

THE ROLE OF LICENSING AND ENFORCEMENT MECHANISMS IN PROMOTING ACCESS AND PROTECTING CONFIDENTIALITY

MARILYN SEASTROM, NATIONAL CENTER FOR EDUCATION STATISTICS,
INSTITUTE OF EDUCATION SCIENCES, U.S. DEPARTMENT OF EDUCATION,
CANDICE WRIGHT, CARNEGIE MELLON UNIVERSITY, AND
JOHN MELNICKI, HARBOR LANE ASSOCIATES

INTRODUCTION

Many federal agencies in the United States and national statistical institutes abroad have legislative mandates that require them to collect specific types of information, including at times individually identifiable data. At the same time, most countries have data confidentiality laws that require the protection of individually identifiable data (e.g., in the United States the Privacy Act of 1974, as amended, and in the European Union Article 285 of the Treaty). Thus organizations holding confidential data tend to limit releases to aggregate data or models and in some cases public-use data files that have been anonymised to protect the confidentiality of individually identifiable data.

The production of an anonymised public-use data file generally involves removing all direct individual identifiers and some combination of data coarsening. The coarsening includes, for example, recoding variables into broader categories, and top and bottom coding to avoid the re-identification of cases on the extreme tails of a distribution; perturbation, including the addition of random noise, changing specific values, rounding, and data swapping; and suppression through the deletion of some variables. These alterations to the original data produce a tension between the need to protect the confidentiality of data and the needs of qualified external researchers to access the same micro-data records that are used to produce official government statistics.

This tension is formally recognized in the laws and regulations of the United States, the European Union and many other individual countries. In the European Union, laws were passed in both 2002 and 2003 to grant access to qualified external researchers to certain types of confidential data held by the Community statistical authority (Eurostat). In the United States, the congressionally mandated OMB Information Quality Guidelines (Section 515 of the Treasury and General Government Appropriations Act for fiscal year 2001 (Public Law 106-554)) require that federal agencies ensure and maximize the quality, objectivity, utility, and integrity of information disseminated by federal agencies.¹ In addition, the Guidelines require the transparency and reproducibility of influential information, where influential information is defined as information that is likely to have a clear and substantial impact on important public policies or private sector decisions. Thus influential data must be accompanied by supporting documentation that allows an external user to clearly understand the steps

¹ (OMB, Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies, February 22, 2002).

involved in producing the information and, to be able to reproduce the information. In the case of influential analytic results, the mathematical and statistical processes used to produce the report must be described in sufficient detail to allow an independent analyst to substantially reproduce the findings using the original data and identical methods. To this end, OMB has suggested that “even in those cases where the original data require confidentiality, a qualified party, operating under the same confidentiality protections as the original analysts, may be asked to use the same data, computer model or statistical methods to replicate the analytic results reported in the original study” (FR, Vol. 67. No. 36 2/22/02, pg. 8456). In December 2002, the Congress enacted legislation in Title V, Subtitle A, Confidential Information Protection (CIP 2002)) of the E-Government Act that includes provisions for sharing confidential statistical data with qualified external researchers.

Government organizations have a number of different mechanisms that allow qualified external researchers access to original data files that are not released to the public because of data confidentiality protections. For example, some organizations have established limited access data centers that approve qualified researchers to come on site at a center to use the data in a controlled and monitored environment. Some organizations also use a remote access approach in which the qualified external researcher is provided with a synthetic data file to use in developing computer programs which are in turn submitted to the agency where the computer programs are run, the output is protected from potential disclosures, and then returned to the external researcher. Other organizations have implemented data sharing or licensing agreements that allow external researchers to use protected original data files in a secure environment at their home institutions, subject to the terms and responsibilities specified in the agreement. This report will explore the governing legislation and the use of data sharing agreements in the United States and abroad.

GOVERNMENT-WIDE LAWS SUPPORTING DATA AGREEMENTS IN THE UNITED STATES

There are three laws that pertain to individually identifiable data at any federal agency: the Privacy Act of 1974, the Computer Security Act of 1987, and the E-Government Act of 2002. In addition, individual agencies use the Code on Crimes and Criminal Procedures, the federal regulation on the “Federal Policy for the Protection of Human Subjects,” and agency-specific laws.

The Privacy Act of 1974. This act protects the privacy of personal data maintained by the federal government. It imposes specific requirements on federal agencies to protect the confidentiality and integrity of personal data, and limits the uses of these data. Federal Information Processing Standard Publication 41, *Computer Security Guidelines for Implementing the Privacy Act of 1974*, offers guidance to ensure that government held individually identifiable information is adequately protected in accordance with federal statutes and regulations. Unlawful disclosure is a misdemeanor and is subject to a fine up to \$5,000.

The Computer Security Act of 1987. This law relates to sensitive information, defined as any unclassified information that could adversely affect the national interest, the conduct of federal programs, or individual privacy covered by the Privacy Act of 1974. This law requires each federal agency to protect all federal computer systems against loss, misuse, disclosure, or modification. Unlawful disclosure under this law is a misdemeanor and is subject to a fine up to \$5,000.

The E-Government Act of 2002. This law provides strategic direction for implementing electronic Government under relevant statutes, including the Privacy Act, and the Government Paperwork Elimination Act.

Title III—Information Security, Federal Information Security Management Act of 2002. This act includes provisions that are intended to ensure the effectiveness of information security controls over federal information resources, where information security refers to the protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction. Included specifically are information and information systems that contain individually identifiable or proprietary information, with requirements that the confidentiality of this type of information be maintained.

Title V—Confidential Information Protection and Statistical Efficiency Act of 2002. Subtitle A, Confidential Information Protection (CIP 2002) of this title provides safeguards to ensure the confidentiality of individually identifiable information acquired under a pledge of confidentiality for statistical purposes. In particular, individually identifiable information obtained for statistical purposes may not be disclosed in identifiable form and may only be used for statistical purposes. In addition, only individuals authorized under this title may have access to such information. Authorized individuals include officers, employees, or agents of the agency. Agency heads may designate external researchers as agents to perform exclusively statistical activities, provided the external researcher agrees in writing to comply with all provisions of law that affect the agency's information. Anyone who willfully discloses information covered under this law to any unauthorized person will be subject to the penalties associated with a class E felony—including imprisonment for up to 5 years, and/or fines up to \$250,000.

It is important to note that while individually identifiable data loaned under data sharing agreements in the United States have usually been de-identified, through the removal of direct identifiers, such as name, address, and social security number, the amount of additional coarsening varies by agency.

LAWS SUPPORTING THE USE OF DATA AGREEMENTS IN THE EUROPEAN UNION

Although individual countries may each have their own laws and regulations guiding the collection, use, and dissemination of confidential statistical information, member countries in the European Union also have a shared set of regulations governing these

matters. In particular, within the European Union, the principle of statistical confidentiality was included in Article 285 of the Treaty that established the European Community. Since the Treaty, the European Union has passed a number of regulations that guide the collection, processing, analysis and dissemination of confidential data. The first of these regulations was passed in 1990 as Council Regulation No 1588/90. That regulation included provisions on the transmission and safe handling of confidential data moving from member state national authorities to the Statistical Office of the European Community. Then in 1995, the Data Protection Directive (No 95/46) was passed by the European Council and the European Parliament to strengthen the protection of personal information in both computerized and manual files.

In February of 1997 the Council of the European Union passed Council Regulation (EC) No. 322/97 on Community Statistics. That regulation, identified as the “EU statistical law,” includes articles that specify the requirements for the use and handling of confidential data collected for statistical purposes. Specifically, per Article 15, if the data are obtained exclusively for the production of Community statistics, Community and national authorities may only use them for statistical purposes (without the consent of the respondents). Data may be shared across member countries providing the standards of protection in the receiving country are at least as stringent as those in the country that collected the data, and the Community authority may distribute data across countries with the consent of the providing country (Article 17). Finally, all employees at the Community and national levels with access to confidential data that are subject to Community statistical confidentiality legislation are required to protect the confidentiality of the data; and Community and national levels are required to take appropriate measures to ensure the physical and logical protection of confidential data and to ensure that no unlawful disclosure occurs when Community statistics are disseminated (Article 18).

Then, in May 2002 the European Commission issued Regulation (EC) No. 831/2002 on the implementation of Council Regulation (EC) No. 322/97 (described above). The 2002 regulation concerns access to confidential data for scientific purposes. Under this regulation, the Community is given the authority to grant access to de-identified confidential data to certain categories of qualified external researchers. In particular, subject to the approval of the relevant country, the Community may grant access on its premises to confidential data obtained from a specific set of surveys (e.g., European Community Household Panel and the Labour Force Survey); alternatively, subject to the approval of the providing country, the Community may release anonymised microdata files for the same set of specified surveys.

Most recently, in June of 2003, the European Parliament and the Council of the European Union issued Regulation (EC) No 1177/2003 concerning Community statistics on income and living conditions (EU-SILC). These statistics include both cross-sectional and longitudinal household and individual level data on income, poverty, social exclusion, and other living conditions. Article 12 of this Regulation speaks to the issue of access to confidential data from this study for scientific purposes. In particular, the Community authority (Eurostat) is given the authority to grant access on its premises to confidential

data and to release sets of anonymised micro-data under the conditions specified in regulation (EC) No. 831/2002 discussed above.

In the European Union access to unaltered microdata files is only allowed at the offices of the Community or a National statistical authority, and then only under strict control. All data that are shared through data use agreements that allow access at the researchers' home institutions are required to be anonymised. Article 2 of Council Regulation (EC) No 322/97 defines anonymised data as "...individual statistical records which have been modified in order to minimize, in accordance with current best practice, the risk of identification of the statistical units to which they relate." Thus, it appears that data available to external researchers for analysis at their home institutions may be more limited under European Union law than in the United States.

COMPONENTS OF A DATA AGREEMENT

There are a number of common elements to the variety of data licensing and data sharing agreements that are currently in use in a variety of countries—including an application, specified security procedures, the designation of authorized users, methods of enforcement, and methods of termination.

Application. The first step in executing a data agreement involves the external researcher filing an application that generally includes a description of the requested data file(s), the location of the proposed site for the analysis, the length of time the data are needed, and a justification of the need for access to the specified data. The justification identifies the members of the research team, describes the proposed research, explains how the proposed use is consistent with the original purpose for the data collection, and provides an explanation of why existing public-use data (if available) are not sufficient.

Security Procedures. Individually identifiable confidential data are sensitive and require high levels of protection to prevent unauthorized disclosure, use, or modification. Thus, before data are shared, the external researchers must have a security plan, usually approved by the agency that is sharing the data. The security plan should specify the exact location where the data and printouts will be stored and the exact location where the data will be used, including the physical security arrangements at both locations. For example, the data should be stored in a locked cabinet when not in use; the location where the data will be used should be a safe, controlled access environment; and access to both should be limited to authorized users when the data are present. The use should be limited to the specified location; no remote access should be allowed; and the data should not be moved to a separate location without prior approval of the loaning agency.

The security plan should also describe the computer security environment where the data will be used. The safest computer environment is based on the use of a stand-alone personal computer that has no active connections to an external network or computer. The users of the stand-alone computer should be password protected, with periodic changes of the passwords. The computer should include a notification or warning either on the machine or displayed during login that indicates that unauthorized access to

individually identifiable data is a violation of law and will result in prosecution. The computer should have an automatic shutdown feature set to three or five minutes, or the computer or room should be locked when the authorized user is away. Finally, the number of backup copies should be limited, the data should not be backed up on a routine basis, and at the end of each session the hard disk must be overwritten.

Authorized Users. All authorized users are usually listed in the agreement. In addition, there should be evidence that each authorized user is aware of the terms of the data agreement. Specifically, each authorized user should be aware of the specific use for the data, the security requirements, and the laws and related penalties associated with the use or misuse of the data. Each authorized user should attest in writing to compliance with these terms of the data sharing agreement, subject to penalties specified in law for misuse of the data.

Agency Control. The agreement also frequently includes terms that allow the loaning agency to maintain control over the protected data. Control may be exerted in several ways—through requirements for periodic reporting to the agency, submission of materials to the agency for a disclosure review prior to any release of analyses based on protected data, and/or periodic onsite security inspections at the external researcher’s facility.

Periodic Reporting. Data agreements should be entered into for a fixed period of time, thus requiring the external users to either submit a request and justification for an extension or to return the protected data. In addition, the primary signatory to a data agreement should be required to notify the agency of any personnel changes among the authorized users, including the submission of statements from proposed new users—attesting to compliance with the terms of the data sharing agreement, subject to penalties specified in law for misuse of the data.

Publishing Results. Each agreement should also describe limitations on publishing results from the protected data. For example while analyzing the data, the authorized users should edit all printouts, tabulations, and reports for any possible disclosures of the data. The general rule is to not publish individual data cells that contain fewer than the agreed to number of cases (e.g., 3 or 5); similarly steps must be taken to ensure that such a cell could not be identified through subtraction either within or across tables. In addition, external researchers granted access to protected data should agree to agency review of reports prior to release or formal publication.

Security Inspections. Each external researcher seeking access to protected data may be asked to agree to unannounced onsite security inspections from an agency representative. The inspection allows the agency to ensure that all aspects of the agreement, including especially the security plan are in effect. The primary purpose of these inspections is to ensure compliance, and when necessary, to assist and advise the external researchers to achieve compliance.

During the course of an inspection, the security officer for the agency reviews the agreement file for a copy of the agreement, the security plan, and a list of authorized users. The security officer reviews the terms of the agreement and compares the current security procedures to those specified in the submitted security plan. Special emphasis is placed on the physical handling, storage, and transporting of the data, and on computer security requirements. The security officer also compares a current list of all project personnel with access to the data to the list of authorized users in the agreement to ensure that all project personnel are authorized users. The inspector also confirms that all current personnel have reviewed a copy of the agreement and the security procedures.

The security inspector may take corrective steps to help external researchers ensure compliance with the data agreement. The security inspector submits a report to the agency summarizing these actions, along with any more egregious outstanding problems. The agency, in turn, sends a formal letter to the agreement signatory summarizing the results of the inspection and, when necessary outlining corrective steps that must be taken.

Termination. Upon completion of the approved use of the protected data, the external researcher notifies the agency that the project is complete, and either returns the original protected data and additional data file documentation to the loaning agency or destroys the data following agreed upon procedures. At this time, the external researcher is also required to overwrite the protected data from any and all computers that were used in the analysis of the protected data.

The agreement should also include a provision that requires the external researcher to return the data when requested if a major violation of the agreement or breach of confidentiality is identified.

USE OF DATA AGREEMENTS IN THE UNITED STATES

An extensive search of government and federally funded university research centers identified 16 different current uses of data agreements that are used to provide qualified external researchers access to confidential or restricted access data files that have been de-identified, but contain data that could be used for indirect identification of individuals² (table 1). Although most of the data sets identified were sponsored by one or more federal agency, a number of the data agreements are administered by other entities, such as research centers. In some cases an agreement is used for only one specific data set, in others the same agreement format is used for multiple data collections within one agency. While the actual name used for the agreement varies across agencies (e.g., restricted-use data license, data agreement, data distribution agreement, restricted data use agreement), there are several stipulations that are common across all of these agreements.

² There are instances in which agencies may allow limited access to additional information required for matching to additional data sets. In other cases, matching to any external data sets is strictly prohibited. Please refer to individual agencies for details.

Universal Terms of the Agreements. Each application starts with a requirement for a research proposal that describes the data requested, the time frame for the request, the analysis proposed, and explains why public use data (if available) are not adequate. It must be clear that the proposed analysis is consistent with the statistical/research purpose for which the data are supplied.

For an agreement to be executed, each authorized data user must agree to release data only in statistical summaries so as to not disclose information about any individual, and to share the individually identifiable data only with members of the immediate authorized research team. Researchers are also prohibited from using the data to learn the identity of any person or other legally protected entity. Researchers may not transfer an agreement to another individual or move the data to another institution without the written consent of the loaning agency or center that is party to the agreement.

Institutional Support of Data Agreements. A number of other features of data agreements are shared across many, but not all, agreements (table 2). The first of these involves support from the researcher's institution. In particular, as part of the research proposal phase, eight of the fourteen require an Institutional Review Board (IRB) approval of the proposed research project from the researcher's home institution. In the remaining agreements, the staff members approving the agreements assume the responsibility for reviewing the proposals. In addition, all but three of the agreements require the approving signature of an individual who has the legal authority to bind the researcher's home institution to the terms of the agreement.

Confidentiality and Data Security. Each agreement requires the principal investigator on the agreement to sign a pledge of data confidentiality. All but three require additional signed data confidentiality pledges from all authorized users on each research team. Each agreement also requires a data security plan, either as prescribed in the agreement or as submitted by the applicant and approved by the agency or center, as a component of the data agreement. The data security plans generally cover both the computer environment where the data will be analyzed and the storage and handling of the data and related extracts and outputs. They are intended to prevent access by unauthorized persons. Related to this, 12 of the 16 agreements require the researcher to notify the agency or center if there are any known disclosures. Finally as an additional security precaution, eight of the agreements require the researchers to agree to on-site inspections of their data security arrangements.

Reporting Standards. Standards for reporting are not as well delineated. Each agreement requires the researcher to only release data as statistical summaries, so as to avoid the disclosure of information about an individual; however, only seven of the agreements include any detailed minimum cell size specifications. Seven of the agreements require the submission of manuscripts to the agency or center prior to formal journal submission or any other release of the data, and three additional agencies require notification of releases of any summary data from the data file. Thus, six of the agreements do not monitor compliance by reviewing reports produced using protected data.

Termination. At the completion of the approved research project, thirteen of the sixteen agreements require the researcher to either return the data or destroy the data under terms specified in the agreement.

Penalties. Fifteen of the agreements include provisions covering the termination or revocation of the agreement. The remaining agreement, used by AHRQ, carries monetary penalties up to \$10,000 and prison terms up to 5 years for violations but is silent on the disposition of the data in the event of a violation. The Privacy Act of 1974, as amended, protects data collected on behalf of federal agencies; under this law a violation of confidentiality constitutes a misdemeanor, which carries a fine up to \$5,000. With the 2002 passage of the Confidential Information Protection and Statistical Efficiency Act, a violation of the confidentiality of any individually identifiable data collected under a pledge of statistical confidentiality from a federal agency constitutes a class E felony, which includes imprisonment for up to 5 years, and/or fines up to \$250,000. There are additional agency specific laws and regulations that may apply as well. For example, agreements for the National Institute on Aging Retirement History study and for protected data from the Center for Medicare and Medicaid Studies or from the National Science Foundation require that the researcher is a current recipient of a federal grant or contract for the approved research, with the understanding that a disclosure may result in a recommendation for the revocation of research funding. Laws governing the violation of confidential data from the Center for Medicare and Medicaid Studies include fines up to \$5,000 and felony prison terms up to 10 years. The federal regulations for agencies within the Department of Justice include penalties up to \$10,000 for a violation of confidential data. The law used by the Bureau of Labor Statistics includes fines up to \$1,000 and prison terms up to 10 years. The law governing the use of protected data from the National Center for Education Statistics makes a violation of confidentiality a class E felony, with fines up to \$250,000 and prison terms up to 5 years.

USE OF DATA AGREEMENTS IN THE EUROPEAN UNION

Under the Treaty and subsequent laws of the European Union, all member countries are required to make their laws, regulations and procedures for the collection, processing, and dissemination of confidential statistical data consistent with the laws of the European Union. To that end, by August of 1997 all member countries of the European Union had enacted Data Protection Acts and Statistical Acts. Most Data Protection Acts permit the release of personal data from national statistical institutes for statistical and research purposes, subject to restricting provisions included in each country's Statistical Act (Holvast 1999).

The European Union laws and individual country laws on the uses and dissemination of confidential statistical data taken together give each member state in the European Union the ability to share anonymised microdata files with qualified external researchers for use at their home institutions, under the terms specified in data use agreements. In addition, these laws include provisions that allow each country to provide more restricted access to

microdata files that have not undergone full anonymization (e.g., de-identified, but not further perturbed).

Not all member countries post their data agreement application procedures and forms on their websites, and in some cases they may be on the website in the country's language, but not in English; as a result of these limitations, the analysis that follows is illustrative, but not exhaustive. A search of government websites identified Eurostat and seven countries that currently use data agreements to provide qualified external researchers access to anonymised microdata files (table 3)^{3,4}. While the actual name used for the agreement varies across countries (e.g., authorization, license, agreement, contract, research contract), there are several stipulations that are common across all of these agreements.

Universal Terms of the Agreements. Each application starts with a requirement for a research proposal that describes the data requested, the time frame for the request, and the analysis proposed. It must be clear that the proposed analysis is consistent with the statistical/research purpose for which the data are supplied.

For an agreement to be executed, each authorized data user must agree to release data only in statistical summaries so as to not disclose information about any individual, and to share the individually identifiable data only with members of the immediate authorized research team. Researchers are also prohibited from using the data to learn the identity of any person or other legally protected entity. Researchers may not transfer an agreement to another individual or move the data to another institution.

Institutional Support of Data Agreements. A number of other features of data agreements are shared across many, but not all, agreements (table 3). The first of these involves support from the researcher's institution. In particular, Eurostat and most of the identified member countries enter into the agreement with someone in a leadership position at the researcher's home institution. This ensures that the agreement is signed by an individual with the authority to legally bind the institution to the terms of the agreement. It also facilitates the ability to grant access to multiple researchers at an institution, subject to the approval of the statistical authority that holds the data.

Confidentiality and Data Security. Each agreement requires the principal investigator on the agreement to sign a pledge of data confidentiality. All but one of the identified agreements also require signed data confidentiality pledges from all authorized users on each research team, and require a data security plan, either as prescribed in the agreement or as submitted by the applicant and approved by the agency or center, as a component of the data agreement. The data security plans cover both the computer environment where the data will be analyzed and the storage and handling of the data and related extracts and outputs. They are intended to prevent access by unauthorized persons. Related to this,

³ Denmark is included because researchers can work at their home institutions, but access is restricted to an interactive form of remote access.

⁴ Norway and Iceland also provide anonymised microdata files under data use agreements, but the necessary information was not available in English on the web.

two agreements require the researcher to notify the agency or center if there are any known disclosures. Finally, as an additional security precaution, six of the seven identified member country agreements require the researchers to agree to on-site inspections of their data security arrangements.

Reporting Standards. Standards for reporting are not as well delineated. Each agreement requires the researcher to only release data as statistical summaries, so as to avoid the disclosure of information about an individual; however, only two of the agreements include any detailed minimum cell size specifications. Four of the seven agreements require the submission of manuscripts to the agency or center prior to formal journal submission or any other release of the data, and Eurostat and one additional country agreement require notification of releases of any summary data from the data file. Thus, Eurostat and five of the seven country agreements monitor compliance by reviewing reports produced using protected data.

Termination. In Denmark, although the researchers access the anonymised data at their home institutions over a secure government run network, the researchers never hold the microdata files. Thus, the agreements for Eurostat and five of the remaining six identified countries include provisions for returning or destroying the data under terms specified in the agreement, at the completion of the approved research project(s).

Penalties. Directive 95/46 of the European Parliament and the Council of the European Union states that sanctions must be imposed on anyone who fails to comply with the national measures associated with the Directive. In particular, under Article 28 the supervisory authority in each member country has the power to engage in legal proceedings when provisions of the Directive have been violated, the establishment of sanctions is left to each member country. The Data Protection Acts and Statistical Acts of each European Union member country include provisions specifying the penalties for a violation of the relevant Acts. In the case of Denmark, the agreement states that any researcher violating the provisions of the agreement will have their authorization to access microdata withdrawn, and will be denied access to government microdata files for at least three years. In Finland persons violating the confidentiality of statistical data are subject to penalties specified in the Finnish Penal Code (Section 1 or 2, Chapter 40 or Section 5, Chapter 5). No penalties are mentioned in the data agreement for Ireland. In the case of the Netherlands, there are no established legal penalties for misusing confidential statistical data, but future access may be denied. In Sweden a breach of confidentiality restrictions is punishable by detention or imprisonment. Persons entering into a data agreement to access anonymised microdata files from a national statistical institute in the United Kingdom are informed that a breach of the provisions of the data agreement will lead to termination of access to all information in the UK Data Archive, and may result in legal action.

USE OF DATA AGREEMENTS IN OTHER COUNTRIES

A search of documents from the United Nations Economic and Social Council provides evidence from a survey of 24 statistical offices in the transition economies that it is

legally possible for external researchers to access microdata for their own purposes in 15 countries, and that only six countries have legal provision that preclude them from providing access to microdata (ECE Secretariat 2003). Consistent with the European Union Commission Regulation No. 831/2002 in the pre-accession countries (which comprised most of the 24 countries), access is limited to specific institutions for purposes related to scientific research and typically a signed agreement specifying the exact conditions for use of the microdata is required.

As is the case among European Union member countries, not all countries that employ data use agreements post their data agreement application procedures and forms on their websites, and in some cases they may be on the website in the country's language, but not in English; as a result of these limitations, the analysis that follows is limited to a nonrandom sampling of countries and is offered as an illustration of the international use of data agreements. Thus six additional countries were identified that currently use data agreements to provide qualified external researchers access to anonymised microdata files (table 4). Although the names used for the agreements vary across countries (e.g., authorization, memorandum of understanding, microdata license, undertakings, contract), there are several stipulations that are common across all of these agreements.

Universal Terms of the Agreements. Each agreement includes requirements that each authorized data user agrees to use the data for research purposes, release data only in statistical summaries so as to not disclose information about any individual, and to share the individually identifiable data only with members of the immediate authorized research team. Researchers are also prohibited from using the data to learn the identity of any person or other legally protected entity, and from transferring an agreement to another individual or moving the data to another institution.

In four of the six countries the agreement includes a requirement for a research proposal that describes the data requested, the time frame for the request, and the analysis proposed. It must be clear that the proposed analysis is consistent with the statistical/research purpose for which the data are supplied.

Institutional Support of Data Agreements. A number of other features of data agreements are shared across many, but not all, agreements (table 3). The first of these involves support from the researcher's institution. In particular, four of the six identified countries enter into the agreement with someone in a leadership position at the researcher's home institution. This ensures that the agreement is signed by an individual with the authority to legally bind the institution to the terms of the agreement.

Confidentiality and Data Security. Each agreement requires the principal investigator on the agreement to sign a pledge of data confidentiality. Three of the six identified agreements also require signed data confidentiality pledges from all authorized users on each research team. Three of the six identified agreements also requires a data security plan, either as prescribed in the agreement or as submitted by the applicant and approved by the agency or center, as a component of the data agreement. The data security plans cover both the computer environment where the data will be analyzed and the storage and handling of the data and related extracts and outputs. They are intended to prevent access

by unauthorized persons. As an additional security precaution, the agreements from the same three countries require the researchers to agree to on-site inspections of their data security arrangements. None of these six agreements require the researcher to notify national statistical institute if there are any known disclosures.

Reporting Standards. Standards for reporting are not as well delineated. Each agreement requires the researcher to only release data as statistical summaries, so as to avoid the disclosure of information about an individual; however, only one of the agreements includes any detailed minimum cell size specifications. Three of the six agreements require the submission of manuscripts to the agency or center prior to formal journal submission or any other release of the data, and one additional agreement requires notification of releases of any summary data from the data file. Thus, four of the six agreements monitor compliance by reviewing reports produced using protected data.

Termination. The agreements each include provisions for returning or destroying the data under terms specified in the agreement, at the completion of the approved research project(s).

Penalties. The penalties for violating statistical confidentiality vary by country. In Australia, Subsection 19(2) of the Census and Statistics Act of 1905 includes a fine of up to \$5000 and/or imprisonment up to 2 years. The Canadian agreement states that the user is responsible for complying with the terms of the use agreement and that “Any infringement of Statistics Canada’s rights may result in legal action.” In the case of Hong Kong, external researchers are provided only with a 20 percent subsample of the survey data held by the government’s Census and Statistics Department. External researchers must submit requests and/or special programs to the Census and Statistics Department to have results run and reviewed for the full data set prior to publishing or otherwise releasing any results based on the microdata file. In India there are explicit requirements for the handling of confidential statistical data, but the agreement does not include an explicit reference to legal penalties that may be imposed for violating the data agreement. In Israel, the government treats microdata files as publications, thus any violation of the terms of the agreement granting external access to microdata files is considered a violation of the copyright law; however, there are no formal penalties associated with the copyright law. There are additional laws under Israel’s Protection of Privacy Law (5741,1981) and the Israeli Statistics Ordinance, but no specific penalties are cited. Finally, in the case of New Zealand, the agreement states that it shall be governed by the current law in New Zealand. In particular, there is a fine of \$1000.

COMPARISONS OF ENFORCEMENT ACROSS AGREEMENTS

In the European Union the release of microdata files to external users is limited to anonymised data files in the European Union. This represents a major departure from practice in the United States, where many federal agencies release anonymised files as public use files on a regular basis. Researchers in the European Union member countries may only access de-identified files that have not had the data otherwise perturbed under special arrangements in facilities controlled by a country’s statistical institute. This

approach is similar to the data center approach adopted in the United States by the Census Bureau and the National Center for Health Statistics. However, in the United States a number of agencies provide external users access to de-identified data that have not been further perturbed under data use agreements. These differences do not however, result in major differences in the basic components of the data agreements.

In the United States the use of confidential statistical data under a data use agreement requires the submission of a research proposal for the proposed use; a data security plan for the safe storage, handling, and analysis of confidential statistical data; a pledge from the researcher to protect the confidential statistical data; a commitment to only publish any resulting analysis as summary aggregates of data; and a commitment to not transfer the confidential statistical data to any unauthorized individuals. All but three of the reviewed data agreements require institutional concurrence and/or IRB approval from the researcher's home institution. In addition, at least three-quarters of the reviewed agreements required that all users of the shared data sign a pledge of confidentiality and/or required researchers to report any known disclosures; and at least one-half included an option for security inspections and/or monitored compliance by reviewing products from data covered by data use agreements.

The situation is similar in the data agreements used by Eurostat and the identified European Union member countries, with requirements for a research proposal, a confidentiality pledge from the researcher, to only publish resulting analysis as summary aggregates, and to not transfer the confidential statistical data to any unauthorized individuals. All but one of the identified agreements requires a security plan and six of the seven identified countries included an option for security inspections. The Eurostat agreement and five of the seven identified member countries' agreements require institutional concurrence from the researcher's institution and monitor compliance by reviewing products from data covered by data agreements. The Eurostat agreement and those of five of the identified member countries require signatures from all authorized data users. In addition, only two agreements require the researchers to report any known disclosures.

Again, in the remaining countries that were identified as using data agreements, the basic terms of data use agreements are similar. Although two of the six identified agreements do not require a specific research plan, they all limit the uses to research. They each require a confidentiality pledge from the researcher, and commitments to only publish resulting analysis as summary aggregates and to not transfer the confidential statistical data to any unauthorized individuals. Two-thirds of the agreements identified require institutional concurrence, two-thirds monitor compliance by reviewing products, and one-half have the option to conduct security inspections, and require data security plans and signatures from all authorized data users. None of these agreements require researchers to report known disclosures.

The successful use of data agreements to provide qualified external researchers access to microdata files is related to the strength of the license. There are a number of elements of an agreement that may influence users' compliance with the terms of a data agreement.

ELEMENTS OF ENFORCEMENT IN DATA AGREEMENTS

Conditions Influencing Compliance. The requirement that the researcher requesting the data sign the data agreement signifies the researcher's legal commitment to the terms of the contract. This requirement was universal across the data agreements. Similarly, the requirement for institutional concurrence ensures that senior officials at the researcher's institution are aware of the responsibilities of projects using confidential statistical data. This requirement was used in most, but not all, of the data agreements. Further, a requirement that each member of the research team who will access the data must attest to maintain the confidentiality of the data with a formal signature ensures that everyone authorized to use the data is aware of the restrictions and responsibilities associated with using confidential statistical data. While most data agreements include this provision, several of the agreements leave the responsibility of educating other users to the principal investigator. The review of publications by the loaning statistical unit, further enforce compliance with the terms of the agreements; this is done by approximately two-thirds of the loaning agencies. Reporting requirements that specify minimum cell sizes also serve to prevent possible disclosures; while this practice is more common in the United States than other countries, less than half of the agencies or centers in the United States, the European Union, and the other countries include cell size limitations in their data agreements.

Security Inspections. On-site security inspections are a strong element of enforcement. In the United States eight of the sixteen agreements include a provision for on-site inspections. In the European Union six of the seven identified member country data agreements require researchers to agree to on-site inspections of their data security arrangements. And in the six other countries that were identified, three have data agreements that include provisions for on-site inspections. While the possibility of an inspection carries some weight on its own, knowing that an inspection is likely to occur arguably offers an even stronger deterrent. There is limited information available, in general, as to how the individual data agreements are enforced; and, in particular, as to how the options to inspect are exercised.

In the United States, for example, there is considerable variability in how the inspections are conducted. One agency conducts a limited number of inspections each year by selecting a specific geographic location and sending in-house professional staff to inspect researcher sites at the selected location. Several other agencies and one center have contracted with data security experts who conduct inspections on an on-going basis to insure that each data agreement holder from the sponsoring agency will be inspected during the life of each data agreement.

Termination/Close Out. Each of the agreements reserves the right for the agency that distributes the data to demand return of the data in the event agreement is terminated as a result of a violation. In addition, all but four of the examined agreements have requirements for the data to either be returned or destroyed at the end of the agreement. If the data may be destroyed, there are usually protocols described for the destruction of the data and related materials, and in some cases, the researchers is required to submit

certification that the data were destroyed. The use of established procedures is another mechanism for the agencies that loan data to monitor the disposition of all data under their control.

MEASURES OF THE EFFECTIVENESS OF DATA AGREEMENTS

External Audits. In the United States, the Government Accounting Office (GAO) conducts audits from time to time on specific programs and practices of an individual agency, or at times, on a specific program and/or practice across agencies within the federal government. In 1999, a federal agency that uses data use agreements was audited for its compliance with the Privacy Act of 1974 and for its ability to protect the confidentiality of individually identifiable data. The GAO found that weaknesses in the implementation of the agency's policies could potentially compromise the confidentiality of the individually identifiable data (U.S. Government 1999).

One of the factors cited was a failure on the part of the agency to routinely monitor contractors and researchers who used individually identifiable data. During the course of the audit, agency officials reported that they did not have an established system for monitoring data users' compliance with the terms of the data agreement. Instead, their policy was reliant upon the data users monitoring their own compliance with the commitments specified in the agreement. The agency was also cited for the failure to track and monitor the return or destruction of data at the end of the approved research project. Both of these shortcomings have the potential for resulting in unauthorized access and use of these confidential data. The importance of these findings can be generalized to any agency or center that shares individually identifiable data with qualified external researchers without monitoring compliance or without monitoring the disposition of the data at the end of the approved research.

The Role of Security Inspections. Short of undergoing an external audit, security inspections provide a mechanism for monitoring researchers' compliance with the terms of data agreements. Limited information is available from individual countries and/or agencies concerning the extent and nature of known violations. In the United States, the National Center for Education has published the results of a 1998 analysis of their inspection reports. That analysis noted violations of varying degrees of severity in terms of risk of disclosure, and reported that no known disclosures existed (Seastrom 2001). In a paper submitted from Statistics Sweden to the Statistical Commission of the Economic and Social Council of the United Nations, Niva, Sundgrid, and Lyberg (2003) noted that "a violation of confidentiality regarding microdata use has in fact hardly ever occurred in the National Statistical Institute data based research projects."

In the absence of more detailed agency or country specific data, an examination of the results of a number of security inspections can provide insights into the issues that arise as multiple external researchers work to implement the terms of data agreements.

Drawing upon several years of security inspections, the violations that occur can be categorized into four categories.⁵

The first category includes a set of minor violations that are easily corrected as part of the security inspection. Two of the violations in this category involve a failure to use proper signage—1.) a door or wall notice indicating that the facility has restricted use data, and 2.) a warning on the computer (or screen) that indicates the presence of restricted use data and delineates the penalties for improper use of restricted data. Another problem in this category occurs when junior members join a research team, and are not given adequate training concerning the requirements for the protection of the restricted use data. Each of these problems may be resolved on the spot with the assistance of a security inspector.

The second category of violations includes problems that while not serious, may not be corrected by a one-time intervention from a security inspector. The items in this category involve violations of the security plan that can be pointed out by a security inspector, but require an ongoing responsibility on the part of the data user. The first violation in this category is the broadest and may serve as the basis for the violations that were cited above and for those that will be discussed in this category; specifically, this violation involves a lack of restricted data oversight and/or the inadequate use of established procedures. Specific problems emanating from this include:

- § Computers left running unattended without active, automatic security measures in place,
- § A failure to maintain a consistent set of logs when the data are taken from or returned to the Project Security Officer.
- § Restricted data and removable storage/media devices and peripherals not properly stored in locked cabinet (including the absence of a locked cabinet,
- § Designated researcher(s) office space lacks basic security measures, and
- § Designated office space not located.

The third group of violations includes activities that result in a more immediate risk of disclosure of individually identifiable data associated with a possibility that unauthorized users may access the confidential data. A number of the specific violations in this category involve computer configurations and data security concerns:

- § Any connection to unsecured, public networks,
- § LAN/WAN connections to unsecured institution or organization server,
- § Unauthorized use of restricted data on unregistered terminals, and
- § Improper restricted data distribution procedures, including unsecured network transfer and/or the use of unsecured media-storage devices, media, or peripherals.

There are also problems associated with the improper use of the data, including the following:

- § File sharing with unauthorized users (this frequently involves the addition of new research assistants),

⁵ One of the co-authors of this report is the principle in a company that provides security inspection services to several of the agencies and centers in the United States that use data use agreements, that experience serves as the basis for this discussion.

- § Use of data at an unauthorized location (e.g., the researcher’s private residence), and
- § Researcher relocates without proper notification.

The final violation in this category involves a failure to return data or submit a Certificate of Destruction at the end of the approved use of the data.

The fourth category includes the most serious violation, which is the identification of an individual using the restricted use data.

NEXT STEPS

Security Inspections. Given the potential importance of security inspections as a means of monitoring and enforcement for the terms of data use agreements, all agencies using data use agreements to provide external researchers access to microdata files should give serious consideration to using security inspections on a regular basis. While periodic inspections cannot identify all violations, the knowledge that an inspection may occur sends a strong signal to the researchers that the agency responsible for the data takes its responsibilities seriously and expects the same for authorized data users.

Inspections can also help inform the agency of potential problems or vulnerabilities in the existing agreements that can be corrected by modifications in policies and procedures. In addition, the use of security inspectors to help correct potential violations in real time adds additional protections for restricted use data.

Termination Procedures. To meet the legal requirements associated with individually identifiable data, entities loaning microdata files with such data to external researchers must have procedures in place for monitoring the final disposition of the data files at the completion of a research project. This can help ensure that the data are not used in the future for unauthorized purposes.

Tracking Database. To run an effective data use agreement program, the agency must have and maintain complete, accurate, and thorough records for each data agreement. This is essential for monitoring the authorized users, the approved uses of the data, and the security of the data. With the advent of electronic record systems, consideration should be given to the use of automated record keeping software. Properly configured such a system could be used to maintain contact with users— to notify users of specific data files update information on those data, by generating letters reminding them of their current list of authorized users, or by reminding them of impending expiration dates.

This paper is intended to promote the exchange of ideas among researchers and policy makers. The views expressed in it are part of ongoing research and analysis and do not necessarily reflect the position of the U.S. Department of Education.

REFERENCES

Consolidated Version of the Treaty Establishing the European Union (2002) Article 285, Official Journal of the European Communities 325, 12/24/2002, pgs. 35-284.

Computer Security Act of 1987 (1987) Public Law 100-255, 100th Congress, Washington, D.C.: U.S. Government Printing Office.

Community Statistics (1997) European Council Regulation (EC) No. 322/97, Official Journal of the European Communities 052, 22/02/1997, pgs. 0001-0007

Community Statistics on Income and Living Conditions (EU-SILC) (2003) European Council Regulation (EC) No. 1177/2003, Official Journal of the European Communities 165, 03/07/2003, pgs. 0001-0009

Computer Security Guidelines for Implementing the Privacy Act of 1974 (1974) Federal Information Processing Standard, Publication 41, Washington, D.C.: U.S. Government Printing Office.

Economic Commission for Europe Secretariat, *Data Confidentiality Results of the Ad-Hoc Survey Carried Out in the Transition Economies* (2003) Conference of European Statisticians, June 2003.

E-Government Act of 2002 (2002) Title III, Information Security, Federal Information Security Management Act Public Law 107-347, 107th Congress, Washington, D.C.: U.S. Government Printing Office.

E-Government Act of 2002 (2002) Title V, Confidential Information Protection and Statistical Efficiency Act Public Law 107-347, 107th Congress, Washington, D.C.: U.S. Government Printing Office.

E-Government Act of 2002 (2002) Title V, Subtitle A, Confidential Information Protection Public Law 107-347, 107th Congress, Washington, D.C.: U.S. Government Printing Office.

Federal Policy for the Protection of Human Subjects (Revised as of July 1, 2003) Title 34, Volume I, Code of Federal Regulations, Part 97, Washington D.C.: U.S. Government Printing Office.

Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies (2002) Office of Management and the Budget, Federal Register, Vol. 67. No. 36 2/22/02, pg.8456, Washington, D.C.: U.S. Government Printing Office.

Holvast, Jan (1999) *Statistical Confidentiality at the European Level*, Conference of European Statisticians, March 1999.

Implementing Council Regulation (EC) No. 322/97 on Community Statistics, Concerning Access to Confidential Data for Scientific Purposes (2002) European Council Regulation (EC) No. 831/2002, Official Journal of the European Communities 133, 18/05/2002, pgs. 0007-0009

Niva, Matti, Bo Sundgren and Ingrid Lyberg, (2003) *Statistical Confidentiality and Microdata Issue Paper*, June 2003.

Privacy Act of 1974 (1974) U. S. Code Section 522a, as amended, Washington, D.C.: U.S. Government Printing Office.

Protection of Individuals with Regard to the Processing of Personal Data and the Free Movement of Such Data (1995) European Union Council Directive (EC) No. 95/46, Official Journal of the European Communities 281, 23/11/1995, pgs. 0031-0050

Seastrom, Marilyn (2001) "Licensing," in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pgs. 279-296, eds. P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L.V. Zayatz, Elsevier Science: Amsterdam, Netherlands.

Transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities (1990) European Union Council Regulation (EC) No. 1588/90, Official Journal of the European Communities L151, 15/06/90, pgs. 0001-0004.

Treasury and General Government Appropriations Act for Fiscal Year 2001 (2000) Public Law 106-554, 106th Congress, Washington, D.C.: U.S. Government Printing Office.