German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD

Working Paper Series

Working Paper No. 113

Balancing Access to Data And Privacy

A review of the issues and approaches for the future

Julia Lane and Claudia Schur

July 2009

Federal Ministry
of Education
and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Balancing Access to Data And Privacy

A review of the issues and approaches for the future[1]

Julia Lane, *National Science Foundation*

Claudia Schur, *Social and Scientific Systems*

Abstract

Access to sensitive micro data should be provided using remote access data enclaves. These enclaves should be built to facilitate the productive, high-quality usage of microdata. In other words, they should support a collaborative environment that facilitates the development and exchange of knowledge about data among data producers and consumers. The experience of the physical and life sciences has shown that it is possible to develop a research community and a knowledge infrastructure around both research questions and the different types of data necessary to answer policy questions. In sum, establishing a virtual organization approach would provided the research community with the ability to move away from individual, or artisan, science, towards the more generally accepted community based approach.

Enclave should include a number of features: metadata documentation capacity so that knowledge about data can be shared; capacity to add data so that the data infrastructure can be augmented; communication capacity, such as wikis, blogs and discussion groups so that knowledge about the data can be deepened and incentives for information sharing so that a community of practice can be built. The opportunity to transform micro-data based research through such a organizational infrastructure could potentially be as far-reaching as the changes that have taken place in the biological and astronomical sciences. It is, however, an open research question how such an organization should be established: whether the approach should be centralized or decentralized. Similarly, it is an open research question as to the appropriate metrics of success, and the best incentives to put in place to achieve success.

JEL Code: C81 - Methodology for Collecting, Estimating, and Organizing Microeconomic Data

1 This is drawn from a report commissioned by AcademyHealth for a conference entitled “Health Services Research in 2020: A Summit on the Future of HSR Data and Methods”.

Introduction

The new Administration promises to focus considerable attention on basing policy decision on empirical evidence. Social science researchers have an unprecedented opportunity to respond to this national imperative by collecting and analyzing new data on human and social behavior. Advances in cyberinfrastructure have created a virtual deluge of new types of data ranging from new data on human interactions through digital imaging, sensors, and analytical instrumentation to new ways of collecting biological and geospatial information from survey respondents and to combining data from different sources, such as surveys and administrative records. New computational capacity has emerged that facilitates the analysis of the data in terms of modeling and simulation with an unprecedented breadth and depth and scale [2, 3]. At the same time, new instrumentation provides unprecedented opportunity for researchers to advance scientific understanding through collaboration with colleagues around the globe[4].

Other disciplines have developed institutions to use the new data collection and analysis capacity provided by cyberinfrastructure advances to respond to similarly pressing needs with great success. Biotechnologists acquired the human genome sequence and used new technologies and analytical methods to identify variations in human DNA that underlie particular diseases; the development of institutional infrastructures, such as the National Center for Biotechnology Information (NCBI), to promote access and analysis has been critical to this response.[2] In response to the concerns with tsunamis, geoscientists advanced their modeling, mapping and assessment techniques by putting together a tsunami-related data archive[3]. Astronomers have developed national and international virtual data observatories of the sky[4] to better compare and combine data from different sources.

Despite the potential recognized and realized by other disciplines, the set of options available to access social science data has remained fundamentally unchanged for decades. It is clear that traditional responses to providing access are unlikely to be sufficient to address the national imperative. Current approaches admit too great a loss of data utility, and too great a risk to confidentiality, to provide the evidence base necessary to guide policy.

The major reason for the current lack of options is that the data that are best suited to guide decision making are collected about human beings. These micro-data, or data collected on an individual unit of analysis, such as a person, household, or firm, are critical to

2 ncbi.nlm.nih.gov/dbgap
3 http://nctr.pmel.noaa.gov/Dart/dart_home.html
4 NVO: http://www.us-vo.org/; IVOA: http://www.ivoa.net/

modelling individual behavior, and hence to studying the marginal effects of interventions. This is particularly true in social science research where there is great interest in investigating different impacts across racial or ethnic groups, or where much of the analytical interest results from studying a small group of individuals. It is obvious that the micro-data are extremely sensitive because the very information necessary to provide policy guidance has a privacy risk in that the information could be used to re-identify individuals.

However, there may well be a chance for new approaches due to a confluence of a number of events.

One is the increased movement to openness and transparency in government, illustrated by the data.gov and open.gov initiatives. The increased emphasis on evidence based policy[5] and accountability is permeating the way in which the federal government does business, and should extend to statistical agencies.

Another is due to advances in multiple scientific disciplines. In the field of information technology, important advances have been made in the technological aspects of cyber-security. In statistical analysis, there is a burgeoning literature on perturbation techniques and synthetic datasets. Finally, our understanding of the behavioral and social factors contributing to data protection has increased, particularly the ways in which social, economic, organizational and legal factors can be combined to reduce the risk of re-identification. [5]

This paper provides an overview of the challenges raised by concerns about data confidentiality in the context of social science research as well as the current environment and the significant issues raised by the advent of new electronic data systems and data linkage technologies. It describes the current methodologies used to ensure data security and privacy together with the impact on data analysis. It argues that the current data access modalities are insufficient to allow a response to the current national need for evidence-based research and provides an overview of successful approaches used in other contexts. It concludes by providing a set of policy recommendations for improving access to data for research purposes while giving appropriate attention to patient privacy.

A conceptual framework for Data Access and Privacy

The basic tension between data access and data confidentiality in the context of studying social science phenomena is well understood [6]. The core challenge is balancing the *risk* of

5 "Building Rigorous Evidence to Drive Policy" Orszag Blog June 8, 2009 http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy/

reidentification with the *utility* associated with data analysis.

The risk[6] from reidentifying individuals in a microdataset is intuitively obvious. Indeed, one way to formally measure the reidentification risk associated with a particular file is to measure the likelihood that a record can be matched to a master file [7]. If the data include direct identifiers, like names, social security numbers, establishment id numbers, the risk is obviously quite high. However, even access to close identifiers, such as physical addresses and IP addresses can be problematic. Indeed, HIPAA regulations under The Privacy Rule of 2003[7] require the removal of 18 different types of identifiers including other less obvious identifiers such as birth date, vehicle serial numbers, URLs, and voice prints. However, even seemingly innocuous information make it relatively straightforward to reidentify individuals, by finding a record with sufficient information that there is only one person in the relevant population with that set of characteristics[8]. In one particularly well known example, voter registration records, which provide information on birthdate, gender and zipcode, were combined with hospital discharge data to locate hospital discharge records to generate diagnosis, treatment, and medication information for former Massachusetts Governor William Weld [8]. It is worth noting that while birthdate, gender, and zipcode ares considered to be "de-identified data," and are permitted to be used under a data use agreement without patient authorization or waiver under HIPAA [9] 87 percent of Americans could be identified based simply on such information [10]. Such risk of re-identification has been increasing due to the increased public availability of identified data and rapid advances in the technology of linking files.[9]

There are two main types of consequences of reidentification that have been described in the literature - (i) financial and (ii) psychosocial. In the former category, one might think about the revelation of an expensive medical condition to an insurer, employer, or potential employer. Such disclosure might lead to denial of insurance coverage or to job loss or lack of job offer. These events could result in serious financial consequences. Not linked directly to health disclosures are the no less worrisome risks of identity theft. In terms of psychosocial impacts, revelation of PHI could lead to embarrassment or stigma in a social or work circle, or loss of reputation resulting in isolation or difficulty obtaining employment.

---

6 In the context of this analysis of health services research, we will combine the term "risk" of reidentification with the term "harm" from being reidentified. Although these are usually conceptually separated, the key concern associated with data access is the "risk of harm"

7 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.

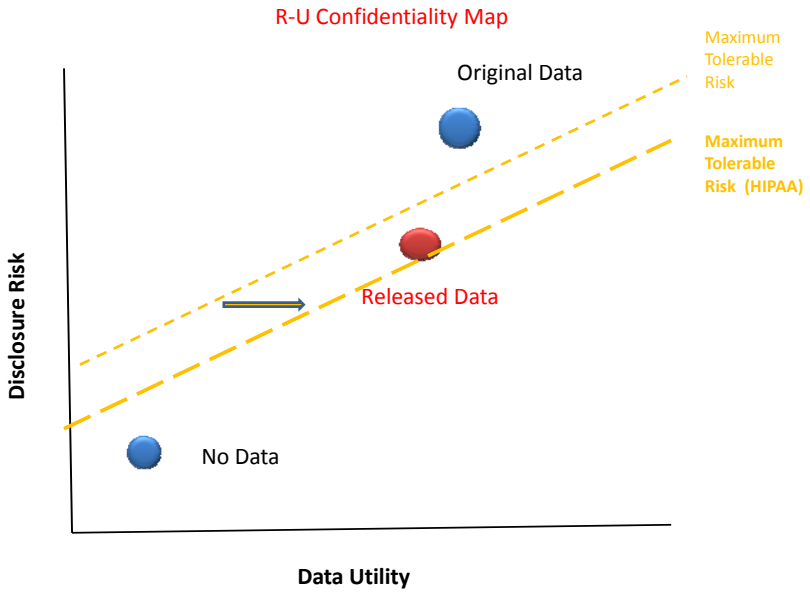
8 Statistical agencies often are even more stringent, and institute a "rule of three", since even if there were only two individuals in a multidimensional cell, the self-identifying respondent could infer information about someone else's characteristics.

9 http://www.fcsm.gov/working-papers/charlesday.pdf

Access to micro-data generates utility in a number of dimensions. [5] Clearly the more information that is provided and the more researchers that have access to the data, the greater the value of the analytical work that can be undertaken. In addition, the more transparent the access, the more likely it is that a body of knowledge will be developed around the dataset, expanding knowledge about the underlying data quality, the correct uses of the data, and the important data gaps. Finally, data access is essential to ensuring that analytical work is generalizable and replicable, which is the essence of scientific endeavor.

Figure 1 provides a graphical representation of this conceptual tradeoff between data risk and data utility. Here the dashed line identifies the maximum tolerable risk; the core guiding principle should be to generate released data that are as close to the frontier as possible.[10] [12]

Figure 1:



Overview of the current Environment

Types of Data, Utility and Risk

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

10 For a good practical implementation of this approach, see 11. Duncan, G., S. KellerMcNulty, and L. Stokes, Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map. 2004, National Institute for Statistical Sciences.

expanded.

There are four main data sources or types that are generally used; the data we discuss are primarily those related to the patient or consumer though we touch on provider-level data where appropriate.

Survey Data Perhaps the most longstanding source of data to guide policy has been survey data collected from individuals and households. The *utility* of survey data lies in its ability to provide detailed information on a wide variety of theoretically based questions: indeed, there are certain types of research questions that can only be addressed using survey data. However, survey data can be limited when examining specific. In-person surveys are also costly and telephone surveys, while less so, are increasingly subject to low response rates and, perhaps of more importance, can be biased due to differences in phone coverage by economic status [13].

The *risk* associated with survey data lies in its strength: the rich contextual information that is provided. Information that is typically important for policy decisions, such as geography, date of birth, marital status and history, number of children, and occupation are sufficient to reidentify not only the respondent, but also possibly others in the household. The risk is ameliorated by the fact that typically a survey is drawn from a subsample of the population, and the smaller the proportion of the population that is sampled, the more difficult it is to reidentify the individual. It is for this reason that surveys can often be released as public use files.

The *maximum acceptable risk* for federally collected survey data has, until quite recently, been determined by the legal mandate of the agency that collected the survey, as well as the Privacy Act of 1974 (5 U.S.C. 552a). In the case of the U.S. Census Bureau, the legal requirement was derived from its Title 13 mandate (if the data were collected using a Census Bureau frame) or from Title 15 (if the data were collected using a frame provided by the survey sponsor). BY contrast, in the case of the Bureau of Labor Statistics, the data collection was covered by a Commissioner's order. In 2002, however, the passage of the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), meant that statistical agencies collecting data under CIPSEA guidelines were formally required to take "reasonable means" to protect data confidentiality. However, OMB guidelines left that definition to agency discretion, and each agency has interpreted the term "reasonable means" differently.

Administrative data. This category of data usually refers to data that is collected or compiled primarily to administer a program or provide a benefit. The *utility* of administrative data is that they offer the advantage of large numbers of observations, so that it is possible to get very precise estimates of certain effects, and to study small groups with statistical

precision. However, administrative data are not a panacea. Since they are collected for programmatic, rather than analytical purposes, there can be important gaps. In addition, they tend to have limited information on individual demographic or socio-economic characteristics: for example, data on race/ethnicity can often be missing, in part because it is not part of the data custodian's mission-related responsibility. Because of these data limitations, studies assessing racial and ethnic disparities in areas of policy interest may be misleading. In addition, it is worth noting that administrative data are probably underutilized, reducing the utility of such data. Getting permission to use administrative data for purposes other than those for which it was collected can be an extremely time-consuming process. Each data custodian is responsible for fulfilling their agency's mission, which is typically programmatic in nature. Allowing researcher access is usually not within their mandate, however important the broader social or research goal.

The nature of disclosure *risk* with administrative data also lies in its strength, and as such is very different from that of survey data. Because the data are universal, a record that links uniquely is reidentified with certainty. In addition, because the program agency retains the administrative file, they always have the possibility of reidentifying the individual for non analytic purposes. For this reason, administrative records are typically not released as public use files, but are provided through licensing agreements or via onsite access.

The "maximum acceptable risk" definition here also depends on the guidelines of the agency. Probably the most well known (and feared) are the rules governing access to IRS administrative records, which are governed by Title 26 of the U.S. code. IRS imposes both physical and statistical security requirements, as well as institutes a formal safeguards review which can be daunting in its level of detail. By contrast, many state agencies have limited oversight of the use of their administrative records.

Linked Administrative and Survey Data Often linking administrative data to survey data can provide the best of both worlds (though it mitigates any advantage of the large size of claims data sets). In practice, survey data are often expanded by linking the data to administrative records, such as provider billing records, medical records, claims data, or employer information.[14]

The *utility* of linked survey and administrative data is substantial. In addition to increasing the accuracy of reporting, survey data can also be linked to administrative records to expand the analytic time horizon; for example, a cross-sectional survey with a follow-up linkage to administrative data can provide a quasi-longitudinal data set. From a methodological view, linkage to administrative data can also help to reduce bias from survey nonresponse (Cohen,

2008). However, the increased *risk* associated with combining the rich contextual survey information with administrative records is substantial. Typically the only access that is provided is onsite.

The *maximum acceptable risk* can be a major challenge to define, since typically multiple legal requirements cover the use of such linked data. It is often the case that it is the intersection, rather than the union of the different requirements that govern the definition.

Social-Spatial data. With the increasing sophistication of technology and geographic information systems, the use of social-spatial data is expanding. These are usually contextual data describing neighborhoods or other small geographic areas. An entire literature has developed in spatially explicit analysis because location, pattern, and spatial structure all matter in understanding human behavior.

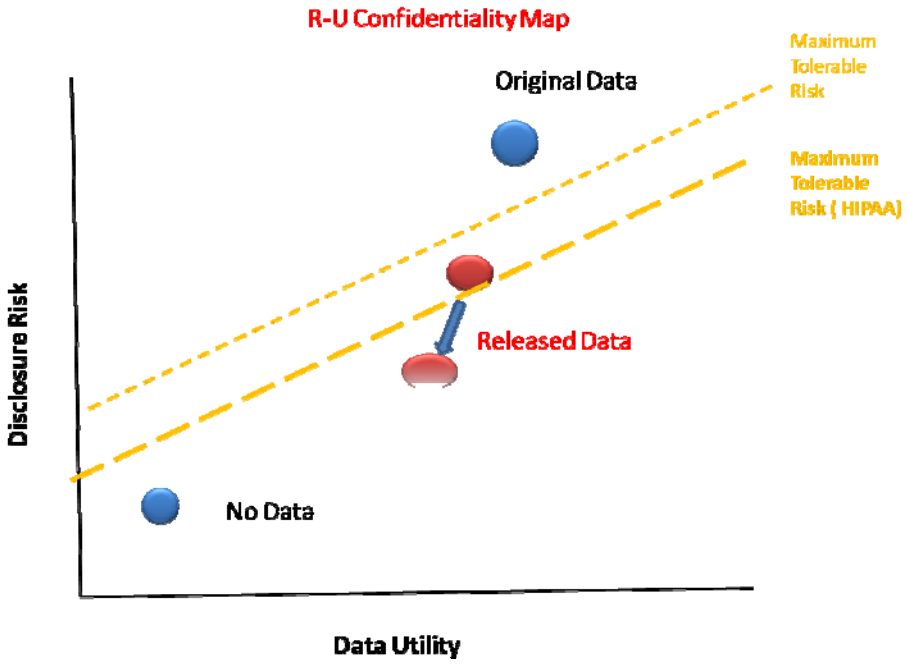
The *utility* of such data is that many insights can be derived from the contextual variables surrounding an individual - the schools they go to, the neighborhoods they live in, the firms they work for, etc, and even the people they interact with. Yet there is substantial *risk* from the use of geo-codes (such as latitude-longitude coordinates) rather than addresses or political units. Clearly, publicly-available data based on real property records - such as lot size or property tax maps - can lead to re-identification of individuals. However, just as new administrative datasets have made it more feasible to link micro-data, so have technological advances such as global positioning system (GPS) instruments and satellite technology made it much easier to link location-specific data at the household or neighborhood level and re-identify individual respondents.[15] Indeed, the capacity to study the inter-relationships among social, demographic, neighborhood, environmental, health supply and other contextual factors may be essential in order to advance our understanding but raises to an even greater level the red flags of confidentiality breaches. A recent study funded by three federal agencies concluded that the use of social-spatial data linked to other person-level data "has created significant uncertainties about the ability to protect the confidentiality promised to research participants" (National Research Council, 2007).

Current Approaches: Access to Social Science data

The following is a brief enumeration of the most common strategies currently being used to balance access and privacy [5]. In terms of the R-U map, the current approaches reduce utility and presumably reduce risk, although despite substantial concern about disclosure of personal

health information, it is worth noting that there are few studies or systematic discussions of the types and consequences of breaches within research.[11]

Figure 2:



Public use files. A large number of federally-sponsored survey data sets are now available in public use files to be downloaded from agency websites. These files, however, contain somewhat restricted versions of the survey data. Geographic information is probably the most frequent type of data to be omitted from public use files - public use files for CPS, SIPP and PUMS do not include county identifiers. The increasing amount of data on the web, combined with better matching software has meant that the release of public use files is increasingly less viable.

More generally, the utility of this type of access is certainly questionable, given the amount of "sanitizing' that occurs before release. There are two types of approaches that are used to reduce the risk of disclosure: reducing information and perturbing information. In the former case information is reduced in a public use file by deleting variables (such as geographic information), recoding categorical variables into larger categories, recoding continuous variables into categories, rounding continuous variables, using top and bottom code and using local suppression and enlarging geographic areas. An excellent survey of the

11 There is some evidence from research conducted earlier in the HIV epidemic that rural patients traveled to urban areas to seek care in order to avoid disclosure about their condition; however, this phenomenon was related to health care delivery rather than research (Schur et al,). Similarly, a typology of confidentiality was developed but applies to health care communications between patients and providers rather than to research (Brann and Matson, 2004).

techniques available to agencies is provided by Duncan et al.[12] who note that considerable effort has gone into developing disclosure limitation methods for tabular data that effectively lower disclosure risk and provide products with high utility to legitimate data users. However, as has been documented in multiple reports, these approaches can lead to biased coefficients (in the case of topcoding) and reduced statistical precision (recoding). In the case of at least one important survey, the Survey of Income and Program Participation, the lack of date of birth information substantially reduced its value for studying retirement decisions (one of the two major rationales for funding the study) and the lack of state specific detail substantially reduced its value for studying welfare program participation (the other major rationale for funding the study).

In the latter case, information is perturbed in a number of ways: noise addition (adding a random error centered on zero to the measure), record swapping, rank swapping, blanking and imputation, micro-aggregation and multiple imputation/modeling to generate synthetic data. In the case of noise addition, the resultant parameter estimates are unbiased, but the standard errors are too large. Thus, for example, it is more difficult to reject the null hypothesis that a treatment has no impact – even if, in fact, it has an impact.

Of even greater concern is the fact that most researchers are unaware that the public use files have been disclosure proofed, and make inferences without understanding the caveats. Of course, this is partly due to the fact that despite the fact that statistical agencies publish extensive and high-quality documentation that informs users of the consequences of different sampling procedures and nonsampling errors, and how to adjust estimates accordingly, there is typically no discussion of the effort to achieve disclosure limitation because of concerns that such information would permit researchers to "back out" the disclosure limitation algorithms.

It is also not clear what the impact of such statistical approaches have on risk. The increased capacity to find identifying information and link to the survey data means that researchers like Latanya Sweeney have been able to reidentify individuals in public use files.

Research data centers. Research data centers - both on-site and remote access - provide access to data in a controlled physical or electronic environment. The nature of the control is such that the researcher can essentially have access to 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. Materials are subject to review before they can be removed from a data center.

---

12 George T. Duncan, Stephen E. Fienberg, Ramayya Krishnan, Rema Padman and Stephen F. Roehrig "Disclosure Limitation Methods and Information Loss for Tabular Data" in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, editors, Pat Doyle, Julia Lane, Laura Zayatz and Jules Theeuwes, North Holland, 2001.

The impact on data *utility* is substantial in terms of reducing the amount of research that can be done, given that utility can be defined as a function of both data quality and the number of researchers using the data [5]. The process used by the Census Bureau RDCs is particularly cumbersome and requires that the research conducted show a benefit to the Bureau's programs. From submission of a final proposal to actual use of the data takes a minimum of 6 months. This long time horizon imposes substantial burdens on researchers, and lessens the usefulness of data for quick turnaround policy studies. Physically accessing data centers can be difficult for those not in the Washington metro area; even for those located nearby, temporarily re-locating can be an imposition and working in the data center makes it difficult to confer with colleagues or have a research assistant do the programming. In some cases, the researcher must pay significantly for use of the data center, usually to have data center staff construct analytic files that the researcher might have been able to create more efficiently on their own.

Licensing arrangements or data use agreements (DUAs). Licensing is used by a variety of different agencies. The approach involves the agency entering into a signed agreement with an external researcher that permits them to access semi anonymized datafiles using a defined set of protocols at their home institution. The license typically includes a Data Security Plan that defines location, security arrangements and access protocols; confidentiality pledges; institutional concurrence, disclosure review, onsite security inspections and terms for termination.

The impact on data utility is substantial, primarily due to the time and financial burdens on the researcher. The application processes for access to data range from the straightforward to the intrusive and cumbersome. In some cases, researchers need to justify the relevance of their research and submit lengthy answers to questions, despite having obtained funding support.

General Issues

In addition to the very specific issues identified about the current environment associated with data access, it is worth noting that the requirements of the Privacy Rule may inhibit other aspects of research. Gaining permission from individuals to gather these types of data is difficult, in part because of the publicity surrounding the implementation of privacy rules and the mis-interpretation of the requirements by providers. Researchers have reported that the requirements for informed consent and the explanation of risks has reduced individuals' willingness to participate in research and concerns over penalties for disclosure of information have made organizations reluctant to make data available to researchers[16]. These changes

may result in fewer research studies or research studies that are less scientifically robust.

Finally, the Privacy Rule may limit the circle of those involved in the research process. Without strong academic partners to facilitate the IRB process, community-based organizations may find it increasingly difficult to participate in research studies.

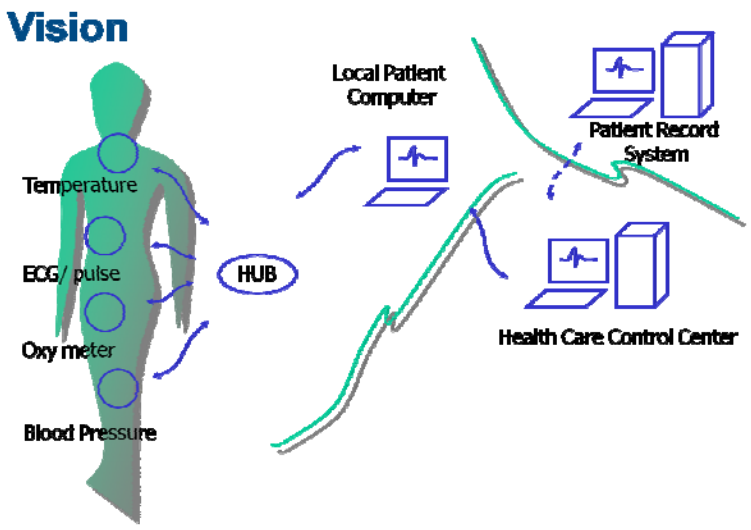
New Types of Data and New Approaches

The new demands for microdata access result from more than a national health care imperative and the new Administration's emphasis on openness and transparency[13]. Social scientists in many areas of research recognize that new ways of collecting data mean that the traditionally ways of providing access are inadequate. The new types of personally identifiable information that are being collected by means of sensors, video imaging, texts and bio-markers 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. Social scientists are also beginning to recognize that the advent of large scale shared datasets in the physical and biological sciences has transformed those disciplines by building scientific communities that share and communicate knowledge. Similar technologies offer a corresponding potential to transform social science research in general and health services research in particular.[2]

New Types of Data

The potential to use biomedical markers to guide social science research has become increasingly obvious. As the FOBIS project has noted, biomedical sensors can be developed that exploit micro and nanotechnology, to monitor body functions and status. These markers, together with development of RFID

Figure 1: Source - Dag Ausen Nordic Innovation[1]



13 Evidence by data.gov and open.gov

(radio frequency identification devices) and video technologies mean that information can be collected in a far more granular fashion than what is available from all the data sources previously available, ranging from the environmental impacts on social behavior to measuring the number and quality of human interactions. In fact similar technologies are already being used for research purposes to great advantage. For example, Schunn uses video data collected from a recent highly successful case of science and engineering, the Mars Exploration Rover, to study the way in which human interactions contributed to the success of the project. While the project both wildly exceeded engineering requirements for the mission and produced many important scientific discoveries, not all days of the mission were equally successful. Schunn uses the video records to trace the path from the structure of different subgroups (such as having formal roles and diversity of knowledge in the subgroups) to the occurrence of different social processes (such as task conflict, breadth of participation, communication norms, and shared mental models) to the occurrence of different cognitive processes (such as analogy, information search, and evaluation) and finally to outcomes (such as new methods for rover control and new hypotheses regarding the nature of Mars).[17] Similar potential should exist to examine how different health intervention teams interact and work together – and the impact (or failure) of different interventions.

Figure 2

Of course, human behavior is increasingly captured through transactions on the internet. For example, most businesses, as well as registering with the tax authority, also create a website. It is now entirely possible to use web-scraping technologies to capture up to date information on what businesses are doing, rather than relying on administrative records and survey information. Historical records on businesses can also be created by delving into the repository of webpages on the Wayback Machine (see Figure 4 for an example of the webpages for Citibank). This archive takes snapshots of the web every two months and stores them in the manner shown, providing a rich archive of hundreds of billions of web pages. Individual as well as business behavior can be studied using this archive. Indeed, major NSF

grants, such as the Cornell Cybertools award[14], have funded the study of social and information networks using these very large semi structured datasets.

This vividly illustrates how new approaches to capturing information could transform social scientists' ability to provide information to policy makers. Imagine a similar exercise being done in the study of health care markets, for example. Real time data collected from the web analysis of online blogs and newspaper articles could have picked up clusters of concern about different types of medicines or treatments and potentially used to describe the information cascades about swine flu that had such an impact in the spring of 2009.

Of course, together with new data, new analytical techniques need to be developed. Standard regression analysis and tabular presentations are often inadequate representations of the complexity of the underlying data generation function. There are a variety of reasons for this inadequacy. First, the units of analysis are often amorphous – social networks rather than individuals, health ecosystems rather than physical health care establishments. Second, the structural relationships are typically highly nonlinear, with multiple feedback loops. Third, theory has not developed sufficiently to describe the underlying structural relationships, so "making sense" of the vast amounts of data is a substantive challenge. There has been substantial effort invested in developing new models and tools to address the challenge, however. For example, since a major national priority is understanding the formation and evolution of terrorist networks through the internet and other communication channels, substantial resources have been devoted to the field of visual analytics. Their research agenda aligns very closely with a potential research agenda for social scientists, focusing as it does on the science of analytical reasoning, visual representations and interaction techniques, data representations and transformations, as well as the production, presentation and dissemination of complex relationships. [18] It is also worth noting that new partnerships are being formed to address the nontrivial computing challenges.[15]

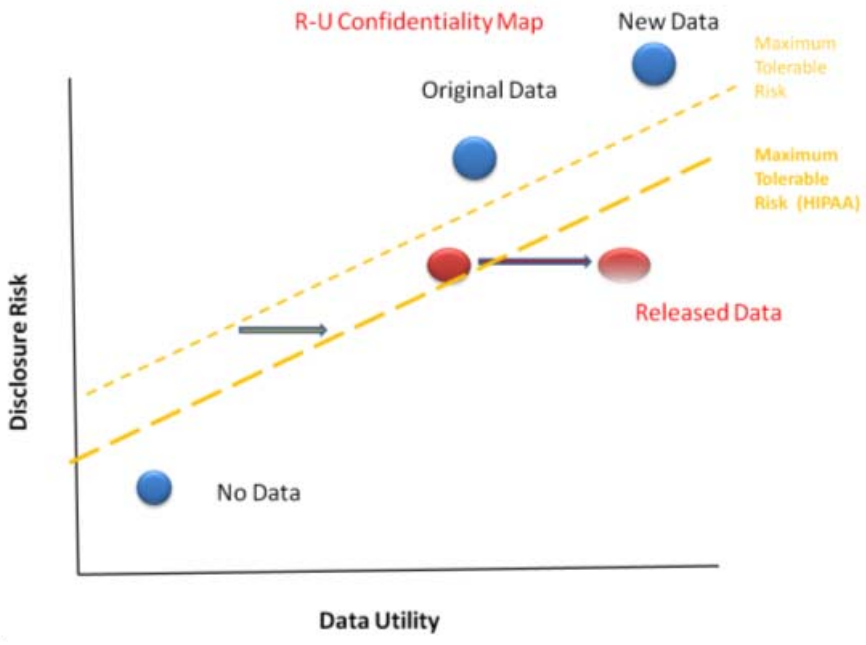
New Approaches: Remote Access and Statistical Approaches

Just as the new types of data could potentially transform the utility (and risk) associated with access to data on human beings, as indicated by the location of the "new data" element on the R-U map in Figure 5, new approaches to providing access have also evolved (as indicated by the "released data" element on the same map).

---

14 Very Large Semi-Structured Datasets for Social Science Research, NSF award 0537606 http://www.infosci.cornell.edu/SIN/cybertools
15 http://www.nsf.gov/news/news_summ.jsp?cntn_id=111470

Figure 5:



These include trustworthy computing: models, logics, algorithms, and theories for analyzing and reasoning about all aspects of trustworthiness - reliability, security, privacy, and usability. Protecting databases against intruders has a long history in computer science[19]. Computer scientists themselves are interested in protecting the confidentiality of the data on which they do research (for example, the Abilene Observatory supports the collection and dissemination of network data, such as IP addresses). Cyberinfrastructure advances have the potential to greatly expand the set of access modalities, particularly with respect to remote access. The Trustworthy Computing initiative at NSF has created a research community that focuses on developing network computers that are more predictable and less vulnerable to attack and abuse, that are developed, configured, operated, and evaluated by a well-trained workforce, and that educate the public in the secure and ethical operation of such computers.[O1] The Department of Defense has developed different levels of web-based access ranging from unclassified (nipr-net) to secret (sipr-net) to top-secret (jwics-net)[7] using off the shelf technology.

There are also scientific advances in ways to state, reason about, and resolve conflicts among privacy policies, and between privacy and security policies, particularly understanding the interplay between people and technology and the evaluation of trustworthiness. A good example of this is the PORTIA project which focuses on both the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners and

data users. Finally, the recent NSF SBE/CISE workshop on cyberinfrastructure[8] outlined a combined computer and social science research agenda for different approaches to access.

Remote Access

Indeed, many national and international statistical agencies have moved towards secure remote access as a way to promote researcher access. These entities, often called "data enclaves" have a portfolio approach to protecting confidentiality. This approach combines statistical, technical, legal and operational controls at different levels chosen by the agencies to optimize the combination of confidentiality protection and data utility in their context. A visualization of this is provided in Figure 6.

Figure 6:

The specific approach can be implemented within a secure data enclave that researchers can access remotely. All access is in compliance with agency-specific and department-specific data sharing requirements and utilize best practices from the data sharing field as well as state-of-the-art information technologies and applications. The specific approach is implemented within a secure data enclave that researchers can access remotely. In addition, the enclave typically has utilities that permit data archiving, indexing and curation.

A typical data enclave provides an information technology solution using a robust set of data access tools that facilitate high-quality researcher interaction with the data, while at the same time ensuring that data confidentiality is protected through a holistic suite of security and auditing measures. With the remote access mode, the data enclave provides external researchers with the ability to access the data in a controlled manner over the internet. Thus when a researcher needs to remotely access the data enclave's online resources, he/she first

initiates an encrypted connection with the data enclave using virtual private network (VPN) technology. VPN technology enables the data enclave to prevent an outsider from reading the data transmitted between the researcher's computer and the enclave's network. Before the VPN connection can be completed, the user must provide a pre-defined user id and password. RSA Smart Card technology can also be used, so that the user must validate his/her identity in real time. Other components of the VPN technology allow the enclave to control which network resources the external researcher can access on the enclave's network. Finally, if it becomes a requirement, the data enclave can also restrict the users to accessing the data enclave from specific, pre-defined IP addresses. So, for example, the researcher would be able to use the remote access tool at work, but not at home or from overseas.

There are typically also statistical protections. Typically data enclaves protect every data set by constructing a set of unique identifiers that can substitute for variables that are explicit personal/organizational identifiers, such as name, address, phone number, Social Security Number and Taxpayer Identification Number. The data enclave is also able to limit researchers' access to the data they need for their specific research questions if necessary. To accomplish this, the data enclave can create custom analytic data files that contain a subset of the columns (and even rows) contained in the master data set.

The *utility* from such an approach is that the new cybertools could be used to provide an opportunity for health services researchers to develop new modes of analysis, such as virtual organizations that study social science data.[16] The opportunity is clear from the way in which ubiquitous information technologies has transformed many facets of human interaction and organization. Tools such as the Grid, MySpace, and Second Life have changed how people congregate, collaborate, and communicate. Increasingly, people operate within groups that are distributed in space and in time that are augmented with computational agents such as simulations, databases, and analytic services which interact with human participants and are integral to the operation of the organization.

The risk is limited because the enclave access modality relies on multiple approaches to reducing risk rather than one single "silver bullet". There typically legal protections, which can be used to reduce the likelihood of a deliberate breach; researchers are trained and instilled with a culture of confidentiality, to reduce the likelihood of an inadvertent breach; and technical procedures are put in place, through IT technologies, to reduce the likelihood of

16 is a group of individuals whose members and resources may be dispersed geographically, but who function as a coherent unit through the use of cyberinfrastructure. A virtual organization is typically supported by, and provides shared and often real-time access to, centralized or distributed resources, such as community-specific tools, applications, data, and sensors, and experimental operations.

an external breach. Finally, organizational procedures are put in place, such as audit logs, trails and webcams to monitor behavior and act as a discipline device.

Statistical Approaches

Synthetic data

A great deal of attention has recently been paid to the potential of using synthetic data as an alternative approach to releasing public use data files.[20] One approach is to shuffle data; another is to develop samples composed of draws from the posterior predictive distribution of the confidential data, given some conventionally disclosure-controlled data. The advantages of these approaches is that they are inference valid in that the synthetic data contain exactly the same statistical information as the micro data. In addition, the effect of disclosure protection on data quality can be measured. Finally, the multiple synthetic data implicates are not identical so the analyst can use the between implicate variation to measure the extent to which confidentiality protection made the inferences less precise.

In practical terms, an important additional value of such inference-valid synthetic data is that multiple public use files can be created from the same underlying data - targeted at different audiences. For example, some users of business data (such as transportation agencies) are particularly interested in geographic detail, while others are interested in industry detail (such as industry analysts). Providing both levels of detail on the same data set immediately re-identifies important businesses. However, inference-valid synthetic data could be used to produce two separate data sets that can not be re-linked for such re-identification. An excellent layman's summary is provided by Norman Bradburn

> … synthetic data sets which have all of the statistical properties of the original data set, but have entirely false data - made-up data, so that you cannot break confidentiality because, in fact, any data set, any data record you have is a synthetic data record. …… possibly the way of the future for lots of very, very confidential data, and maybe because the … the ability to protect confidentiality … is being eroded by the internet …this is probably where we are going to be driven to, although, I hope not.[21]

The *risk* is reduced since the synthetic data record does not reflect the respondent's actual data record, so identity disclosure is impossible. However, it is quite possible that an individual's attribute could be disclosed, and with extreme values, the re-identification of a source record might occur.

There is also a strong possibility for reduced *utility* in some cases. In particular, since the synthetic data approach relies on the conditioning variables to generate the released data, any analysis on the synthetic data will be in error if the synthesizing model is wrong – there may also not be analytical validity for small subgroups. Synthetic data will not be able to release

the outliers that are often critical to understanding important rare events. And synthetic data take a very long time to generate, since there are very few people trained to create such files. Finally, since it is necessary to use quite sophisticated techniques to work with synthetic data (working with 10 or more implicate files), the typical user may not be able to use the dataset correctly.

Spatially based methods A new and parallel literature has developed using spatially based methods and algorithms.[22] These use Geographic Information Systems (GIS), rather than individuals or households as the unit of analysis but then can be used to link individuals with their geographic location to such measures as environmental exposures, the locations of health resources and the demographic characteristics of populations.[23] The same set of challenges arise with geospatial data as with other obviously reidentifiable measures. The use of geo-codes (such as latitude-longitude coordinates) rather than addresses, political units can create risks to respondents because publicly available data based on real property records - such as lot size, property tax maps – can lead to re-identification. However, just as new administrative datasets have made it more feasible to link micro-data, so have technological advances such as global positioning system (GPS) instruments and satellite technology made it much easier to link location-specific data at the household or neighborhood level and re-identify individual respondents[15].

It is quite striking that the approaches that have been used to protect confidentiality at the geo-spatial level mimic those that have been used to protect micro-data. In particular, researchers rely on geographical aggregation and removal of spatial context to protect confidentiality, but have similarly serious concerns about the impact of these measures on data quality[15]Other protection approaches, such as data masking are, mean that locations are “offset” by a parameter that moves the geo-coded location off the centerline to a “plausible” (approx) location on the correct side of the street or “squeezed” by a compression factor that moves locations inward on block face to ensure they are on correct street. Similarly, inverse address matching approaches to measure the degree of re-identification risk are very similar to record linking approaches.[24]

A core ethical question is raised in the use of the data: the expanding use of spatial technologies in combination with communication technologies via location based services (LBS), poses a particular challenge to increase beneficial uses and grow the industry, while protecting users. The core assumption of the LBS industry is that corporations and industry will own and control location and related information about individuals, individual choice limited to “opt-in” or “opt-out” of our services and boilerplate conditions [25] and leads to

very different technical challenges and research questions than those that will be addressed by the market place. In particular, there is a very strong case to be made for research into the public goods aspect of protecting privacy - particularly development of a legal/ethical code of conduct. An excellent review piece, providing a summary of institutional and technical approaches to ensuring confidentiality, the techniques employed by various agencies and a set of recommendations. [24]

Recommendations

The Administration's focus on evidence based policy means that new approaches must be taken to improve the utility derived from current and existing data, while at the same time protecting confidentiality. The evidence produced in this paper provides the basis for the following recommendations.

1. Access should be provided to data using remote access data enclaves. These enclaves should be built to facilitate the productive, high-quality usage of microdata, and should support the most useful elements of traditional, hands-on data analysis collaborative environment.[17] The goal of the enclaves, drawing on the experience on 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. In sum, establishing a virtual organization approach would provided the health services research community with the ability to move away from individual, or artisan, science, towards the more generally accepted community based approach adopted by the physical and biological sciences. It would provide the community with a chance to combine knowledge about data (through metadata documentation), augment the data infrastructure (through adding data), deepen knowledge (through wikis, blogs and discussion groups) and build a community of practice (through information sharing). This opportunity to transform health services research through such a organizational infrastructure could potentially be as far-reaching as the changes that have taken place in the biological and astronomical sciences. It is, however, an open research question for the health services data community as to how such an organization should be established: whether the approach should be centralized (like the UK's JISC) or decentralized (like the U.S.

17 See, for example, Building Effective Virtual Organizations http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

National Science Foundation approach). Similarly, it is an open research question as to the appropriate metrics of success, and the best incentives to put in place to achieve success. However a recent solicitation[18] as well as the highlighting of the importance of the topic in NSF's vision statement,[19] suggests that there is substantial opportunity for health service researchers to investigate the research issues.

2. Delays associated with access to data for research should be reduced. These delays act to reduce both data utility and do not reduce the risk associated with data access. 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 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. As noted by the Institute of Medicine Committee on Health Research, the current rules do not distinguish "between the unique needs of information-based research and interventional clinical research, which involves people who participate in experimental treatment. Applying the same protections in these two fundamentally different scenarios is neither appropriate nor justifiable."[26]
3. 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 personally identifiable information including standards for data security, so that each data provider does not have to "reinvent the wheel". Beyond this, we should rely on legal sanctions for anyone who intentionally tries to re-identify or disclose information.

18 www.nsf.gov/pubs/2008/nsf08550/nsf08550.htm
19 NSF Cyberinfrastructure Vision for 21st Century Discovery, March 2007

References:

1. Ausen, D., FOBIS: Foresight Biomedical Sensors, in FOBIS - NICE meeting. 2006: Nice.
2. Lazer, D., et al., Computational Social Science. Science, 2009: p. 721-723.
3. Lane, J., Administrative Transactions Data, in Rat fur Sozial und WissenschaftsDaten 2009: Berlin.
4. NationalScienceandTechnologyCouncil, Harnessing the Power of Digital Data for Science and Society. 2009: Washington DC.
5. Lane, J., Optimizing the Use of Microdata: An Overview of the Issues. Journal of Official Statistics, 2007. 23(3).
6. Doyle, P., et al., eds. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. 2001, North-Holland: Amsterdam.
7. Winkler, W., Overview of Record Linkage and Current Research Directions, in U.S. Bureau of the Census, Statistical Research Division Report. 2005.
8. Sweeney, L., Privacy Technologies for Homeland Security, in Testimony before the Privacy and Integrity Advisory Committee of the Department of Homeland Security. 2005: Washington DC.
9. Duncan, G.T., M. Elliot, and J.J. Salazar, Statistical Confidentiality: Principles and Practice. 2007.
10. Ferris, N., The Search for John Doe, in Government HealthIT. 2009.
11. Duncan, G., S. KellerMcNulty, and L. Stokes, Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map. 2004, National Institute for Statistical Sciences.
12. Duncan, G., et al., Disclosure limitation methods and information loss for tabular data, in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, et al., Editors. 2001.
13. Corey, C. and H. Freeman, Use of Telephone Interviewing in Health Care Research. Health Services Research, 1990.
14. Lane, J., Administrative and Survey Data, in Handbook of Survey Research, P. Marsden and J. Welsh, Editors. 2009, Oxford University Press.
15. Balk, D., Confidentiality issues arising from integrating social and health behavioral data with geospatial data", in NSF Confidentiality Workshop, J. Lane, Editor. 2003: Washington DC.
16. Shen, J., et al., Barriers of HIPAA Regulation to Implementation of Health Services Research. Journal of Medical Systems, 2006.
17. Schunn, C., Integrating Social and Cognitive Elements of Discovery and Innovation, N.S. Foundation, Editor. 2008.
18. Thomas, J. and K. Cook, Illuminating the Path: The Research and Development Agenda for Visual Analytics. 2005.
19. Dobkin, D., A. Jones, and R. Lipton, Secure Databases: Protection Against User Influence. ACM Transactions on Database Systems (TODS), 1979.
20. Abowd, J. and J. Lane, New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers, in Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra, Editors. 2004.
21. Duncan, G., Statistical Confidentiality: Is Synthetic Data the Answer?, in UCLA IDRE. 2006: UCLA.
22. Cassa, C.A., et al., A Context-sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection. J Am Med Inform Assoc, 2006. 13(2): p. 160-165.
23. Rushton, G., PUBLIC HEALTH, GIS, AND SPATIAL ANALYTIC TOOLS. Annual Review of Public Health, 2003. 24(1): p. 43-56.
24. Guttman, M.P. and P. Stern, eds. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. 2007, National Academies Press: Washington, D.C.
25. Onsrud, H., Privacy in the Use of Spatial Technologies: Ethics as a Driver of Technological Research Priorities, in NSF Confidentiality Workshop, J. Lane, Editor. 2003: Washington DC.
26. Nass, S.J., L.A. Levit, and L.O. Gostin, eds. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research. Institute of Medicine. 2006, National Academies Press: Washington DC.