CORPORATE PERSPECTIVE ON DATA STEWARDSHIP FOR STATISTICAL DATA

Wendy A. Visscher and Richard A. Kulka
RTI International, Research Triangle Park, North Carolina 27709

KEY WORDS: Confidentiality, privacy, data security, Institutional Review Board, IRB

Our paper focuses on data stewardship from a corporate perspective. We bring this perspective from Research Triangle Institute, a not-for-profit research company that conducts research studies for federal, State, and private clients.

Background

We find ourselves in an information-rich world. Our capabilities to collect and disseminate information are increasing exponentially. Computers and the Internet are becoming more and more sophisticated every day. For researchers, this astounding progress has competing implications. We can collect more information, faster, and more efficiently. This allows us to discover relationships between variables easier, and to develop policies based on our findings faster. On the other hand, the increased ease of data collection and dissemination has multiplied the opportunities for these data to be intercepted and confidentiality breached. In our research studies, we make promises to participants about how their personal information will be used and with whom it will be shared. It is our responsibility, then, to minimize the risks for breaches of confidentiality given modern information technology.

Violations of privacy and confidentiality have long been considered the major risks associated with social research. People participate in these types of research studies because they feel the information they provide will help advance a specific research field, or may benefit society in general. The informed consent form they sign tells them that their information will be kept confidential and they accept this promise. But in reality, complete confidentiality can never be guaranteed and this is more true now than ever. We need to carefully access the risk of a breach of confidentiality in every research study, as well as the potential consequences of such a disclosure.

If the probability of disclosure is increasing, what are the potential harms that study participants may experience? These will vary both by data content and by participant characteristics. They may include emotional distress, social stigmatization, financial problems associated with loss of job or insurance, and legal implications if illegal activities become known. Even though breach of confidentiality is recognized as the major risk for social research, such a breach may actually cause physical harm if, for example, an abusive spouse learns than his or her partner reported something the spouse did not want revealed. This could be of special concern for a child of an abusive parent, or perhaps for an elder with an adult child caretaker.

Finally, the issue of whether third parties can become "secondary respondents" in research studies has become an issue for Institutional Review Boards (IRBs) recently. These are persons about whom the study respondent gives information, such as family members. These people have not given their consent for this information to be given, but will be vulnerable to the same harms should it fall into the wrong hands. Examples of third party information include interview reports of mental illness in a respondent's family, or genetic information about family members gleaned from biospecimens that are collected from the primary study respondent. Thus, the risks associated with violations of confidentiality may be quite complicated and can involve more than just the study respondent him or herself.

Methods for Protecting Confidentiality

How can confidentiality be protected? There are a number of methods that could be employed. Of course, the highest level of protection would be offered if the data were collected without identifiers or were rendered completely anonymous later. Unfortunately, such data files may be of limited utility to researchers so other methods are more commonly used. These vary from a simple separation of the main study data from participant identifiers to very sophisticated statistical techniques to minimize disclosure risks. Other methods include encryption and other data security measures, audit trails, and data sharing agreements. Federal Certificates of Confidentiality can also be obtained to provide some legal protection to researchers against the release of study data under subpoena.

The new federal privacy law, the Health Insurance Portability and Accountability Act, or HIPAA, is intended to provide additional protection for patient medical information. This law, which will go into effect next year, will have a significant impact on whether and how these data can be released by health care providers to researchers.

RTI does a wide array of research, spanning health,

social, statistical, and laboratory science. Our studies vary by subject area, client, study population, and by the sensitivity of the data involved. The types of data we collect also vary tremendously—from extremely sensitive information about child abuse, potentially damaging information about drug and alcohol use, totally identifiable genetic data from biospecimens, to quite innocuous information such as meat handling practices in the home.

At RTI, we use a three-pronged approach to address confidentiality in the studies we conduct. Specifically, we expect three parties to share the responsibility for stewardship of research data—the client, the project director and his/her team, and the Institutional Review Board or IRB. The role of each of these parties in protecting research data is described below.

Role of the Client

In addition to funding the research, our clients most often design the study, define the target population, specify the types of data to be collected, and determine how the data are to be used. In the majority of the studies that RTI conducts, particularly those done under research contracts (rather than grants), the study data are ultimately delivered to the client who is then responsible for their long-term storage and any future use or dissemination. The client's role in research studies at other institutions may differ from this, especially if they engage in a different mix of contract- versus grant-funded research than does RTI.

Role of the Project Team

For many RTI studies, the project team can be a large and quite diverse group of people. All team members must be fully cognizant of the importance of maintaining confidentiality, as well as potential harms to respondents should confidentiality be breached. We accomplish this awareness in a number of ways. All staff who come in contact with research participants or with identifiable data must complete a human subjects tutorial. This is required not only for substantive research staff, but also for programmers, data clerks, and field and telephone interviewers. Team members also sign project-specific confidentiality agreements. These document their understanding of, and agreement to abide by, the project's confidentiality requirements. Also critical to increasing staff awareness is RTI's corporate culture, which is strongly committed to the conduct of ethical research. This support is demonstrated by an institute-wide ethics program and through a visible and interactive IRB. RTI operates a separate office for research protection, which further underscores its corporate support for protecting study respondents and their data.

Role of the IRB

Finally, the Institutional Review Board (IRB) shares responsibility with the client and the project team for protecting research data. At RTI, the IRB has always had an important role in helping researchers protect their study respondents and the data they provide. This role has evolved over the years and the IRB program now operates under a corporate-level office for research protection. We maintain three separate IRB committees to handle our large research volume. We review at least 200 new pilot or full studies each year, along with about 500 amendments and continuing reviews for ongoing research.

All RTI studies involving human subjects must be reviewed by the IRB as specified in the assurance we hold with the federal Office for Human Research Protections (OHRP). The IRB reviews each study for compliance with human subjects regulations—either the "Common Rule", or the corresponding FDA regulations for drug or device studies. These regulations are based on three ethical principles that were delineated in Belmont Report of 1978—beneficence, justice, and respect for persons. Maintaining confidentiality is an expression of the last principle—respect for persons and for the personal data they provide.

The IRB reviews the confidentiality procedures proposed for each study to determine if they are adequate and feasible, yet reasonable given the sensitivity of the data and potential harms to respondents if the data were inadvertently disclosed. The IRB also reviews the consent form to be sure it accurately informs the study participant about the measures that will be used to protect their data, as well as any data sharing that is planned.

The IRB must also consider if there are any possible exceptions to confidentiality in the study. These may include mandatory reporting of child abuse or neglect, imminent harm, or communicable disease reporting.

Protecting Confidentiality in RTI Studies

How we operationalize our three-pronged approach to maintaining confidentiality is described here for three studies that RTI conducts for federal agencies.

Our largest federally-funded study is the National Survey on Drug Use and Health which is funded by SAMHSA. This is a national household survey that asks respondents to report on their use of tobacco, alcohol, and illegal drugs, as well as mental health conditions and

some illegal activities. Confidentiality procedures in place for this study include password-protected access to field laptops, transfer of data from RTI's public to private network after receipt from the field, staff confidentiality agreements, and unsigned informed consent form. RTI is responsible for preparing the public use files for this study. Our statisticians are using new statistical techniques for avoiding deductive disclosure of respondent identities in the creation of such files.

We do a number of studies for the National Center for Education Statistics (NCES). These studies involve collecting academic and related information from faculty and students from various educational institutions. Although these data are less sensitive than those collected in the previous study, their disclosure could still be potentially damaging to respondents. In some of these studies, staff sign not only a confidentiality agreement but a notarized affidavit of nondisclosure. Data access for this study involves three levels: (1) a public use file that is created using standard methods to avoid disclosure risks, (2) a restricted use file that is only released to parties who sign a Licensing Agreement with significant penalties if violated, and (3) an encrypted table generator that runs off the NCES website.

An important issue arose for these educational surveys recently. The client, study team, and IRB had to re-assess the promises of confidentiality given to respondents in light of the Patriot Act of 2001. This act was passed by Congress after the September 11 attacks and provides for the release of certain information, such as data about flight school students, to the government that might be relevant to terrorist investigations. Although a possibility, to date we have not been asked to release any study data under this Act.

Perhaps the most sensitive study that RTI conducts is the National Survey of Child and Adolescent Well-being. This study is funded by the Administration on Children, Youth, and Families and focuses on child abuse and neglect. Interviews are conducted with both children and with their parents or caregivers. Questions are asked about child abuse, domestic violence, alcohol and drug use, and illegal activities, all extraordinarily sensitive and very damaging if disclosed. The confidentiality procedures for this study include audio-computer-assisted self interviewing (A-CASI) so that the interviewer does not hear the answers to the questions and a federal Certificate of Confidentiality. In addition, the parents are asked to sign separate permission forms (in addition to their main consent forms) to indicate whether their data—or their child's data—can be linked with data from other sources. Such data linkage would, of course, increase the chance that the family could be identified.

This study also has an elaborate data release plan in place. This plan specifies three tiers of data release, all of which require IRB review by the data user. The first tier file is essentially a "safe" file. This file contains no identifiers, strata, geographic variables, or any other potentially identifying information such as agency-level indicators. Extensive data disclosure analysis is done in the creation of this file. The second tier file contains agency-level data and some geocodes and requires that the user sign a Licensing Agreement. The third tier contains all data and all identifiers. This file is used to link to other data sources (if the respondent has agreed to this linkage) and is never released outside RTI. Even the client is not granted access to this highest, most identifiable level of data.

A final confidentiality issue that affects this study is the mandatory reporting of possible child abuse or neglect detected through the survey to local agencies. The procedures for reporting these cases, including the consent language that explains this confidentiality exception to the parent and child respondents, were developed jointly by the project team and the IRB.

Balancing Confidentiality with Data Needs

How can we balance the competing desires of avoiding data disclosure and producing useful research data? Of course, we can and do delete direct identifiers from the study data files. We also use standard techniques for recoding, suppressing, masking, or even substituting for other potentially identifying variables. We can also conduct various levels of disclosure analysis and have these plans reviewed by a disclosure review board. There is constant pressure between data users and data producers to do as little perturbation of the data as possible. However, even a single breach of confidentiality could damage the reputation of both parties, may call into question the legitimacy of the research, as well as potentially causing significant harm to the study respondent. RTI statisticians are developing new techniques for statistical data limitation based on survey sampling theory. We hope these methods will further increase the statistical precision of data, while still protecting confidentiality.

What are some issues and challenges before us that will affect our abilities to be good data stewards? First, we must affirm our commitment to do ethical research. We need to balance ethical principles and the probability and consequences of a breach of confidentiality with feasibility and potential societal benefits from the research findings. Second, researchers must design

studies carefully and consider how they will protect the data during every phase of the study. Next, we must keep up with changing information technology. Researchers and IRB members cannot be expected to have this level of technical expertise. Instead, persons who are knowledgeable in this area must be consulted to assess threats to data security and ways to reduce these threats. Some research institutions constitute specialized privacy boards, or disclosure review boards, that are separate from the IRB to judge these threats. At RTI, the IRB seeks out this expertise as needed for specific studies.

Traditionally, the IRB has been very involved with data collection and consent procedures on the front-end of the project, then more concerned with specific data disclosure procedures as the project progresses. It is clear that more attention needs to be given up-front about how the data will be protected throughout the entire life of the project to further minimize disclosure risks and associated harm to study participants.

We are entering a new era of research which poses significant challenges. The gains we will make in understanding social, environmental, and health problems through the use of statistical data promise to be exciting. To achieve these strides, however, we need to maintain the trust of the public in the importance of our research and in our promises to protect their information to the fullest extent possible. Thus, our continued stewardship of statistical data will be critical to the future success of our research.