data archives and consider encouraging such archives where a significant need is not now being met.
>
> Recommendation 10: Journal editors should require authors to provide access to data during the peer review process.
>
> Recommendation 11: Journals should give more emphasis to reports of secondary analysis and to replications.
>
> Recommendation 13: Journals should strongly encourage authors to make detailed data accessible to other researchers.

Twenty-five years later, the recommendations in this report have been widely adopted by research funding organizations and scientific journals. As a result, the availability of rich microdata,

often from longitudinal panels, has greatly expanded, making possible more in-depth analysis of important topics in basic and applied SBS research. The case for data access has been made; the challenge is to identify responsible practices for data sharing that do not result in intentional or inadvertent breaches of data confidentiality and that do not make data access more onerous than necessary.

PRIVATE LIVES AND PUBLIC POLICIES (1993)

This report, prepared under the aegis of the National Research Council's Committee on National Statistics and the Social Science Research Council, focused on (p. 1): "developing recommendations that could aid federal statistical agencies in their stewardship of data for policy decisions and research. Three areas were of paramount concern . . .: protecting the interests of data subjects through procedures that ensure privacy and confidentiality, enhancing public confidence in the integrity of statistical and research data, and facilitating the responsible dissemination of data to users."

> Recommendation 4.2: Federal statistical agencies should seek to improve the access of external users to statistical data, through both legislation and the development and greater use, under carefully controlled conditions, of tested administrative procedures.
>
> Recommendation 5.1: Statistical records across all federal agencies should be governed by a consistent set of statutes and regulations meeting standards for the maintenance of such records including . . . a guarantee of confidentiality for data.
>
> Recommendation 5.3: There should be legal sanctions for all users, both external users and agency employees, who violate requirements to maintain the confidentiality of data. [This recommendation was subsequently endorsed in *Expanding Research Data* (2005), recommendation 8.]
>
> Recommendation 6.1: The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis . . . . Major statistical agencies should actively encourage and participate in scholarly statistical research in this area.
>
> Recommendation 6.2: Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers.
>
> Recommendation 6.4: Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.

This report was completed in a time when there was a large and growing body of microdata collected by statistical and other agencies of the federal government with great value for research; a growing recognition of the challenges in adequately protecting data confidentiality while not rendering data unusable for research; and a sense that well-thought-out guiding principles and practices and stronger legislation were needed because of the differences in statutes and practices then existing across agencies.

The 2002 Confidential Information Protection and Statistical Efficiency Act (CIPSEA) responded to the report's recommendations for legislation guaranteeing confidentiality for all federal statistical collections and providing penalties for breach of confidentiality applicable not only to agency staff, but also researchers who were provided access to microdata as agents of a statistical agency. The U.S. Office of Management and Budget (OMB) and many statistical agencies acted to develop best practices for data access and confidentiality protection and to establish or enhance different mechanisms for providing access under appropriate conditions to data that could not be provided in the form of public use microdata. Such mechanisms include research data centers, remote data access arrangements, and licensing of principal investigators at universities. See further discussion below under *Expanding Access to Research Data* (2005) and *Protecting and Accessing Data from the Survey of Earned Doctorates* (2010).

PROTECTING PARTICIPANTS AND FACILITATING SOCIAL AND BEHAVIORAL SCIENCES RESEARCH (2003)

This report of the National Research Council's Committee on National Statistics and Board on Behavioral , Cognitive, and Sensory Sciences responded to a growing concern in the social science community that (p. 1), "The U.S. system for protecting people who volunteer to participate in research is widely perceived to need improvement. A major concern is that the linchpins of the protection system—institutional review boards (IRBs)—are overloaded and underfunded and so may not be able to adequately protect participants from harm in high-risk research. . . . [while] the review process may delay research or impair the integrity of research designs, without necessarily improving participation protection, because the type of review is not commensurate with risk—for example, full board review for minimal-risk research that uses such methods as surveys, structured interviews, participant observation, laboratory experiments, and analyses of existing data." The report documented the variability in IRB workloads and practices (for example, treatment of expedited review that ranged from "never allowed" to "always allowed"). In addition to issues of informed consent and appropriate handling of minimal-risk research, the report gave substantial attention to the challenges of providing data access and protecting confidentiality of responses.

> Recommendation 5.1: Because of increased risks of identification of individual research participants with new methods of data collection. . . federal funding agencies should support research on techniques to protect the confidentiality of SBES data that are made available for research use; and the Office for Human Research Protections should regularly promulgate good practices in analyzing disclosure risks and limiting those risks.
>
> Recommendations 5.2 and 5.3: To facilitate secondary analysis of public-use microdata files, the Office for Human Research Protections . . . should establish a new confidentiality protection system for these data. . . . Participating [statistical agencies] and archives in the new public-use microdata system protection system should certify . . . whether data sets are sufficiently protected against disclosure to be acceptable for secondary analysis. IRBs should exempt such secondary analysis from review on the basis of the certification provided. [This recommendation was subsequently endorsed in *Expanding Access to Research Data* (2005), recommendation 6.]

Since the issuance of this report, several university IRBs have adopted a policy whereby microdata files that are released from statistical agencies and established data archives for public use are exempt from IRB review because they do not represent identifiable "human subjects". For example, the Purdue University IRB exempts from IRB review public use data sets from the following sources

(http://www.purdue.edu/research/vpr/rschadmin/rschoversight/humans/docs/101Existing_Public_Use_Datasets.pdf):

- Inter-University Consortium for Political and Social Research (ICPSR)
- Better Access to Data for Global Interdisciplinary Research (BADGIR)
- National Center for Health Statistics
- National Center for Education Statistics
- National Child Development Study
- National Election Studies
- Roper Center for Public Opinion Research
- University of Wisconsin-Madison Data and Information Services Center (DISC)
- U.S. Bureau of Census
- U.S. Bureau of Labor Statistics
- The University of Michigan Health and Retirement Study (HRS) (unrestricted data sets only)
- Panel Study of Economic Dynamics (PSID)
- Survey of Consumers (SCA)
- Integrated Public Use Microdata Samples – International (IPUMS-i)
- Luxembourg Income Study Project Archive

EXPANDING ACCESS TO RESEARCH DATA: RECONCILING RISKS AND OPPORTUNITIES (2005)

This report of the Committee on National Statistics was issued in response to a request from the National Institute on Aging for an assessment of "competing approaches to promoting exploitation of the research potential of microdata—particularly linked longitudinal microdata—while preserving confidentiality . . . and how microdata should be optimally (from a societal standpoint) be made available to researchers."

"The panel concludes that no one way is optimal for all data users or all purposes. To meet society's needs for high-quality research and statistics, the nation's statistical and research agencies must provide both unrestricted access to anonymized public-use files and restricted access to detailed, individually identifiable confidential data for researchers under carefully specified conditions" (p.2).

> Recommendation 2: Data produced or funded by government agencies should continue to be made available for research through a variety of modes, including various modes of restricted access to confidential data and unrestricted access to public-use data altered in a variety of ways to maintain confidentiality.
>
> Recommendation 3: The National Science Foundation, the National Institutes of Health, and major statistical agencies should support research to guide more efficient allocation of resources among different data access modes.
>
> Recommendation 5: Agencies that sponsor data collection should conduct or sponsor research on techniques for providing useful, innovative public-use data that minimize the risk of disclosure . . . [including]: (1) developing measures for quantifying disclosure risk; (2) estimating the effect on disclosure risk of adding selected variables from confidential data files to public-use files; (3) estimating and improving the utility-disclosure limitation tradeoffs of alternative disclosure limitation methods, including synthetic data . . . .

> Recommendation 9: To achieve the research potential and cost-effective operation of the Census Bureau [research] data centers, the Census Bureau should (1) broaden the interpretation of the criteria for assessing the benefits of access to data; (2) maintain the continuous review cycle; and (3) take account of prior scientific review of research proposals by established peer review processes.
>
> Recommendation 10: Statistical agencies and other agencies that sponsor data collection should conduct or sponsor research on cost-effective means of providing secure access to confidential data by means of a remote access mechanism, consistent with their confidentiality assurance protocols.
>
> Recommendation 11: Statistical and other agencies that provide data for research and do not yet use licensing agreements for access to confidential data should implement such an access mechanism. . . .

Since the issuance of this report, statistical and research agencies have expanded not only their use of innovative techniques for producing suitably anonymized public use files, but also their arrangements to provide access to identifiable confidential information through research data centers and/or licensing and/or remote access. For example, confidential data sets from the Health and Retirement Study sponsored by NIA are available on a restricted access basis at the University of Michigan, and the Census Bureau has expanded its RDC network and added data files from other agencies to the RDC holdings. This report recognized that what constitutes identifiable and not identifiable data is evolving and that different mechanisms are necessary to provide different kinds of access—for example, the provision of public use data sets with geographic identifiers may require perturbation of other data fields to safeguard against breach of confidentiality, while some confidential data can only be provided under restricted conditions such as through a license with stiff penalties for disclosure or other means.

PUTTING PEOPLE ON THE MAP: PROTECTING CONFIDENTIALITY WITH LINKED SOCIAL-SPATIAL DATA (2007)

This report of the National Research Council's Committee on the Human Dimensions of Global Change responded to the interests of the National Science Foundation, National Institute of Child Health and Human Development, and National Aeronautics and Space Administration in addressing the added challenges for providing data access and protecting confidentiality when very accurate spatial data (from remote sensing or GPS location devices) are included in social science data sets. The report concluded and recommended that:

> Conclusion 1: Recent advances in the availability of social-spatial data and the development of geographic information systems (GIS) and related techniques to manage and analyze those data give researchers important new ways to study important social, environmental, economic, and health policy issues and are worth further development.
>
> Conclusion 2: The increasing use of linked social-spatial data has created significant uncertainties about the ability to protect the confidentiality promised to research participants. Knowledge is as yet inadequate concerning the conditions under which and the extent to which

the availability of spatially explicit data about participants increases the risk of confidentiality breaches.

Recommendation 7: Data enclaves deserve further development as a way to provide wide access to higher-quality data while preserving confidentiality. This development should focus on the establishment of expanded place-based enclaves, "virtual enclaves," and meaningful penalties for misuse of enclaved data.

Recommendation 8: Data stewards should develop licensing agreements to provide increased access to linked social-spatial datasets that include confidential information.

The value of and issues raised by linked social-spatial data are obvious. Substantial additional research is needed to determine the risks of confidentiality and privacy breaches and to develop techniques for reducing those risks.

CONDUCTING BIOSOCIAL SURVEYS: COLLECTING, STORING, ACCESS, AND PROTECTING BIOSPECIMENS AND BIODATA (2010)

This report of the National Research Council's Committee on National Statistics and Committee on Population responded to the request of the National Institute on Aging to consider the added challenges for data access and confidentiality protection from the addition of biospecimens, such as blood, urine, and saliva, as part of large-scale household surveys intended for SBS research. The report cites best practice reference documents, including *Best Practices for Repositories: Collection, Storage, Retrieval, and Distribution of Biological Materials for Research*, prepared by the International Society for Biological and Environmental Repositories (2008); *National Cancer Institute Best Practices for Biospecimen Resources* (2007); and *OECD Best Practice Guidelines for Biological Resource Centres* (2007) and makes recommendations for data sharing plans and confidentiality protection for biosocial data sets:

Recommendation 3: [NIH] should publish guidelines for principal investigators containing a list of points that need to be considered for an acceptable data sharing plan. In addition to staff review, Scientific Review Panels should read and comment on all proposed data sharing plans. In much the same way as an unacceptable human subjects plan, an inadequate data sharing plan should hold up an otherwise acceptable proposal.

Recommendation 4: NIA and other relevant funding agencies should support at least one central facility for the storage and distribution of biospecimens collected as part of the research they support.

Recommendation 7: Both rich genomic data acquired for research and sensitive and potentially identifiable social science data that do not change (or change very little) with time should be shared only under restricted circumstances, such as licensing and (actual or virtual) data enclaves.

Recommendation 8: [NIH] should develop new standards and procedures for licensing confidential data in ways that will maximize timely access while maintaining security and that can be used by data repositories and by projects that distribute data.

PROTECTING AND ACCESSING DATA FROM THE SURVEY OF EARNED DOCTORATES: A WORKSHOP SUMMARY (2010)

This summary of a workshop convened by the Committee on National Statistics responded to a request of the NSF National Center for Science and Engineering Statistics for expert discussion of confidentiality protection techniques for tabular data from the NSF Survey of Earned Doctorates. While not designed to produce consensus conclusions or recommendations, the workshop highlighted the issues. It also identified and discussed valuable guidance on disclosure protection from:

- U.S. Office of Management and Budget, *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002* (June 2007, http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf);
- U. S. Office of Management and Budget, *Report on Statistical Disclosure Limitation Methodology*, Federal Committee on Statistical Methodology, Statistical Working Paper #22 (2006, http://www.fcsm.gov/working-papers/SPWP22_rev.pdf); and
- Professional association guidelines, such as the American Statistical Association's statement, *Data Access and Personal Privacy: Appropriate Methods of Disclosure Control* (December 6, 2008, http://www.amstat.org/news/statementondataaccess.cfm).

BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH (2009)

This report of the Institute of Medicine responded to a request from the National Institutes of Health, the National Cancer Institute, the Robert Wood Johnson Foundation, the American Cancer Society, the American Heart Association/American Stroke Association, the American Society for Clinical Oncology, the Burroughs Welcome Fund, and C-Change for a study committee to undertake two tasks:

> (1) to assess whether the HIPAA Privacy Rule is having an impact on the conduct of health research, defined broadly as "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge"; and (2) to propose recommendations to facilitate the efficient and effective conduct of important health research while maintaining or strengthening the privacy protections of personally identifiable health information.
>
> The committee concluded the following:
>
> The HIPAA Privacy Rule does not protect privacy as well as it should, and that, as currently implemented, the HIPAA Privacy Rule impedes important health research. The committee found that the Privacy Rule (1) is not uniformly applicable to all health research, (2) overstates the ability of informed consent to protect privacy rather than incorporating comprehensive privacy protections, (3) conflicts with other federal regulations governing health research, (4) is interpreted differently across institutions, and (5) creates barriers to research and leads to biased research samples, which generate invalid conclusions.

Recommendation I. Develop a New Approach to Protecting Privacy in All Health Research

The committee's first and foremost recommendation (Recommendation I) is that Congress should authorize HHS and other relevant federal agencies to develop a new approach to protecting privacy in health research that would apply uniformly to all health research. When this new approach is implemented, HHS should exempt health research from the HIPAA Privacy Rule. The new approach should enhance privacy protections through improved data security, increased transparency of activities and policies, and greater accountability, while also allowing important health research to be undertaken with appropriate oversight. The new approach should do all of the following:

- Apply to any person, institution, or organization conducting health research in the United States, regardless of the source of data or funding.
- Entail clear, goal-oriented, rather than prescriptive, regulations.
- Require researchers, institutions, and organizations that store health data to establish strong data security safeguards.
- Make a clear distinction between the privacy considerations that apply to interventional research and research that is exclusively information based.
- Facilitate greater use of data with direct identifiers removed in health research, and implement legal sanctions to prohibit unauthorized reidentification of information that has had direct identifiers removed.
- Require ethical oversight of research when personally identifiable health information is used without informed consent. HHS should develop best practices for oversight that should consider:
    - o Measures taken to protect the privacy, security, and confidentiality of the data;
    - o Potential harms that could result from disclosure of the data; and
    - o Potential public benefits of the research.
- Certify institutions that have policies and practices in place to protect data privacy and security in order to facilitate important largescale information-based research for clearly defined and approved purposes, without individual consent.
- Include federal oversight and enforcement to ensure regulatory compliance.

---

*NOTE: The material in this document was prepared by National Academies staff from published reports of study panels and committees. It is faithful to those reports and does not go beyond them.

Reports on Privacy and Confidentiality and Access to Research Data from the National Academies

The challenge of providing for data access while protecting privacy, ensuring confidentially, and minimizing risk of advertent or inadvertent disclosure has engaged the attention of federal agencies, the National Academies, and the scientific community for over three decades. Below is a chronological list of major reports from the National Research Council and the Institute of Medicine.

National Research Council. (1979). *Privacy and confidentiality as factors in survey response*. Washington, DC: National Academy of Sciences.

National Research Council. (1985). *Sharing research data*. Washington, DC: National Academy Press.

National Research Council. (1993). *Private lives and public policies: Confidentiality and accessibility of government statistics*. Washington, DC: National Academy Press

National Research Council. (2000). *Improving access to and confidentiality of research data: Report of a workshop*. Washington, D.C.: National Academy Press.

Institute of Medicine. (2000). *Protecting data privacy in health services research*. Washington, DC: National Academy Press.

Institute of Medicine. (2002). *Responsible research: a systems approach to protecting research participants.* Washington, DC: The National Academies Press.

National Research Council. (2003). *Protecting participants and facilitating social and behavioral sciences research.* Washington, DC: The National Academies Press.

National Research Council. (2005). *Expanding access to research data: Reconciling risks and opportunities.* Washington, DC: The National Academies Press.

Institute of Medicine. (2006). *Effect of the HIPAA privacy rule on health research: Proceedings of a workshop presented to the National Cancer Policy Forum*. Washington, DC: The National Academies Press.

National Research Council. (2006). *Improving business statistics through interagency data sharing: Summary of a workshop*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Putting people on the map: Protecting confidentiality with linked social-spatial data*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Engaging privacy and information technology in a digital age.* Washington, DC: The National Academies Press.

National Research Council. (2007). *Understanding business dynamics: An integrated data system for America's future.* Washington, DC: The National Academies Press.

Institute of Medicine. (2009). *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. Washington, DC: The National Academies Press.

National Research Council. (2009). *Ensuring the integrity, accessibility and stewardship of research data in the digital age*. Washington, DC: The National Academies Press.

National Research Council. (2009). *Principles and practices for a federal statistical agency*. Washington, DC: The National Academies Press.

National Research Council. (2009). *Protecting student records and facilitating education research: A workshop summary.* Washington, DC: The National Academies Press.

National Research Council. (2010). *Conducting biosocial surveys: Collecting, storing, accessing, and protecting biospecimens and biodata.* Washington, DC: The National Academies Press.

National Research Council. (2010). *Protecting and accessing research data from the Survey of Earned Doctorates: a research summary*. Washington, DC: The National Academies Press.