Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future

Julia Lane and Claudia Schur

Background. There has been a dramatic increase in the types of microdata, and this holds great promise for health services research. However, legislative efforts to protect individual privacy have reduced the flow of health care data for research purposes and increased costs and delays, affecting the quality of analysis.

Aim. This paper provides an overview of the challenges raised by concerns about data confidentiality in the context of health services research, the current methodologies used to ensure data security, and a description of one successful approach to balancing access and privacy.

Materials and Methods. We analyze the issues of access and privacy using a conceptual framework based on balancing the *risk* of reidentification with the *utility* associated with data analysis. The guiding principle should be to generate released data that are as close to the maximum acceptable risk as possible. HIPAA and other privacy measures can perhaps be seen as having had the effect of lowering the “maximum acceptable risk” level and rendering some data unreleasable.

Results. We discuss the levels of risk and utility associated with different types of data used in health services research and the ability to link data from multiple sources as well as current models of data sharing and their limitations.

Discussion. One particularly compelling approach is to establish a remote access “data enclave,” where statistical protections are applied to the data, technical protections ensure compliance with data-sharing requirements, and operational controls limit researchers’ access to the data they need for their specific research questions.

Conclusion. We recommend reducing delays in access to data for research, increasing the use of remote access data enclaves, and disseminating knowledge and promulgating standards for best practices related to data protection.

Key Words. Administrative data uses, confidentiality/privacy issues, health policy/politics/law/regulation, research ethics/institutional review boards/publication, dissemination issues

The health care reform agenda of restraining costs, improving quality, and enhancing population health will require data to guide decision making. The

most useful data for health services research are microdata—collected on individuals or households; only microdata allow the estimation of the impact of different policies or clinical protocols on individuals with relatively rare conditions, specific demographic profiles, or who are subject to a unique delivery system.

The dramatic increase in the types of microdata—from biological to geospatial—holds great promise for health services research. And there are new ways of combining data from different sources: information scientists have better technical ways of protecting data; statisticians have better statistical techniques; and social scientists have created virtual collaboratories to promote safe, secure access (Lane 2007).

However, legislative efforts to protect individual privacy have reduced the flow of health care data for research purposes and increased costs and delays, affecting the quality of analysis. The Privacy Rule in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) both created new procedural requirements for the use of protected health information (PHI) and defined PHI more broadly than had been done in the past (American Statistical Association 2009).

The American Recovery and Reinvestment Act of 2009 directs the new National Coordinator for Health Information Technology to pay attention to *both* data access and data confidentiality. This paper provides an overview of the challenges raised by concerns about data confidentiality in the context of health services research, the current methodologies used to ensure data security, and a description of one successful approach to balancing access and privacy.

A CONCEPTUAL FRAMEWORK FOR DATA ACCESS AND PRIVACY

The basic tension between data access and data confidentiality in the social science context is well understood (Doyle et al. 2001). The core challenge is balancing the *risk* of reidentification with the *utility* associated with data analysis.

Address correspondence to Julia Lane, Ph.D., National Science Foundation, 4201 Wilson Blvd, Arlington, VA 22230; email: jlane@nsf.gov. Claudia Schur, Ph.D., is with the Center for Health Research and Policy, Social & Scientific Systems, Silver Spring, MD.

The risk from reidentifying individuals can be measured by the likelihood that a record can be matched to a master file (Winkler 2005). If the data include direct identifiers, like names, social security numbers, or establishment id numbers, the risk is high. The harm associated with reidentification can be financial (disclosure might lead to denial of insurance coverage, job loss, or lack of job offer) and psychosocial (e.g., revelation of personal information leading to stigma in a social or work circle, or loss of reputation resulting in isolation or difficulty obtaining employment).

Access to microdata generates utility by increasing the value of the analytical work that can be undertaken, the likelihood of improvements in data quality, and the opportunity for replication of research (Duncan et al. 2001). Unfortunately, increased access simultaneously increases the risk of reidentification. As a result, data producers typically use some disclosure limitation techniques to protect individuals from being identified. Figure 1 provides a graphical representation of this conceptual tradeoff between data risk and data utility. Here the dashed line identifies the maximum acceptable risk; the core guiding principle should be to generate released data that are as close to the frontier as possible (Duncan et al. 2001).

The maximum acceptable risk for federally collected data is determined by the legal mandate of the agency that collected the survey, the Privacy Act of 1974 (5 U.S.C. 552a), and under the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) by the agency's interpretation of taking "reasonable means" to protect data confidentiality. While the privacy

Figure 1: The Duncan/Fienberg Risk-Utility Trade-off

rule does not generally apply to health services researchers (except where they are employed by a covered entity or provide health care during a clinical trial), there have been a number of rulings to give guidance to researchers and data providers about how to reduce risk. HIPAA regulations under The Privacy Rule of 2003¹ require the removal of 18 different types of identifiers (Gunn et al. 2004).

HIPAA can perhaps be seen as having had the effect of moving the “maximum acceptable risk” level down toward the lower dashed line in Figure 1, and, unless action is taken to move the characteristics of the data to the right, rendering some data unreleasable. In addition to HIPAA’s direct effects, adverse publicity surrounding its implementation and misinterpretation of the requirements by providers may result in increased difficulty in gaining permission from individuals or organizations to release data. This may lead to fewer research studies or studies that are less scientifically robust. In a national survey of clinical scientists, only a quarter perceived that the rule has enhanced participants’ confidentiality and privacy, whereas the HIPAA Privacy Rule was perceived to have a substantial, negative influence on the conduct of human subjects health research, often adding uncertainty, cost, and delay (Ness 2007).

OVERVIEW OF THE CURRENT ENVIRONMENT

As data capture and computing capabilities have become more sophisticated, the types of data used in health services research and the ability to link data from multiple sources has expanded, with different associated levels of risk and utility.

Survey Data

The *utility* of survey data collected from individuals lies in its detail on socio-demographic characteristics as well as perceptions, preferences, health behaviors, and health risks. However, survey data are less useful for examining specific medical conditions or expenditures, because of inaccurate reporting (Berk, Schur, and Mohr 1990). Surveys are also costly and increasingly subject to low response rates (Corey and Freeman 1990; Machlin et al. 2009). The *risk* associated with survey data lies in the rich contextual information—inclusion of detail on geography and personal characteristics that may be sufficient to reidentify respondents. Because survey data are collected on a population sample, risk of reidentification is somewhat limited.

Administrative Data

Administrative data are compiled primarily to administer a program or provide a benefit, and they include both private and public payer claims records as well as hospital discharge data.

The *utility* of administrative data lies in the large numbers of observations, permitting the study of small groups with statistical precision, and the greater reliability of certain fields—for example, diagnostic codes from providers or charges used for billing purposes (Berk et al. 1990; Machlin et al. 2009). While Medicare or Medicaid claims are only available for persons covered by those programs, hospital discharge data are available regardless of payer (or uninsured) status. On the other hand, key elements such as race/ethnicity can be missing, since the data are collected for programmatic, rather than analytical purposes.

The nature of disclosure *risk* with administrative data differs from survey data; because the data are universal, a record that links uniquely is reidentified with certainty. And, because the program agency retains the administrative file, there is the possibility of reidentifying the individual for nonanalytic purposes.

Linked Administrative and Survey Data

Survey data are often linked to administrative records, such as provider billing records, medical records, claims data, or employer information (Lane 2010).

The *utility* of linked survey and administrative data is substantial. Linkage can increase the accuracy of reporting, reduce bias from survey nonresponse, and expand the analytic time horizon. For example, the National Health Interview Survey has been linked to mortality data, turning a cross-sectional survey into a quasi-longitudinal dataset. The Medicare Current Beneficiary Survey linked to beneficiary claims has been used to study, for example, how different benefits structures affect utilization of services or the effectiveness of a given treatment by race/ethnicity. Researchers have linked data from the Health and Retirement Survey (HRS) to Medicare claims records and W-2 tax forms to address the interrelationships between earnings (from tax data), health conditions and personal attributes (from HRS), and health care use and expenditures (from claims).

The *risk* associated with combining the rich contextual survey information with administrative records is greater than from either type of data alone. The MCBS is considered a limited dataset; when linked to claims records, it becomes a research identifiable file and obtaining it becomes somewhat more arduous. The *maximum acceptable risk* can be a major challenge to define, since

typically multiple legal requirements cover the use of linked data. Often, it is the intersection, rather than the union, of the different requirements that governs the definition.

Clinical Data

Data from medical records or disease registries include clinical measures such as blood levels, laboratory test results, or indicators of cancer stage. Utility derives from more robust assessment of quality of care and health outcomes. Where electronic medical records are available, clinical data are easier to access though riskier from the privacy perspective. Where data are not electronic, medical abstractors must extract data from individual records, a complex and costly process. Disease registries track individuals with a specific disease; their *utility* comes from supporting the analysis of factors affecting disease incidence, prevalence, and survival.

The higher level of *risk* associated with clinical data is due to the increased harm with reidentification, since data may include highly personal details on disease and treatments, such as stigmatizing diagnoses or pharmaceutical use that reveals sensitive health conditions. Clinical records themselves may contain personal identifiers, increasing the perception of potential for risk.

Social-Spatial Data

The increasing sophistication of technology and geographic information systems has expanded the use of social-spatial data—usually contextual data describing neighborhoods or other small geographic areas.

The *utility* of such data is that many insights can be derived from the contextual variables surrounding individuals—the schools they go to, the neighborhoods they live in, the firms they work for, etc., and even the people they interact with. Data can encompass environmental indicators such as air pollution levels (to study health effects), local measures of health care supply (to analyze impact on health care use), counts of fast food retailers (to determine associations with obesity), or community health behaviors (to understand social network effects). Yet there is substantial *risk* from the use of geocodes (such as latitude–longitude coordinates) rather than addresses or political units. Publicly available data based on real property records—such as lot size or property tax maps—can lead to reidentification of individuals, and technological advances such as global positioning system (GPS) instruments and satellite technology have made it easier to link location-specific data at the household or neighborhood level and reidentify individual respondents (Balk

2003). Indeed, the capacity to study the interrelationships among social, demographic, neighborhood, environmental, health supply, and other contextual factors may be essential to advancing our understanding, but it raises to an even greater level the red flags of confidentiality breaches.

CURRENT ACCESS

The most common strategies used to balance access and privacy both reduce utility and risk, although there are few studies of the types, frequency, and consequences of confidentiality breaches.

Public Use Files

Some federally sponsored survey datasets are available as public use files, which contain somewhat restricted versions of the survey data. Geographic information and dates are the most frequent types of data to be omitted from public use files: MEPS, NHIS, and NAMCS omit state and county identifiers. MEPS public use files do not include fully specified ICD-9, industry, or occupation codes, or asset information. AHRQ's hospital discharge data exclude information on hospital admission, discharge, or procedure dates. These constraints limit data *utility*; for example, service dates are required to construct episodes of care, the basis for comparing overall resource use or charges for patients with a given illness subject to different treatment approaches. Other data protection approaches, such as topcoding, can lead to biased coefficients and reduced statistical precision.

Research Data Centers

Research data centers—both on-site and remote access—provide access to data in a controlled physical or electronic environment. Data *utility* is reduced by the burdens on users and the reduction in the number of researchers with data access. The processes used by the data centers vary, but NCHS, AHRQ, and the Census Bureau RDCs require an approval process. The Census Bureau's process is especially onerous: approval requires demonstrating that the research will benefit the Bureau's programs and, from submission of a proposal to actual data use, takes a minimum of 6 months. Authorized researchers can access the full range of existing data items but must either submit code electronically to process the data or must physically sit in a secure space. The NCHS data center can be used on-site (at several locations) or remotely, while the AHRQ has only an on-site component. Microdata from the insurer/

employer component of MEPS are only available through the Census Bureau RDCs.

Licensing Arrangements or Data Use Agreements

Licensing entails a signed agreement between an agency and the external researcher, permitting access to data files using a defined set of protocols at their home institution. The impact on data utility is through the time and financial burdens on the researcher. For Medicare beneficiary claims, purchase of the data is not only quite costly but requires a fairly extensive application, review, and approval process. Data must be obtained through a CMS contractor (the Research Data Assistance Center), and there is a 14-item checklist that must be submitted.

NEW DATA AND NEW ACCESS

Demand for microdata access also results from newly available data. Innovative approaches to collecting biomedical markers in the home—either through self-collection or using trained data collection personnel—is expanding the reach of health services research. An NIA-funded study at the University of Chicago collects data on sensory function and sexual health as part of an investigation into the role that social networks play in aging and health (Waite et al. 2009). On a much broader scope, the National Children's Study will collect a variety of biological material from over 100 communities nationwide to study the developmental health of children.

These new types of personally identifiable information, which also include those collected by means of sensors, video imaging, and texts, cannot be provided by means of public use files; licensing agreements are too insecure and risky; and research data center access is too slow, difficult, and costly to be a generalizable solution (Guttman and Stern 2007).

One approach is to create synthetic public use data files that add systematic noise to the microdata (Abowd and Lane 2004). The effect of disclosure protection on data quality can be measured. The *risk* is reduced since the synthetic data record does not reflect the respondent's actual data record, so identity (but not attribute) disclosure is impossible. Data *utility* may be reduced: it may not be possible to study small subgroups or examine outliers and the typical user may not be able to use the dataset correctly. Also, synthetic data take a very long time to generate, since there are very few people trained to create such files.

Alternatively, many national and international statistical agencies (such as the U.S. Economic Research Service and the National Science Foundation, the U.K. Office of National Statistics, Statistics Netherlands, and Statistics Sweden) have moved toward secure remote access entities as a way to promote researcher access. These entities, often called “data enclaves,” have a portfolio approach to protecting confidentiality, based on the notion that safe data result from safe projects, safe people, safe settings, and safe conduct. Thus, enclaves combine statistical, technical, legal, and operational controls with researcher training at different levels chosen by the agencies.

One particular approach that has been adopted by many agencies is that used by the NORC data enclave (<http://www.norc.org/dataenclave>). Researchers access confidential microdata from their offices. Statistical protections are applied to the data by constructing unique identifiers that substitute for explicit personal/organizational identifiers. Technical protections ensure that access is in compliance with agency-specific and department-specific data-sharing requirements. Legal protections require that researchers and their institutions sign access agreements. Operational controls limit researchers’ access to the data they need for their specific research questions, and audit logs, trails, and webcams can be used to monitor researcher activities. Finally, researchers are trained to reduce the likelihood of an inadvertent breach.

The *utility* from such an approach is that the number of researchers who can analyze the data is increased precisely because data access costs are low. In addition, the quality of analysis is increased: the NORC data enclave provides researchers with a virtual environment for metadata documentation, data augmentation, and knowledge sharing. The NORC enclave also has utilities that permit data archiving, indexing, and curation. The *risk* is limited because the enclave access modality relies on multiple approaches to reducing risk rather than one single “silver bullet.”

RECOMMENDATIONS

The expansion of coverage under health care reform heightens the need for new approaches to improve the utility derived from existing data, while at the same time protecting confidentiality. The evidence produced in this paper provides the basis for the following recommendations.

In the short term, delays—particularly those that reduce data utility but do not reduce risk associated with access—should be reduced. Often research has been funded and review of usefulness is redundant; these reviews serve to

prolong the approval process and discourage use of data, but they do not lead to enhanced protection. This is particularly true for the information-based research which is the focus of this paper, rather than interventional clinical research—the former uses existing data, records, or specimens, with no direct patient treatment (Duncan et al. 2004).

Moving forward, the use of remote access data enclaves should be promoted in order to facilitate the productive, high-quality usage of microdata and to support the most useful elements of traditional, hands-on data analysis in a collaborative environment. The goal of the enclaves, drawing on the experience in the physical and life sciences, should be to develop a research community and a knowledge infrastructure around both research questions and the different types of data necessary to answer policy questions.

Finally, a broad body of knowledge should be built about the availability of existing technologies for data access. Standards should be promulgated that facilitate use of best practices in protection of PHI, including standards for data security, so that each data provider does not have to “reinvent the wheel.”

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Funding for preparation of this paper was provided by AcademyHealth as part of its Summit on the Future of HSR Data and Methods. We gratefully acknowledge their support. All opinions expressed in the paper are those of the authors and do not necessarily reflect the views of either the National Science Foundation or Social & Scientific Systems. Dr. Lane was the Principal Investigator for establishment of the NORC Data Enclave.

Disclosures: None.

Disclaimers: None.

NOTE

1. Under the Privacy Rule, organizations that hold health care data such as health plans or providers (referred to as covered entities) are bound by specific rules with respect to the sharing or use of “protected health information” (PHI). While researchers are not considered covered entities and so are not directly bound by the rule, because much of the data traditionally used for health services research must be obtained by these covered entities researchers ultimately must adhere to its requirements.

REFERENCES

- Abowd, J., and J. Lane. 2004. "New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and V. Torra, pp. 282–289. Berlin: Springer-Verlag.
- American Statistical Association Open Letter on Proposed Changes to CMS Part D Public Use Files, September 17, 2008 [accessed on March 15, 2009]. Available at <http://www.amstat.org/outreach/pdfs/CMSPartDPUF.pdf>
- Balk, D. 2003. "Confidentiality Issues Arising from Integrating Social and Health Behavioral Data with Geospatial Data." In *NSF Confidentiality Workshop*, edited by J. Lane. New York: Trustees of Columbia University.
- Berk, M., C. Schur, and P. Mohr. 1990. "Using Survey Data to Estimate Prescription Drug Costs." *Health Affairs* 9 (3): 146–56.
- Corey, C., and H. Freeman. 1990. *Use of Telephone Interviewing in Health Care Research*. Health Services Research. University of California, Los Angeles.
- Doyle, P., J. Lane, J. Theeuwes, and L. Zayatz., eds. 2001. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North Holland.
- Duncan, G., S. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. 2001. "Disclosure Limitation Methods and Information Loss for Tabular Data." In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle et al., pp. 135–166. Amsterdam: North Holland.
- Duncan, G., S. KellerMcNulty, and L. Stokes. 2004. *Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map*. Los Alamos National Laboratory, NM: National Institute for Statistical Sciences.
- Gunn, P. P., A. Fremont, M. Bottrell, L. Shugarman, J. Galegher, and T. Bikson. 2004. "The Health Insurance Portability and Accountability Act Privacy Rule: A Practical Guide for Researchers." *Medical Care* 42: 321–327.
- Guttman, M. P., and P. Stern, eds. 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington, DC: National Academies Press.
- Lane, J. 2007. "Optimizing the Use of Microdata: An Overview of the Issues." *Journal of Official Statistics* 23 (3): 299–317.
- . 2010. "Administrative and Survey Data." In *Handbook of Survey Research*, edited by P. Marsden and J. Welsh, sec. 21. Oxford, England: Oxford University Press.
- Machlin, S., J. Cohen, A. Elixhauser, K. Beauregard, and C. Steiner. 2009. "Sensitivity of Household Reported Medical Conditions in the Medical Expenditure Panel Survey." *Medical Care* 47 (6): 618–625.
- Ness, R. B. 2007. "Influence of the HIPAA Privacy Rule on Health Research." *Journal of American Medical Association* 298 (18): 2164–2170.
- Waite, L. J., E. O. Laumann, A. Das, and L. P. Schumm. 2009. "Sexuality: Measures of Partnerships, Practices, Attitudes, and Problems in the National Social Life, Health and Aging Study." *Journals of Gerontology: Social Sciences* 64B (1S): I56–I66.

Winkler, W. 2005, *Overview of Record Linkage and Current Research Directions*. Washington, D.C.: U.S. Bureau of the Census, Statistical Research Division Report.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.