Ethical Use of Administrative Data for Research Purposes

Paul G. Stiles, Ph.D., J.D.

Roger A. Boothroyd, Ph.D.

Department of Mental Health
Law and Policy
Louis de la Parte Florida
Mental Health Institute
College of Behavioral
and Community Sciences
University of South Florida

Contents

Introduction

The goal of the paper it to provide an overview of the ethical issues and considerations associated with the maintenance, integration, and use of administrative data for research purposes. The paper is intended as a guide for data custodians (a.k.a., data owners, data stewards) as well as for other individuals who may be granted permission to use these data for research purposes (i.e., data users/researchers).

The paper is divided into three primary sections. Section one provides a brief introduction regarding the growth of administrative data and the ethical concerns and challenges that have emerged associated with the increased collection and use of electronic information records (i.e., administrative data). This section also includes a brief historical summary regarding the core ethical research principles that have emerged over time along with definitions of concepts discussed in this paper. Section two primarily focuses on the ethical considerations of integrating and using administrative data for research purposes from the data owners' or custodians' perspectives. Section three parallels the previous section but examines these ethical considerations from the data users' perspectives. In organizing the paper this way, we have tried to include relevant information with minimal redundancy – we encourage all readers (e.g., data owners, researchers, academics) to read all three sections to gain a better understanding of the pertinent issues before engaging in sharing data (data owners) or "borrowing" data to use (researchers). Throughout the paper, recommendations or "practice standards" are discussed, that we believe both data owners and users should adopt, to ensure that the individuals whose personal data are contained within these administrative databases are afforded the highest degree of protection available, with respect to their privacy and confidentiality, and that the research conducted using these data is of a high degree of scientific integrity. Finally, this paper focuses primarily on the use of health records because of the rapid changes and acute privacy issues involved. This is not to imply that integration of other types of existing data and information does not have ethical implications. The principles discussed here are directly applicable to a broad array of administration data sources.

RESEARCH ETHICS – A BRIEF HISTORICAL PERSPECTIVE AND CORE PRINCIPLES

Although issues associated with the ethical conduct of research have been debated for centuries, the development, codification, and acceptance of core principles is a much more recent phenomenon.

One of the first attempts to codify principles regarding the ethical conduct of research occurred after the Nuremberg Military Tribunal's verdict in the case of the United States v. Karl Brandt et al. (1947). Several of the German physicians accused of conducting inhumane experimentation during the war argued that their experiments differed little from pre-war research conducted by American and German physicians and that no international laws or guidelines differentiating between legal and illegal experiments existed. Drs. Andrew Ivy and Leo Alexander, who were working with the U.S. Counsel, became concerned with this defense and decided to develop six principles outlining the conduct of ethical research. According to their guidelines, human experimentation is justified only if its results benefit society and it is conducted in accordance with basic principles that "satisfy moral, ethical, and legal concepts."_Their original six points were expanded to ten and were included in the trial's verdict. Known as the "Nuremberg Code" (1949), these principles became accepted throughout the world, despite the fact that the legal force of this document was never established.[1] Among the ten points, the code required that human experimentation include voluntary informed consent and the right of the individual to control his/her own body. This code also recognized that risk must be weighed against expected benefit, and that unnecessary pain and suffering must be avoided. Like many codes, Nuremberg did not detail specific procedures and/or processes to guide researchers regarding the operationalization and implementation of these principles.

In 1978, the National Commission for the Protection of Human Subjects of Medical and Behavioral Research issued the Belmont Report, which provided three general guiding principles governing human subjects research. The first principle is **respect for persons**, and stipulates that individuals should have autonomy with respect to their decision-making and that persons with diminished autonomy are entitled to additional protections. This principle is operationalized through actions such as careful adherence to the best practices in securing informed consent. The second principle, **beneficence**, requires that persons be actively protected from negative outcomes or harm and that positive outcomes or benefits be promoted. This principle is operationalized through actions such as timely, responsible, and objective risk-benefit analyses. The third principle is **justice**, and it stipulates that people should be treated equally and thus share the burden and benefits associated with research. This principle is operationalized through actions such as the implementation of procedures to equitably select subjects for inclusion in the research.

[1] Despite the general acceptance of Nuremberg Code, some researchers initially assumed that it only applied to the Nazi atrocities and thus was not applicable to their own work (e.g., U.S. researchers conducting the Tuskegee Syphilis studies); however acceptance is now virtually universal.

The specific regulations that embody these principles are promulgated in the U.S. Code of Federal Regulations (CFR). Almost twenty federal agencies have adopted this "common rule" for the ethical conduct of human subject research including the Department of Health and Human Services (45 CFR 46), the Food and Drug Administration (21 CFR 50 and 56), and the Veterans Administration (38 CFR 16). As of this writing, the Department of Health and Human Services is considering changes in standards of protections for human subjects under the Common Rule. The specific changes being considered were published in the July 25, 2011, *Federal Register*. Among the revisions to the current regulations under consideration are: (1) revising the existing risk-based framework to better align the level of review to the level of risk, (2) using a single IRB review for domestic sites of multi-site studies, (3) updating informed consent forms and processes, (4) establishing mandatory data security and information protection standards for studies involving identifiable data, and (5) extending federal protections to all research conducted at institutions receiving funding from the Common Rule agencies.

Although the principles set forth in the Nuremberg Code and the Belmont Report do not specifically address the ethical issues associated with the use of administrative data for research purposes, these principles nonetheless provide guidance for the conduct of research using administrative data. For example, when using administrative data for research purposes, the issue of harm to individuals is unlikely to be manifested in the form of physical harm as stressed in the Nuremberg trial, but rather in harm resulting from a breach of confidentiality which may lead to an individual being stigmatized or incurring economic harm. Similarly, the principle of autonomy raises the question of whether individuals need to consent in order for their data to be used for research purposes. Is a blanket consent for data use permissible? Can individuals opt out? How are individual rights balanced against societal benefit?

AVAILABILITY AND USE OF ADMINISTRATIVE DATA: THE ISSUES

In recent years, there has been a dramatic increase in the availability and use of individual-level, large administrative databases for research purposes, due in part to the growth in computerized clinical records in conjunction with their ease of analysis associated with new advances in technology and software packages (Drake & McHugo, 2003; Sørensen & Olsen, 1998). Administrative data have become readily available, inexpensive to acquire, computer readable, and often are amassed on a very large number of individuals (Iezzoni, 2004). In fact Mason (1986) argued nearly twenty-five years ago that in "…western societies more people are employed collecting, handling, and distributing information than in any other occupation."

(pg. 5). Iezzoni (2004) noted that due to rapidly evolving information technologies, the definition, content, and scope of administrative data would change dramatically over the next several years – and it has.

While historically these administrative data have been used primarily for program operations and monitoring purposes, there has been an increasing interest and trend to use administrative data for secondary purposes, including for research. Safran et al. (2007) defined secondary use of health-related data as the use of individuals' personal health information for purposes not related to the direct provision of health care services. Kass et al. (2003) noted that the use of medical records has become an important source of data for health services, epidemiologic, and clinical studies. In addition, the growth in the development of disease-specific registries has made them powerful tools for researchers **in** estimating the prevalence and incidence of disease, resource utilization and clinical outcomes (Rabeneck et al., 2001).

This trend of using individuals' personal information for research purposes with broader societal benefits has sparked considerable debate from both critics and proponents. The issues cluster into three broad categories: (1) issues associated with the individuals whose information resides in these databases and disease registries (e.g., rights, confidentiality, privacy, harm), (2) issues related to the data owner (i.e., access, copyright) and (3) issues associated with the scientific merit of the research conducted using administrative data (e.g., data accuracy, appropriateness). An important question is how the core research ethical principles outlined in the Belmont Report are operationalized when administrative data are used for research purposes. Greenberg (2002) highlighted the need to find an appropriate balance between individuals' rights to privacy and protection of their personal records on the one hand and, on the other, providing professionals access to these data for education, research, and public health surveillance. Lane and Shur (2009) noted the most common strategies currently used to provide this balance between access and privacy include the (1) creation of public use data files, (2) establishment of research data centers, and (3) use of licensing and data sharing agreements.

With respect to issues associated with the individuals whose information resides in these databases, Mason (1986) noted that the ethical issues associated with the growth in electronic data were many; however, he highlighted four that he considered most critical. Two of these issues, (1) privacy and (2) property, are issues related to the individuals whose data reside in these databases. Some of the questions that must be addressed are: What are individuals' rights regarding who can access these data and under what conditions? Do the individuals whose data comprise these administrative databases need to provide permission prior to others accessing the information? Who owns these data? Is the owner the entity that collects, stores, and maintains the administrative data? Is it the individuals whose data are maintained? How will individuals be protected from harm?

Kass et al. (2003) surveyed 603 individuals with serious genetic disorders and chronic medical conditions to determine if they would be willing to have their health records used for research purposes without their knowledge. Over 55 percent expressed their disagreement with this abstract use of their data. However, when the qualifications were added that the database would be created anonymously and that access to the data for research would be controlled, an overwhelming majority were supportive of such a registry, highlighting the importance for the development and acceptance of an agreed upon set of practice standards. Results of Robling, Hood, Houston, Fay, and Evans (2004) focus groups on the use of medical data for research purposes highlighted fears related to unauthorized access to their records and anxiety associated with current data collection practices. The issue becomes even more complex when one considers that some persons who would be fine with their data being used for some types of research (e.g. curing cancer or improving Medicare services) may not want their data used for other types of research (e.g., abortion or stem cell studies).

Chamberlayne et al. (1998) and Broemeling, Kerluke, and Black (2009) described the creation of a population-based provincial registry in British Columbia, CA.[2] Both articles highlighted the importance of protecting individual privacy while recognizing the value of data linkage and population-based registries. To this end, the authors have proposed recommendations for developing a comprehensive set of best practice standards (Black, McGrail, Fooks, & Masdlove, 2005). In essence, data owners need to practice **due diligence**, i.e., protections that a reasonable person would implement to avoid harm to self or others.

The use of administrative data for research purposes also has raised a number of considerations for data owners. Mason (1986) discussed how data owners have an important responsibility for controlling data access. Other issues for data owners were highlighted in “PHS Policy on Instruction in the Responsible Conduct of Research” (U.S. Department of Health and Human Services, 2000).[3] As noted by Pimple (2002), the first of the nine core training areas detailed in this policy is

[2] The British Columbia Linked Health Development Data Project is funded by the British Columbia Ministry of Health and links six types of health data indexed with a code unique to each health care recipient.

[3] It should be noted that although innovative, the policy has been suspended.

data acquisition, management, sharing, and ownership. More specifically, the report notes that within this core area, training should include; (1) accepted practices for acquiring and maintaining research data, (2) proper methods for record keeping and electronic data collection and storage in scientific research, including what constitutes data, (3) maintenance of data notebooks or electronic files, (4) data privacy and confidentiality, (5) data selection, retention, sharing, ownership, and analysis, and (6) data as legal documents and intellectual property, including copyright laws.

One concern is the extent to which organizations and entities owning and managing administrative databases have established policies that are known to employees. The results of Hilton's (2000) survey of 123 information systems' employees highlighted the importance of ethics as it relates to information. His findings indicated that 35 percent of the respondents reported their organization's ethical guidelines for access to information and computer use were "not well known or nonexistent" and another 30 percent reported they were known but were not written. There have been several efforts to codify best practices and develop policies related to the access and use, benefits and challenges, privacy and data security, and technical difficulties associated with the secondary use of health data. In 2006, the American Medical Informatics Association convened a panel to explore the issues associated with the use of health information for secondary purposes such as research. The panel published a white paper (Safran et al., 2007) that was intended to serve as a foundation on which a national framework governing the secondary use of health data would be developed. More recently, Karp et al., (2008) convened a panel of bioethicists, scientists, and legal experts to specifically examine the ethical issues associated with linking health databases and for developing guidelines permitting the aggregation of databases. In short, the panel recommended that initial consents should address the potential that information might be aggregated with other data sources, mechanisms be put in place to ensure data security and protect privacy interests, efforts be implemented to standardize data, data sharing policies be established, and a set of best practices be adopted for the merging of multiple data sources.

There are also a series of issues related to the scientific merit of the research conducted using administrative data. Proponents have argued that administrative and operational databases have many research advantages over narrowly focused, special-purpose data collection (Pandiani & Banks, 2003). Among the advantages that they noted are the comprehensiveness of these databases which include, among others, (1) minority populations in sufficient numbers to provide confident subgroup analyses and findings, (2) reduced problems of subjects lost to contact,

(3) opportunity to identify relevant comparison groups, and (4) the ability to replicate studies at minimal cost because the data already exist. Perhaps most importantly, administrative data, unlike experimental studies, permit the examination of interventions as they are typically provided in community settings where best practices may not be universal.

In contrast, others have argued that the increased use of administrative data for research purposes has created problems, particularly for the research community (Drake & McHugo, 2003). These challenges include (1) poor quality of administrative data, (2) statistical significance without meaningfulness, and (3) the use of multiple statistical tests that capitalize on chance, and post hoc interpretations. Many other authors have expressed concerns associated with the quality of administrative data (Broemeling et al., 2009; Mason, 1986; Rabeneck, et al., 2001), however, Segal (2003) argued that some of these concerns of scientific merit are not limited to research conducted using administrative data, but rather apply to how well any research methods allow the investigator to adequately address the question at hand.

FEDERAL ACTS GOVERNING ACCESS TO ADMINISTRATIVE DATA

In addition to the codification of the core research ethics principles through the Nuremberg Code and the Belmont Report, there have been three Federal Acts passed by the government that have important and direct implications related to access and sharing of information and data.[4] First, in 1966 President Johnson signed into law the *Freedom of Information Act* (FOIA; Public Law 89-554, 80 Stat. 383; Amended 1996, 2002, 2007), which requires the full or partial disclosure of information, records, and/or documents controlled by the U.S. Government upon written request unless the government substantiates that the information requested can be lawfully withheld under one of nine specific exemptions in the act. FOIA carries a presumption of disclosure and the right of access is ultimately enforceable in federal court.

A second act, the *Family Educational Rights and Privacy Act* (FERPA; 20 U.S.C. § 1232g; 34 CFR Part 99) passed in 1974, is a Federal law that protects the privacy of student education records. The law applies to all schools that receive funds under programs administered by the U.S. Department of Education. FERPA provides

[4] Each of these laws is also discussed later in this paper as they specifically impact data owners and researchers using administrative data. We note that there are certainly other laws and regulations that impact the use and sharing of data (e.g., other federal rules, state laws, foreign regulations), however these three are those typically encountered in academic research, and are examples of the types of laws and regulations in other contexts.

parents and eligible students (i.e., 18 years or older) certain rights regarding their children's education records. In general, schools must obtain written permission from the parent or eligible student prior to release of any information from a student's education record. However, FERPA allows schools to disclose information from student records, without consent, to certain parties and under certain conditions, which include officials in cases of health and safety emergencies, state and local authorities within the juvenile justice system, school officials with legitimate educational interests, or to schools to which a student is transferring. Parents or eligible students have the right to inspect and review the student's education records maintained by the school. Schools may disclose, directory information such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance without consent, but must inform parents and eligible students and allow them a reasonable amount of time to request that the information not be disclosed.

The Department of Education recently amended the regulations of the FERPA to increase the effective use of data in statewide longitudinal data systems (SLDS). The amendments increase a state's ability to evaluate education interventions and to build on successful programs to create a culture of continuous educational improvement. More specifically, the proposed changes enable authorized representatives of state and local educational authorities to conduct research using SLDS data by expanding the requirements for written agreements.

In 1996 Congress passed the *Health Insurance Portability and Accountability Act* (HIPAA; P.L.104-191). While the act primarily protects health insurance coverage for workers and their families when they change or lose their jobs, the provisions also address the security and privacy of health data. Entities subject to the HIPAA, known as covered entities, are required to protect individuals' health records and other identifiable health information by requiring appropriate safeguards to protect privacy and setting limits and conditions on the uses and disclosures that may be made of such information without patient authorization. Similar to FERPA, HIPAA rules grant patients' rights over their health information that include the right to examine and obtain a copy of their health records and to request corrections. The HIITECH provisions of the American Recovery and Reinvestment Act of 2009 (P.L. 111-5) provide incentives intended to promote the adoption of electronic health records. Given the anticipated increase in the electronic exchange of protected health information, these provisions also broaden the reach of HIPAA privacy and security protections, as well as enforcement mechanisms.

Definitions of Key Concepts

This section provides brief definitions of several concepts that are importantly associated with the ethical considerations related to the creation, maintenance, and utilization of secondary data. Given the variability associated with these definitions in various contexts and usages, we offer them to provide the reader with insight into our usage and meaning of them within the context of this paper.

Administrative (a.k.a., operational or secondary) data – data collected in the course of programmatic activities for the purposes of program operation, client-level tracking, service provision, or decision-making—essentially: non-research activities (Goerge & Lee, 2002). Iezzoni (2004) defined administrative data within a health context as data resulting from administering health care delivery, enrolling members into health insurance plans, and reimbursing for services. She noted the primary producers of administrative data are the federal, state, and local governmental entities, and private health care insurers. We use the terms administrative, operational, and secondary data as synonymous throughout this manuscript.

Data Owners (a.k.a., data custodians, data stewards) – the entity or organization with authority to collect, maintain, and use individuals' information for program monitoring and management.

Data Users (i.e., researchers) – individuals or entities external to the data owners, using the information for purposes other than program monitoring and management.

Due diligence – a legal phrase used to describe a range of assignments, obligations, reports and investigations, which take place in business, manufacturing, and law. In other words, due diligence refers to the standard of care that a reasonable person would exercise to avoid harm to self or other persons.

Ethics – Resnik (2010) noted that when the majority of people hear the word ethics, they think of rules distinguishing right from wrong. Ethics can be defined as the norms for conduct that differentiate acceptable and unacceptable behavior. To many people, these norms are so obvious that they are considered simple commonsense.

Research Ethics – The codification and application of norms, standards and/or professional codes differentiating acceptable and unacceptable behavior associated with the conduct of research. Central to these standards of practice are an individual's voluntary participation, right to privacy, confidentiality, equitable selection of subjects, and informed consent.

Ethics and Best Practices from a Data Owner/ Organizational Perspective

As discussed above, as the world has become more digitized and it seems like electronic data are everywhere, the owners and custodians of such data are increasingly concerned about how they should protect the data from inappropriate disclosure, and how they should determine who can have access to and use the data for research and other purposes. These concerns are generated not only from a professional ethics perspective (i.e., to protect the privacy of individuals whose information is contained in the data sets), but a professional liability position – no one wishes to be on the wrong end of litigation that asserts data in their custody and care were inappropriately disclosed or used.

As the use of electronic data for research purposes has exploded,[5] a number of recent attempts have been made to identify best practices for data protection and management from a data owner/custodian perspective (e.g., Bloomrosen & Detmer, 2008; Karp et al., 2008; Safran et al., 2007), however efforts to highlight principles for best practices have occurred for more than a quarter century. In 1986, Mason discussed four broad ethical concerns that arise as advancing technology allows for broader use of electronic information in research. Mason was concerned about privacy as information from disparate data sets are linked creating a more complete picture of individuals. He also discussed problems with the accuracy of administrative data when it is used for purposes it was not originally intended for, and ownership rights in an age where information can have great economic value. Finally, Mason (1986) discussed his concerns about when and to whom data access ought to be offered, balancing the three initial concerns (privacy, accuracy and ownership) with our ethical imperative to improve society, including the use of such data to do so.

In the two decades after Mason's (1986) original article, several other organizations described guidelines for using administrative data for research purposes. For example, Chamberlayne et al. (1998) described a government-based approach at the British Columbia Ministry of Health with explicit policies and procedures for addressing ethical issues, and Hotz et al. (1998) offered a comprehensive report on research uses of administrative data generated from an academic viewpoint (the Northwestern University/University of Chicago Joint Center for Poverty Research).

[5] Indeed a recent study of PubMed entries found a six-fold increase of published research that used existing medical records data between 2000 and 2007 (Dean et al., 2009).

More recently, Kelman, Bass and Holman (2002) provided basic guidance for establishing necessary agreements, constructing valid data sets, and protecting the confidentiality and privacy of individuals whose information is contained in data sets. Although these initial guidelines were a good start, discussion and guidance for data owners/custodians remained sparse, and more comprehensive guidelines were still needed.

In 2006 and 2007, the American Medical Informatics Association assembled panels of stakeholders and experts to discuss a variety of issues associated with the use of administrative or secondary health data for research purposes. The resulting two white papers (Bloomrosen & Detmer, 2008; Safran et al., 2007) discussed a framework upon which to build a national consensus addressing the secondary use of health data. Recommendations in the Safran (2007) paper included a call for on-going discussions and development of educational activities and policies related to the access and use of such data as well as ethical concerns (especially privacy issues). The Bloomrosen and Detmer (2008) paper extended the discussion by proposing a “data stewardship” concept that would allow more effective and streamlined use of administrative data for research. The work of the AMIA is important and both white papers offer a good discussion of broader issues that provide an innovative framework (particularly from a health care industry perspective). They have greatly added to the conversation about secondary use of health data.

More recently, Karp et al. (2008) convened a panel that explored the ethical and practical issues associated with use secondary data – particularly when linking and aggregating data sets. Their seven recommendations addressed concerns in three broad areas: 1) legal and ethical permissions (e.g., Did initial consents [if any] permit secondary use of information, and do future consent processes include permission for such secondary use?); 2) data security and confidentiality issues (e.g., Is personal information contained in the data sets protected from disclosure through appropriate confidentiality and data security processes?); and 3) appropriate and effective use of data (e.g., Does the data user appropriately understand the unique challenges posed by secondary data analysis, including standardization of data and data sharing protocols?). By offering some framing principles for the data owner/custodian-research relationship, Karp’s panel also (just as the AMIA panels) moves the discussion toward practical principles for use of administrative or secondary data for research purposes.

Finally, Stiles et al. (2011) attempts to move the discussion further by first providing findings from a national survey of Medicaid authorities on their current practices for the use of Medicaid data in research, and second, building on Mason (1986), the AMIA panels and Karp et al., to propose four primary ethical parameters that data owners/custodians must consider when allowing the use of their administrative data for research purposes. The recommended considerations also provide more practical advice or approaches that data owners/custodians can apply or adapt to their individual contexts. Although Stiles et al. (2011) attempts to add to the conversation about ethical and practical considerations when using secondary data in research, it is certainly not the end of the discussion. As Karp et al. (2008) points out "Data protection standards will evolve, and a methodology that was appropriate at one time may not be appropriate later." (pg. 1337) The conversation must continue as technologies, methodologies, and even cultural standards change and evolve.

PRINCIPLES IN USING ADMINISTRATIVE DATA FOR RESEARCH PURPOSES

Much of what follows is summarized from the Stiles et al. (2011) paper. The four principles or parameters that data owners/custodians need to consider are 1) **security** of the data; 2) **confidentiality** of information contained in the data; 3) **permission to use** data for research purposes; and 4) **appropriate/ethical use** of the data by the researchers. Each of these principles is discussed below.

Data Security

The first principle outlined by Stiles et al. is the security of the data. All human data containing private or potentially personal information, including administrative data, should be secured to protect against inappropriate disclosure, and data that are used for research purposes (where direct benefits will likely not accrue to the individuals whose information is contained in the data) must be especially protected. Data owners/custodians must secure data according to applicable laws (e.g., HIPAA, FERPA, or state law) or absent statutory provisions, according to standards established by tort law (e.g., case law principles reflecting industry custom). As technology and techniques are ever-advancing, it would not be prudent to outline detailed standards, as they could easily become outdated or obsolete in a short time. A better approach is for data owners/custodians to establish their own security protocols based upon industry standards, summarized by organizations such as the CERT Coordination Center at Carnegie Mellon University (CERT, 2011) or the System Administration, Networking and Security Institute (SANS Institute, 2011). Data owners/custodians should also confirm that any researchers who they might transfer data to also have such safeguards in place. Stiles et al. (2011) recommends balancing protocols in at least three areas to secure electronic data:[6]

a. **Training (expertise).** Popular media has noted that a major weakness in any security plan is the human element, thus a critical first consideration is a well-trained staff. The people who have access to, or can grant access to, sensitive data must understand the risks involved with disclosure of information and have the expertise to secure the data. There should be regular, mandatory training to ensure that all staff receive **foundational security education** on data privacy and security including detailed discussions of data owner/custodian security policies. In addition, opportunities for more advanced **professional development/training** in using and maintaining sensitive data should be offered. Finally, any organization maintaining and/or sharing sensitive data should establish an information security **awareness program** to reduce security lapses by both staff that have access to sensitive data and staff who do not have direct access, but may be able to use the organizational electronic network. Most importantly, data owners/custodians should also confirm that any entities with which they are sharing data have their own ongoing training programs.

b. **Policies (processes).** The second line of defense is the establishment of well-crafted policies and procedures that provide clear processes for ensuring that data are secure. Stiles et al. (2011) outline that policies should address data procurement and use (including the appropriate use of encryption); data security and access; security incident and disaster recovery procedures; recording and monitoring of system activity; and policy enforcement and training. Several models for security guidelines and policies are available online (e.g., SANS Institute, 2011; Litwak, 2011). Again, as with training, data owners/custodians should confirm that entities with which they are sharing data for research purposes have their own established policies and processes for handling and securing sensitive data. Any entity that will house or access the sensitive data should have comparable security to the data owner/custodian itself.

[6] The HIPAA Security Rule is another model for formulating security procedures. The Security Rule requires that all protected health information be secured according to industry standards in three areas: administrative safeguards (e.g., policies, procedures); physical safeguards (e.g., hardware, locked doors); and technical safeguards (e.g., encryption) (*Federal Register*, 2003).

c. **Technology (tools).** Technological security is what we typically think of first (before training and policies), and although it is third in this list, it is by no means the least important. However, technological safeguards are not sufficient if they are the only line of defense. The owner/custodian must decide where on the continuum of technological options it falls in order to determine whether enough safeguards are in place. If disclosure could be disastrous for the organization (e.g. highly sensitive data) or the organization is particularly risk-averse, then perhaps the data should be stored on a physically secured, password protected, isolated system that is disconnected from any networks and external links such as the Internet. A less secure, but still strong, option is to maintain a secured server behind a firewall with filtering and activity logging. Two-factor authentication is also advisable.[7] As technology is constantly advancing and those wishing to break into secured servers or networks are becoming more skilled, data owners/custodians should regularly consult experts to confirm current options and industry standards (e.g., SANS Institute, 2011; or CERT, 2011).

Confidentiality of Information

Most professional relationships require or are enhanced when personal information is kept in confidence (i.e., confidentiality); indeed the Surgeon General's Report on Mental Health (Mental Health: A Report to the Surgeon General, 1999) considered confidentiality as a core value created on notions that society desires the reduction of stigma and embarrassment, the fostering and maintenance of trust, and the protection of personal autonomy and privacy.[8] Such beliefs also extend to the research context and the use of administrative data systems in research,[9] with

[7] Two-factor authentication involves the requiring of two types of identification before access to secure areas is allowed. Typically two of the following three parameters are needed: 1) something the person **knows** (e.g. a password, a birth date, a zip code); 2) something the person **has** (e.g., an ID card, a key FOB); and 3) who the person **is** (e.g., a biometric/fingerprint scan, voice recognition). Two-factor authentication is becoming more widely used even in everyday transactions, so use in securing sensitive data is good practice. For example, many gas stations now require a person using a credit card at the pump to enter his/her billing zip code – the two factors being the card (i.e. having an object) plus entering the zip code (i.e., having unique knowledge).

[8] The terms confidentiality and privacy are terms used often in discussing the protection of private information. "Confidentiality" typically involves the expectation that certain information about a person will not be disseminated to others, while "privacy" refers to the avoidance of violating a person's body, space or liberty (Stiles & Petrila, 2011). In the discussion here, we primarily address issues of confidentiality.

[9] Indeed the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (45 CFR 160 and 164) focuses on such ideas that private health information should be kept confidential.

researchers often being held to some of the highest standards of conduct in this area (Stiles & Petrila, 2011). Thus anyone using administrative data for research purposes (including data owners/custodians) should clarify processes to safeguard the confidentiality of private information contained in data sets through written policies and procedures. Provisions may include auto-logging out of idle computers, and protocols for de-identifying, encrypting and disclosing data. The SANS institute (2011) and CERT (2011) offer a variety of models for policies. All staff coming into contact with data should be required to sign and execute agreements clarifying confidentiality policies and procedures that must be followed.

When developing policies for confidentiality, it can useful to consider provisions of several federal laws that might impact one's ability to protect against inappropriate or involuntary disclosure of data and information. As described previously, the HIPAA Privacy Rule (45 CFR 160 and 164) governs disclosures of protected health information (PHI) by "covered entities;" however many data owners apply the processes established by the rule to other types of sensitive data as well. Oversight protocols for disclosure of data under HIPAA are fairly well established at this point (Stiles & Petrila, 2011), allowing use of PHI for research purposes with patient authorization, with a waiver, or as a limited data set. An online publication by the Department of Health and Human Services (DHHS, 2003) provides an excellent discussion of the impact and processes of the HIPAA Privacy Rule on research efforts, and Lane and Schur (2010) offer an interesting framework for health data access and privacy through the use of novel concepts as "data enclaves". The Family Educational Rights and Privacy Act (FERPA)[10] provides guidelines for protecting the confidentiality of educational records. While school directories can be released without consent from students and parents, all other student information requires student/parent consent before release. There are a number of explicit exceptions however, including release of information to "school officials with legitimate educational interest;" "specified officials for audit or evaluation purposes;" and "organizations conducting certain studies for or on behalf of the school" (DOE, 2010). Thus, although FERPA establishes strong protection for the confidentiality of educational records, access to such information for research is allowable if consent is obtained, if the data are de-identified, or if the researcher is conducting a study for the school (Arwood, 2010).[11] Finally, the Freedom of Information Act (FOIA) has generated the concern among researchers that all federally funded research information could be discoverable under the act – especially after the Shelby

[10] 20 U.S.C. 1232g; 34 CFR Part 99

[11] For more detailed information about the Provisions and application of FERPA, see the Department of Education Website (ED.gov).

Amendment was promulgated in the early 1990s, specifically allowing public access to research information (Stiles & Petrila, 2011).[12] However, the Federal Office of Management and Budget has implemented a very narrow procedure only requiring disclosure of data in published studies that are used by a federal agency. There have not been any serious court challenges yet, so the concerns have subsided somewhat, and when the provision is challenged, the court will likely use a balancing test weighing the privacy of the individuals whose information is contained in the data versus the public's need to know the information (Stiles & Petrila, 2011).

One final confidentiality consideration is the common situation where data sets are integrated to form a new, more comprehensive source of information.[13] Each data set alone may not provide enough information to identify or disclose a sensitive pattern of behavior, but the combined data set may offer a more comprehensive picture, which is likely why the data sets were integrated. However, data owners/custodians must be careful to protect the information generated in these integrated data sets and ensure that researchers with whom data are shared also protect such data. Treating such "new" data sets similar to other sensitive data is probably the easiest and most conservative approach to dealing with integrated data systems. Indeed, many integrated data sets combine information in such a way as to provide enhanced information about individuals and even an increased possibility of re-identification of persons, therefore treating such combined data as other sensitive information could be necessary as a minimum standard of care.[14] Owners/custodians

[12] This has also concerned data owners/custodians using or sharing data in their care for research purposes.

[13] For example, a data owner/custodian, or a researcher the owner/custodian shared data with, might combine a service/demographic file from one entity (e.g. Medicaid), with a file from another source (e.g. Corrections or Department of Motor Vehicles).

[14] With ever-evolving technologies, re-identification of individuals, when two or more seemingly de-identified data sets are integrated, is a real possibility. From software that mines data sets for unique combinations, to probabilistic matching techniques, to cutting-edge facial recognition software (e.g., "Face Recognition: Anonymous no More", The Economist, July 30, 2011) protecting against re-identification becomes more complex. Wjst (2010) highlighted several strategies that could be used -to "de-anonymize" a publically released de-identified, large-scale genetic dataset and concluded that there was a "good chance" that the identities of some individuals in the data file could be established from these "anonymous" data. Similarly, Sherman and Fetters (2007) described the re-identification issues and concerns with spatially explicit (i.e., geocoded) data, indicating the most frequently used strategy to ensure confidentiality is to purposely compromise micro level accuracy. Data owners can require that any one cell in the matrix of data cross tabulation findings must be of a certain level (for example, the Center for Medicare and Medicaid Services requires that all disseminated data tables must have at least ten individual data points per cell), but that does not address the issue of the researcher having the ability to identify individuals through data integration and not reporting it. The data owner can modify the data extract or access provided to attempt to limit the researcher from possibly re-identifying individuals through integration with other data sets; however, that can render the resulting extract unusable or the interpretation of findings misguided. Once

may also wish to include notice provisions in any data use agreements executed with researchers using their administrative data requiring the researchers to inform the owner/custodian when their data are integrated with other information.[15]

Disclosure Permissions

Beyond the need for securing data and ensuring confidentiality, the appropriate use of the data should be managed. This involves obtaining adequate permission to use the data for research purposes and then ethically analyzing and interpreting findings. We will address the permissions issue here, and discuss ethical use in the section "Ethics and Best Practices from a Researcher Perspective".

When data owners/custodians, or others with whom the owners/custodians are sharing data, desire to access and use such data for research purposes, adequate permissions should be secured and well documented. While permissions to disclose and/or use data involve principally the individual rights of persons whose information is contained within the data sets, data owners/custodians need to address appropriate use concerns at the **institutional level**. Owners/custodians should clarify reasonable protocols addressing such use in their data use policies, including whether, and under what circumstances, data may be used for research purposes or disclosed to others who may conduct research (Stiles et al., 2011).[16]

Arguably more important than the institutional considerations are the permissions concerns at the **individual level**. Karp et al. (2008) warn that before any secondary research use of existing data occurs, a determination should be made whether the original consent allows for such use, and if not, any future consent processes with the same population should directly address the issue of secondary research uses of individually recorded data. The inclusion of permission for secondary use of data for research purposes in the original consent process is good practice (if such use

again, reliance on the ethics and assurances of the researcher and periodic monitoring by the data owner is perhaps the only reasonable solution at this point. Perhaps new techniques allowing for integration of data without the possibility of re-identification will be developed, but nothing like that is currently available.

[15] Another unfortunate issue is the obligations of the data owner and researchers when a breach of confidentiality occurs. When/how should individuals whose data were inadvertently disclosed be notified? Who should do the notifying? Since administrative data sets often contain millions of individuals' records, would the risk of a possible breach of confidentiality and subsequent need for notifying the individuals be cost prohibitive (i.e., would the risk stop the research from going forward)? This is an evolving area that should at least be discussed among the data owner and researchers, and perhaps included in a final data use agreement.

[16] Approaches to crafting data use policies could be influenced by considerations such as liability for disclosure of proprietary information or the institution's public image/political position after findings are disseminated.

is a possibility), and indeed is commonly done in health care settings, but many data owners/custodians who collect administrative data do not secure such direct consent.

As discussed previously, several federal laws indicate standards and processes for secondary use of certain kinds of data. For example, the HIPAA Privacy Rule permits PHI to be used for research after Privacy Board review and oversight.[17] Similarly, FERPA allows use of identified educational records provided specific consent is provided or the research is in direct support of the school program (i.e. on behalf of the school).[18] The bottom line is that owners/custodians, and researchers with whom owners/custodians have shared data, should well document that 1) the data were obtained appropriately; and 2) that adequate permission has been secured to use the data for the specific research purposes intended to be pursued. Otherwise, they may not have fulfilled their due diligence, and potential liability increases for all parties.

Appropriate Use

Provided that security and confidentiality protection has been established, and permission to disclose and use the data has been obtained, the fourth and last ethical principle or parameter comes to the fore. The appropriate use of data from the researcher perspective is discussed in more detail below, but it is also a great concern for data owners/custodians – not only in making sure researchers with whom data is shared are following ethical protocols, but also when the owner/custodian uses the data for its own research purposes. As Stiles et al. (2011) points out, appropriate use of administrative data for research purposes involves not only possessing the skills to organize, manipulate, and analyze the data, but also entails approaching the whole process of administrative data use in an ethical manner (that is, ensuring that the data and their limits are well understood so that interpretation of findings are adequately informed). Of the four principles, this is the least controlled by the data owner/custodian, as assertions of ethical behavior and approaches by researchers must be relied on. Nevertheless, there are a number of things an owner/custodian can do to facilitate appropriate use of the data.

[17] See DHHS (2003) for a more detailed discussion of the HIPAA Privacy Rule and Research.

[18] See DOE (2010) for a more detailed discussion of the provisions of the Family Educational Right and Privacy Act.

Although it is clearly the researchers' ethical responsibility to understand the data used in research (see discussion below), data owners/custodians can facilitate the appropriate use of data by providing assistance and intelligence regarding the validity and reliability of available variables, as well as offering information about the contexts in which data were collected. Owners/custodians should clarify how the researcher plans to assess the fidelity of the data set--if a researcher cannot detail how he or she will organize and check the fidelity and validity of the administrative data used, the owner/custodian should ask for a more formal data management plan.

The ethical researcher must take the time to understand the data and how it was collected in order to assess what questions it can be used to answer. Some have proposed specific techniques for determining the appropriateness of research questions (van Eijk, Krist, Avorn, Porsius & de Boer, 2001), however owners/custodians must also periodically monitor the use of their data by others. If resources are available, inclusion of representatives of the data owner/custodian on research teams, to assist with design and interpretation, can be helpful. At a minimum, owners/custodians should require the review of reports and other study publications before they are widely disseminated, to avoid misinterpretations and analytic surprises. These reviews need not involve censoring, but at least the data owner/custodian can inform the researcher about misinterpretations or prepare for the dissemination of findings not flattering for the owner/custodian – or even make suggestions regarding how the findings are reported (e.g., a more positive spin). The extent to which an owner/custodian goes in monitoring use of administrative data after disclosure to researchers, depends on the extent of concern about inappropriate use of the data.[19]

CONCLUSIONS FOR DATA OWNERS/CUSTODIANS

Data owners and custodians, in many respects, have great control over the data in their care through limiting access and carefully crafting data sharing agreements. However, they must also rely on the integrity and skills of the researchers whom

[19] Data owners are not necessarily obligated to monitor appropriate use as long as security, confidentiality, and permission issues are addressed. The ethical use of data typically falls on the researcher. Nevertheless, it is hoped that owners/custodians would want to decrease the misuse or misinterpretation of data provided to researchers. Implementation of periodic auditing of researcher data handling and use (including policy, training and technology activities) by data owners can help ensure the data they are sharing is appropriately utilized. Inclusion of liquidated damages language in data use agreements can convey the data owners' seriousness about the researchers meeting adequate standards of care in storing, compiling, and using the data.

use their data – whether the researchers reside within the owner/custodians' own organization, or if they are independent of the organization. Considering the principles and parameters raised above, along with following any local data sharing laws, regulations or policies will go far in ensuring administrative data are appropriately handled and used in research contexts. Operating according to industry standards will not only protect the organization from potential liability, but more importantly should protect the persons whose information is contained in the data sets from harm, and protect the data itself from misuse and misinterpretation.

As Stiles et al. (2011) note, there are models where a balance across all four parameters allow reasonable use of administrative data for research purposes, without "opening the store" to inappropriate analyses that could harm individuals and generate bad research with incorrect conclusions. The Manitoba Centre for Health Policy at the University of Manitoba is one such example, where in coordinating access to a data repository, the Centre facilitates good use of the data through initiatives like an online glossary and concept dictionary, so that researchers using their data can share insights, findings, and interpretations (MCHP, 2011). Such initiatives can go far in promoting high quality and ethical research with the rich information contained in existing and evolving administrative data systems.

Ethics and Best Practices from a Researcher Perspective

The ethical and best practice considerations when using administrative data for research purposes from a **researcher's perspective** is not greatly different from the issues faced by data owners and custodians – indeed many data owners/custodians are researchers themselves. Nevertheless, there are some important things to highlight when researchers – particularly those who are not the original data owners or custodians – use administrative data in their research protocols.

Non-owner/custodial researchers must ensure that both **internal** resources and **external** connections are well developed and maintained so that studies are informative and valid while protecting the integrity of the data and the privacy/confidentiality of individuals whose information is contained in the secondary data sets. Internal resources include establishing the infrastructure and expertise to reposit, compile, and analyze the data, taking into consideration hardware, software, and "fleshware"[20] requirements to adequately conduct administrative data

[20] The human part of the equation is often overlooked as technological advances make maintenance

research. The most crucial external connections involve partnering with individuals representing data owners/custodians, to not only establish a trusting relationship regarding data security and use,[21] but to provide an invaluable source of intelligence about the context of the data collection, how to use the data (e.g., which variables are useful), and the best approaches to interpreting various findings.

PRINCIPLES FOR RESEARCH USE OF ADMINISTRATIVE DATA

The same four principles for use of administrative data from Stiles et al (2011) apply (Security, Confidentiality, Permissions, and Appropriate Use), however while the first three are still important, the last principle (Appropriate Use) is particularly critical for researchers. There is not much literature on the ethical compilation and use of administrative data by researchers, thus much of what follows is based on the authors' own experience in conducting such research for many years.

Data Security

To ensure data received from a data owner or custodian are safe, the researcher should provide as good or better security for the data than that implemented by the data owner/custodian. The same considerations for training, policies, and technology that apply to owners/custodians also apply to researchers who are given permission to use the data in studies. Thus adequate **training** to ensure expertise should be implemented for all research team members having access to the data, including the principal researchers themselves. When a research team is small this is manageable, but when larger data center operations are involved, more formal documentation of expertise and training is advisable. In academic settings, even graduate students or other part-time team members who have access to data should receive at least foundational security education on data privacy and security including detailed discussions of any security policies and procedures.[22] Inconvenience is not an excuse to forgo properly training team members, and could open the researcher to liability for inappropriate disclosure and mistrust of the researcher when seeking access to data in the future.

and manipulation of data deceptively easy, and yet the individuals who are handling and analyzing the data are critical in the understanding and interpretation of findings.

[21] The researcher should secure and respect the administrative data as well or better than the data owners or custodians who allowed them access to the data.

[22] Larger data analytic operations should also provide opportunities for professional development training and implement a broader awareness program for staff who may not have direct access to data, but do have access to any networks on which data reside.

In smaller research teams, formal **policies and processes** for handling data are often not documented. However, basic intake and security policies can go far in not only creating awareness of the need for security, but in ensuring adequate security is actually realized. Indeed many data owners/custodians will want to make sure that their data are properly protected from improper disclosure, and even simple written policies can establish a relationship of trust that can ultimately produce better research analyses and interpretations. As mentioned above, Stiles et al. (2011) outline that policies should address data procurement and use (including the appropriate use of encryption); data security and access; security incident and disaster recovery procedures; recording and monitoring of system activity; and policy enforcement and training. Several models for security guidelines and policies are available online (e.g., Litwak, 2011; SANS Institute, 2011).

Finally, **technological security** needs to be carefully considered. For highly sensitive data, smaller research teams may wish to secure the data by keeping it on a separate stand-alone computer (avoiding intrusions from network connections), or they may wish to invest in secure network controls (e.g., logged firewalls), encryption of data, and at least two-factor access protocols.[23] Larger operations that must maintain data on a server or de-centralized network environment must implement appropriate technological controls, and organizations such as the SANS Institute (SANS Institute, 2011) and CERT Coordination Center (CERT, 2011) can provide information about current security standards.

Confidentiality

Confidentiality concerns for researchers using administrative data also parallel the discussion of the issue for data owners and custodians. Confidentiality is a core consideration in professional relationships, especially when the person whose private information is involved does not know about the use of the data by researchers. Clear confidentiality policies should be in place with all research team members agreeing to them before data access is allowed. Techniques such as de-identification and encryption should be standard procedures, with exceptions implemented only when absolutely necessary (e.g., when identifiers are needed for linking data sets).

[23] As noted above, two-factor authentication involves requiring two types of identification before access to secure areas is allowed. Typically two of the following three parameters are needed: 1) something the person **knows** (e.g. a password, a birth date, a zip code); 2) something the person **has** (e.g., an ID card, a key FOB); and 3) who the person **is** (e.g., a biometric/fingerprint scan, voice recognition).

The same federal laws that impact data owners/custodians can similarly impact researchers including the HIPAA Privacy Rule[24] and FERPA;[25] however, the researchers are usually the ones trying to gain access to the protected data rather than the party charged with preserving confidentiality. Although the federal rules are less direct with researchers' obligations to preserve confidentiality once they possess data,[26] they should still ethically protect the private information contained in any data originally subject to these federal laws (civil liability still can prevail). As with owners/custodians, researchers should also pay heed to concerns and advice regarding protecting "new" information gleaned from linked data sets (e.g., Kelman, Bass & Holman, 2002; Dokholyan et al., 2009). Special consideration and protection may be needed to appropriately preserve the confidentiality of information created by linking two or more data sets from one or more data sources (see previous discussion of this issue above).

Permissions

Obtaining the appropriate permissions to retain, access, and use administrative data sets is crucial for researchers before any data are accessed or transferred and analyses are conducted. Clarification should be made about whether all the needed permissions can be secured from the data owner/custodian, or if consent/authorization must be obtained from the individuals whose information is contained in the administrative data.[27] A good practice for the researcher is to make a **formal and detailed request for data** from the data owner/custodian.[28] In addition to permission clarification, this request should include things such as an introduction to the study, a list of needed variables, timespan and population parameters for the data, transfer media/methods, and a checklist for owner/custodian staff to follow to ensure better accuracy of data extraction. A **written agreement** (which can be executed before or after the formal request) between the researcher and the data owner/custodian should not only clarify the terms of data possession and what analyses and studies are allowed

[24] 45 CFR 160 and 164

[25] 20 U.S.C. 1232g; 34 CFR Part 99

[26] For example, once protected health information is disclosed to another entity by a covered entity, it is no longer considered protected health information; although any data use agreements clearly would still apply.

[27] Use of administrative data without individual consent is justifiable in a number of circumstances including statutory authorization and program evaluation. Regidor (2004) provides an interesting discussion of the debate in this area.

[28] Before a formal request is made, the researcher should Learn about the data (e.g., review data dictionaries, and layouts; read documents that used the data), and attempt to develop a good relationship with owner/custodian staff. This will help in subsequently compiling and using the data in research studies.

using the data, but also how the data will be secured and confidentiality maintained, whether additional permissions or consent will be required at an individual level (i.e. directly from individuals whose information is contained in the data sets), and how long the researcher may retain the data. The agreement should also address with whom the researcher may share the data (if anyone).[29]

Federal laws (e.g. HIPAA, FERPA) can influence what specific permissions and documentation are needed for disclosure of administrative data. FERPA is fairly clear on when specific authorization is required from individual students and parents before disclosure of educational data (see discussion above). HIPAA is also fairly clear on when individual authorization is needed, or when a waiver of authorization can be granted by the overseeing privacy board (DHHS, 2003). However, on the institutional level, if the researcher is not a business associate of the covered entity,[30] initial access should be secured through a limited data set and data use agreement. Specific requirements for such an agreement are detailed by the HIPAA regulations (DHHS, 2003).

Appropriate Use

Once the security and confidentiality safeguards are in place and required permissions are secured, the final parameter of concern for researchers is appropriate handling and use of the administrative data. While this is an area that is largely out of control of the data owner/custodian, it is probably the most critical for the researcher – and it is often overlooked and minimized by those conducting studies. It is an essential part of a researcher's ethical obligations that can be lost in "technospeak" and a focus on hardware and software security, and it is founded on principles of professional ethics and the obligation to conduct good research (or at least not knowingly conduct bad or inappropriate research). Simply put, obtaining access to and securing the data are only part of the researcher's obligations. The researcher must also ensure that he/she has **adequate capacity to use the data**, that the **data received are valid and useful for research** (i.e., to answer the research questions), and that the **research team has adequate understanding of the data** and the context within which they were collected to appropriately interpret findings.

[29] In the experience of the authors, most agreements indicate that any third party requests to the researcher for data extracts should be referred back to the original owner/custodian. To avoid misunderstandings and potentially legal liability, the researcher should only transfer data to a third party when a clear written authorization to do so is provided by the original data owner/custodian.

[30] Business associates of the covered entity under HIPAA are allowed broader access to data because they are essentially acting on behalf of the covered entity (e.g., conducting evaluations or operational research for the covered entity).

Establishing the capacity to maintain and use administrative data sets requires the researcher to not only procure the technology to handle and secure potentially large data sets, but also assemble the human expertise to effectively compile and analyze the data. Cross-platform knowledge, analytic software expertise, and change control processes (particularly in larger research teams) are especially important when using data from multiple systems. Part of possessing the analytic capacity involves understanding what the appropriate analyses to pursue are, and knowing that data should only be used for purposes documented in the data use agreement. With relatively large, population based data sets, researchers may experience the temptation to mine the data. Limited data mining may be justifiable in an exploratory study, but the testing of theoretically based research questions is typically a more informative experimental approach. In no instance is “fishing” for significance justifiable.[31]

Potential problems with using administrative data are well documented (e.g., Drake & McHugo, 2003; Ray, 1997), so any data compiled and analyzed, where the researcher is not in control of the data collection, should be carefully explored for fidelity and validity. This does not involve analyzing the data to answer research questions or to discover some new piece of information. Rather it simply involves ensuring that the data set received is adequate to use for the research purposes desired. There are no standards for assessing data fidelity and validity, as any assessment is necessarily specific to the context of the administrative data purpose and research questions being pursued.[32]

Perhaps one of the most under-discussed, yet potentially critical issues, is the obligation of the research team to adequately understand the data and the context within which they were collected in order to appropriately interpret findings. Understanding and appreciation of the limitations, strengths, and idiosyncrasies of the data are critical in determining appropriate analytic approaches as well as accurate and informed

[31] “Fishing” for significance, or running multiple significance tests (e.g. t-tests), on data to try to find interesting results is unethical, and even using procedures such as the Bonferroni correction does not make the practice more palatable. There is a vague line between exploratory data mining and unethical fishing, and if such uses of administrative data are pursued by researchers, they should be careful to document their intentions and processes to avoid producing invalid or misleading findings.

[32] The process of assessing fidelity and validity need not be difficult, but it certainly can take time. The assessment can involve such activities as reading all documentation about the data set, creating production/locked data sets (so only one version of the data is used in analyses), running key frequencies and crosstabs for reasonableness, assessing out of range and missing values, and verifying numbers with data owner/organizational audits or reports. Program logs can be used to document data fidelity testing. The researcher is essentially “playing” with the data to better understand quality of variables and potential of administrative or secondary data for research purposes (must avoid temptation to “fish” for substantive findings however).

interpretation.[33] Such understanding can be gained through the process to establish data fidelity and validity outlined above, but more importantly, should involve the ongoing communication with, and intelligence gathering from, individuals who have worked with the data before (e.g., staff of the data owner/custodian, other researchers who have analyzed similar data). Developing and maintaining on-going relationships with persons who already understand the data is crucial to appropriately designing analyses and then interpreting findings.[34] Maintaining such relationships can also help the researcher to recognize issues and findings that might be sensitive operationally or politically for the data owner/custodian, so that the researcher may be more discreet when the findings are disseminated. This does not suggest that the data owner/custodian should be authorized to censor or change the findings, but at least the owner/custodian should not be surprised by perceived sensitive results and conclusions.

Conclusions

Administrative data and other types of integrated data can provide insights not available from any other source – however users of such data (organizations and researchers) must ensure the security of the data and confidentiality of the information contained in the data, as well as carefully document the custodial and disclosure permission. Finally, the end user (researcher) must show due diligence to ensure that the data are appropriately used for the research purpose desired, which involves assessing the validity/fidelity of the data used, and taking the time to adequately understand the context from which the data come, in order to inform the interpretation of findings. The unique risks and benefits of integrating and analyzing administrative data need to be recognized by both data owners and researchers in

[33] For example, incentives for administrative clerks who entered the data to be accurate may not align with research analysis, leading to incomplete or over-representative data, or varying fidelity/accuracy. Commonly named fields (e.g. “gender”) may not have the same values across systems complicating integration. Finally, if significance testing is needed, large Ns and/or use of the population universe (which is often possible with administrative data) instead of a sample can greatly impact the usefulness of p-values.

[34] The authors unfortunately were witness to a scholar from a preeminent northeast university using Florida Medicaid claims data to examine a policy issue and who completely misinterpreted the analytic findings, because the scholar did not take the time to understand that the services that were the focus of the analyses would not be covered by Medicaid, but rather the state mental health authority paid for the services. The scholar disseminated the interpretation that persons in the state of Florida were not well served because they did not receive the service, when in fact they were receiving it through another funding source. A simple conversation with Medicaid staff or staff of the state mental health authority would have clarified the inaccuracy of the interpretation, but unfortunately that did not happen. It is incumbent upon the researcher (and even ethically required of the researcher) to make sure he or she understands the context of the data he or she is analyzing.

order to insure the ethical use of these data while protecting the confidentiality of the individuals whose private information is contained in the data. There is little that can deter a misguided administrative data user from unethical practices. However, the careful implementation and monitoring of the four principles of data, **security**, **confidentiality**, access **permissions**, and **ethical use,** through adequate training, established policies, and appropriate technology (Stiles et al., 2011) will greatly minimize the risk of this happening and will ensure that data owners have demonstrated the requirement of due diligence.[35]

[35] While a simple checklist would be nice (and indeed is utilized to show due diligence in other areas of law), this context of data collection and accessibility, as well as the ever-changing technologies involved could make such a simple checklist obsolete fairly quickly. The four principles, however, will remain applicable regardless of change, and thus data owners and researchers are encouraged to keep cognizant of the evolving standards in each of these broader areas, and apply reasonable approaches to ensure they are met.

References

Arwood, T. (2010). Educational Research and FERPA. Retrieved from http://www.clemson.edu/administration/ogc/documents/FERPA.pdf.

Black, C., McGrail, K., Fooks, C., & Maslove, L. (2005). *Data, Data Everywhere: Improving Access to Population Health and Health Services Data in Canada.* Vancouver and Ottawa: Centre for Health Services and Policy Research/Canadian Policy Research Networks.

Bloomrosen, M. & Detmer, D.E. (2008). "Advancing the Framework: Use of Health Data – A report of a working conference of the American Medical Informatics Association." *Journal of the American Medical Informatics Association*, 15, 715-722.

Broemeling, A. M., Kerluke, K., & Black, C. (2009). "Developing and maintaining a population research registry to support primary healthcare research." *Healthcare Policy*, 5, 65-74.

CERT. (2011). The CERT Coordination Center Homepage. Retrieved from http://www.cert.org/

Chamberlayne, R., Green, B., Barer, M.L., Hertzman, C., Lawrence, W.J., & Sheps, S.B. (1998). Creating a population-based linked health database: A new resource for health services research. *Revue Canadienne De Sante Publique*, 89, 270-273.

Council for International Organizations of Medical Sciences. (2002). *International Ethical Guidelines for Biomedical Research Involving Human Subjects.* Geneva, Switzerland: Council for International Organizations of Medical Sciences.

Dean, B.D., Lam, J., Natoli, J.L., Butler, Q., Aguilar, D., & Nordyke, R.J. (2009). Use of electronic medical records for health outcomes research. *Medical Care Research and Review*, 66, 611-638.

DHHS. (2003). Protecting personal health information in research: Understanding the HIPAA Privacy Rule. Retrieved from http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf.

Drake, R.E. & McHugo, G.J. (2003). Large data sets can be dangerous. *Psychiatric Services*, 54, 133.

Drake, R. E., & McHugo, G. J. (2003). Large data sets are powerful. *Psychiatric Services*, 54, 746.

DOE (2010). Family Educational Rights and Privacy Act (FERPA). Retrieved from U.S. Department of Education website: http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

Dokholyan, R.S., Muhlbaier, L.H., Falletta, J.M., Jacobs, J.P., Shahian, D., Haan, C.K. & Peterson, E.D. (2009). Regulatory and ethical considerations for linking clinical and administrative databases. *American Heart Journal*, 157, 971-982.

Federal Register (2001). 34 CFR Part 99, Part V, Family Education Rights and Privacy, Final Rule. Office of Family Policy Compliance, Family Education Rights and Privacy Act (FERPA). Retrieved December 27, 2010 from http://www.ed.gov/print/policy/gen/guid/fpco/ferpa/index.html .

Federal Register (2003). 45 CFR Parts 160, 162, and 164, Health Insurance Portability and Accountability Act of 1996 (HIPAA), Public Law 104–191, section 1176. Retrieved December 27, 2010 from http://aspe.hhs.gov/admnsimp/final/fr03-8334.pdf

Goerge, R. M., & Lee, B. J. (2002). Matching and cleaning administrative data. *New Zealand Economic Papers*, 36, 63-64.

Green, D. S. (2002). U.S. Senate weighs proposals on medical privacy. *Lancet*, 359, 1585.

Hotz, V.J., Goerge, R., Balzekas, J. and Margolin, F. (1998). Administrative data for policy relevant research: Assessment of current utility and recommendations for development. Chicago: Northwestern University/University of Chicago Joint Center for Poverty Research.

Iezzoni, L. I. (1997). Assessing quality using administrative data. *Annals of Internal Medicine*, 27(8S) Supplement, 666-674.

Karp, D.R., Carlin, S., Cook-Deegan, R., Ford, D.E., Geller, G., Glass, D.N., Greely, H., Guthridge, J., Kahn, J., Kaslow, R., Kraft, C., MacQueen, K., Mallin, B., Scheuerman, R.H., & Sugarman, J. (2008). Ethical and practical issues associated with aggregating databases. *PLoS Medicine*, 5(9), 1333-1337.

Kass, N. E., Natowicz, M. R., Hull, S. C., Faden, R. R., Phdnga, L., Gostin, L. O., & Slutsman, J. (2003). The use of medical records in research: What do patients want? *Journal of Law, Medicine and Ethics*, 31(3), 429-433.

Kelman, C.W., Bass, A.J. & Holman, C.D.J. (2002). Research use of linked health data – a best practice model. *Australian and New Zealand Journal of Public Health*, 26, 251-255.

Lane, J. & Schur, C. (2010). Balancing access to health data and privacy: Issues and approaches for the future. *Health Services Research*, 45, 1456-1467.

Litwak, P. (2011). A Pathway to HIPAA Compliance. Retrieved from http://www.hipaacomplianceguide.com/about_guide.htm.

Mason, R.O. (1986). Four ethical issues of the information age. *MIS Quarterly*, 10(1), 5-12.

MCHP. (2011). Manitoba Centre for Health Policy Webpage. Retrieved from http://umanitoba.ca/faculties/medicine/units/community_health_sciences/departmental_units/mchp/.

Mental Health: A Report to the Surgeon General. (1999). *Chapter 7: Confidentiality of mental health information: Ethical, Legal and Policy Issues*. Retrieved from http://www.surgeongeneral.gov/library/mentalhealth/pdfs/c7.pdf.

Nuremberg Code. (1949). Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No.10, 181-182.

Pandiani, J. A., & Banks, S. M. (2003). Large data sets are powerful. *Psychiatric Services*, 54, 745.

Pimple, K. D. (2002). Six domains of research ethics: A heuristic framework for the responsible conduct of research. *Science and Engineering*, 8, 191-205.

Ray, W.A. (1997). Policy and program analysis using administrative databases. *Annals of Internal Medicine*, 127, 712-718.

Rabeneck, L., Menke, T., Simberkoff, M. S., Hartigan, P. M., Dickinson, G. M., Jensen, P. C., George, L., Goetz, M. B., & Wray, N. P. (2001). Using the national registry of HIV-infected veterans in research: Lessons for the development of disease registries. *Journal of Clinical Epidemiology*, 54, 1195-1203.

Regidor, E. (2004). The use of personal data from medical records and biological materials: Ethical perspectives and the basis for legal restrictions in health research. *Social Science & Medicine*, 59, 1975-1984.

Resnik, D. B. (2010). What is ethics in research and why is it important? Retrieved December 16, 2010 from http://www.niehs.nih.gov/research/resources/bioethics/whatis.cfm

Robling, M. R., Hood, K., Houston, H., Fay, J., & Evans, H. M. (2004). Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. *Journal of Medical Ethics*, 30, 104-109.

Safran, C., Bloomrosen, M., Hammond, E., Laboff, S., Markel-Fox, S., Tang, P.C., & Detmer, D.E. (2007). Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association*, 14(1), 1-9.

SANS Institute. (2011). The SANS Institute Homepage. Retrieved from http://www.sans.org/.

Segal, S. P. (2003). Large data sets are powerful. *Psychiatric Services*, 54, 745-746.

Sherman, J.E., & Fetters, T.L. (2007). Confidentiality concerns with mapping survey data in reproductive health. *Studies in Family Planning*, 38, 309-321.

Sørensen, H. T., Sabroe, S., & Olsen, J. (1998). A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology*, 25(2), 435-442.

Stiles, P.G., Boothroyd, R.A., Robst, J. & Ray, J.V. (2011). Ethically using administrative data in research: Medicaid administrators current practices and best practices recommendations. *Administration & Society*, 43, 171-192.

Stiles, P.G. & Petrila, J. (2011). Research and confidentiality: Legal issues and risk management strategies. *Psychology, Public Policy & Law*, 17, 333-356.

U.S. Department of Health and Human Services [DHHS] (2000). P*HS Policy on Instruction in the Responsible Conduct of Research*. Washington, DC: Office of Research Integrity, DHHS.

van Eijk, M.E.C., Krist, L.F.G., Avorn, J., Porsius, A. & de Boer, A. (2001). Do the research goal and databases match? A checklist for a systematic approach. *Health Policy*, 58, 263-274.

Wjst, M. (2010). Caught you: threats to confidentiality due to the public release of large-scale genetic data sets. *BMC Medical Ethics*, 11, 21-24.