Chapter 4

MANAGING INFORMATION SECURITY AND PRIVACY IN HEALTHCARE DATA MINING

State of the Art

Ted Cooper[1] and Jeff Collman[2]

[1]Department of Ophthalmology, Stanford University Medical School, Palo Alto, California, 94304; [2]ISIS Center Georgetown University School of Medicine; Department of Radiology; Georgetown University Medical Center, Washington D.C., 20057

Chapter Overview

This chapter explores issues in managing privacy and security of healthcare information used to mine data by reviewing their fundamentals, components and principles as well as relevant laws and regulations. It also presents a literature review on technical issues in privacy assurance and a case study illustrating some potential pitfalls in data mining of individually identifiable information. The chapter closes with recommendations for privacy and security good practices for medical data miners.

Keywords

information security; privacy; confidentiality; integrity; availability; HIPAA; anonymity; risk management; privacy appliance; human subjects protection

1. INTRODUCTION

As the health care delivery system adopts information technology, vast quantities of health care data become available to mine for valuable knowledge. Health care organizations generally adopt information technology to reduce costs as well as improve efficiency and quality. Medical researchers hope to exploit clinical data to discover knowledge lying implicitly in individual patient health records. These new uses of clinical data potentially affect healthcare because the patient-physician relationship depends on very high levels of trust. To operate effectively physicians need complete and accurate information about the patient. However, if patients do not trust the physician or the organization to protect the confidentiality of their health care information, they will likely withhold or ask the physician not to record sensitive information (California HealthCare Foundation, 1999). This puts the patient at risk for receiving less than optimum care, the organization at risk of having incomplete information for clinical outcome and operational efficiency analysis, and may deprive researchers of important data. Numerous examples exist of inappropriate disclosure of individually identifiable data that has resulted in harm to the individual (Health Privacy Project, 2003). Concerns about such harm have resulted in laws and regulations such as the privacy rules of the Health Insurance Portability and Accountability Act (HIPAA) of 1996 directly governing the use of such information by most health care providers, health plans, payors, clearinghouses, and researchers. These laws and regulations may also indirectly govern the use of this data by the business partners of these entities. None of these laws forbid research or using technologies such as data mining. All require medical investigators, whether conducting biomedical research or quality assurance reviews, to take sound precautions to respect and protect the privacy and security of information about the subjects in their studies.

Data mining especially when it draws information from multiple sources poses special problems. For example, hospitals and physicians are commonly required to report certain information for a variety of purposes from census to public health to finance. This often includes patient number, ZIP code, race, date of birth, gender, service date, diagnoses codes (ICD9), procedure codes (CPT), as well as physician identification number, physician ZIP code, and total charges. Compilations of this data have been released to industry and researchers. Because such compilations do not contain the patient name, address, telephone number, or social security number, they qualify as de-identified and, therefore, appear to pose little risk to patient privacy. But by cross linking this data with other publicly available databases, processes such as data mining may associate an

individual with specific diagnoses. Sweeney (1997) demonstrates how to re-identify such data by linking certain conditions with the voting list for Cambridge, Massachusetts which contains demographic data on over 50 thousand voters. Birth date alone can uniquely identify the name and address of up to 12% of people on such compilations with birth date and gender up to 29%, birth date and 5-digit ZIP code up to 69%, and full postal code and birth date up to 97%.

Recent work has demonstrated ways to determine the identity of individuals from the trail of information they leave behind as they use the World Wide Web (Malin and Sweeney, 2001) (Malin et al., 2003). IP addresses from online consumers have been linked with publicly available hospital data that correlates to DNA sequences for disease. Data collected as individuals use the Internet to obtain health information, services and products also pose hazards to privacy but much less law and regulation governs the use and disclosure of this type of information (Goldman and Hudson, 2000).

In this chapter we explore issues in managing privacy and security of healthcare information used to mine data by reviewing their fundamentals, components and principles as well as relevant laws and regulations. We also present a literature review on technical issues in privacy assurance and a case study illustrating some potential pitfalls in data mining of individually identifiable information. We close the chapter with recommendations for privacy and security good practices for medical data miners.

2. OVERVIEW OF HEALTH INFORMATION PRIVACY AND SECURITY

Often voluminous, heterogeneous, unstructured, lacking standardized or canonical form, and incomplete, as well as surrounded by ethical considerations and legal constraints, the characteristics of patient health care records make them “messy.” Because they originate primarily as a consequence of direct patient care with the presumption of benefit for the patient, their use for research or administrative purposes must happen with care to ensure no harm to the patient. Inappropriate disclosure, loss of data integrity, or unavailability may each cause harm (Cios and Moore, 2002). Recent laws and regulations such as HIPAA provide patients with legal rights regarding their personally identifiable healthcare information and establish obligations for healthcare organizations to protect and restrict its use or disclosure. Data miners should have a basic understanding of healthcare information privacy and security in order to reduce risk of harm to individuals, their organization or themselves.

2.1 Privacy and Healthcare Information

The term "privacy" bears many meanings depending on the context of use. Common meanings include being able to control release of information about one's self to others and being free from intrusion or disturbance in one's personal life. To receive healthcare one must reveal information that is very personal and often sensitive. We control the privacy of our healthcare information by what we reveal to our physicians and others in the healthcare delivery system. Once we share personal information with our caregivers, we no longer have control over its privacy. In this sense, the term "privacy" overlaps with "confidentiality" or the requirement to protect information received from patients from unauthorized access and disclosure. For example, the HIPAA Privacy Standard (Department of Health and Human Services, 2002) requires healthcare providers, health plans and health plan clearinghouses to establish appropriate administrative, technical, and physical safeguards to protect the use and disclosure of individually identifiable health information. HIPAA draws on ethical standards long developed in the health care disciplines that identify protecting the confidentiality of patient information as a core component of the doctor-patient relationship and central to protecting patient autonomy. Thus, ethics, laws and regulations provide patients with certain rights and impose obligations on the healthcare industry that should keep patient health information from being disclosed to those who are not authorized to see it.

2.2 Security and Healthcare Information

Use of the Internet has resulted in recognition that information technology security is of major importance to our society. This concern seems relatively new in healthcare, but information technology security is a well established domain. A large body of knowledge exists that can be applied to protect healthcare information. A general understanding of security can be obtained by understanding:

1. Security Components
2. Security Principles
3. Threats, Vulnerabilities, Control Measures and Information Assurance
4. Achieving Information Security: Administrative, Physical, Technical Safeguards

2.2.1 Security Components

Security is achieved by addressing its components: confidentiality, integrity, availability and accountability.

1. Confidentiality is the property that data or information is not made available or disclosed to unauthorized persons or processes.
2. Integrity is the property that data or information have not been altered or destroyed in an unauthorized manner.
3. Availability is the property that data or information is accessible and useable upon demand by an authorized person.
4. Accountability is the ability to audit the actions of all parties and processes which interact with the information and to determine if the actions are appropriate.

2.2.2 Security Principles

In 1997 the International Information Security Foundation published the latest update to this set of generally-accepted system security principles (International Security Foundation, 1997):

1. Accountability Principle

The responsibilities and accountability of owners, providers and users of information systems and other parties concerned with the security of information systems should be explicit.

2. Awareness Principle

In order to foster confidence in information systems, owners, providers and users of information systems and other parties should readily be able, consistent with maintaining security, to gain appropriate knowledge of and be informed about the existence and general extent of measures, practices and procedures for the security of information systems.

3. Ethics Principle

Information systems and the security of information systems should be provided and used in such a manner that the rights and legitimate interests of others are respected.

4. Multidisciplinary Principle

Measures, practices and procedures for the security of information systems should take account of and address all relevant considerations and viewpoints, including technical, administrative, organizational, operational, commercial, educational and legal.

5. Proportionality Principle

Security levels, costs, measures, practices and procedures should be appropriate and proportionate to the value of and degree of reliance on the information systems and to the severity, probability and extent of potential harm, as the requirements for security vary depending upon the particular information systems.

6. Integration Principle

Measures, practices and procedures for the security of information systems should be coordinated and integrated with each other and with other measures, practices and procedures of the organization so as to create a coherent system of security.

7. Timeliness Principle

Public and private parties, at both national and international levels, should act in a timely coordinated manner to prevent and to respond to breaches of security of information systems.

8. Reassessment Principle

The security of information systems should be reassessed periodically, as information systems and the requirements for their security vary over time.

9. Equity Principle

The security of information systems should be compatible with the legitimate use and flow of data and information in a democratic society.

2.2.3 Threats, Vulnerabilities, Control Measures and Information Assurance

Numerous threats exist to computer systems and the information they contain originating from within and outside organizations. Some common threats include malicious code such as viruses, Trojan horses, or worms. Malicious code often takes advantage of vulnerabilities in operating system software but depends, too, upon organizational weaknesses such as the failure to deploy, update or train workers in the use of antivirus software. Malicious code may enable denial of service attacks, impersonation, information theft and other intrusions. Attacks by famous malicious code such as the Melissa or Lovebug viruses highlight the threat of “hackers”, outsiders with intent to harm specific organizations or network operations in general. Insiders with privileged access to network operations and a grudge against their employer actually wreak the most harm to say nothing of ill-trained workers unintentionally making mistakes.

For individuals with responsibility for protecting the security of computerized information assets, the important point to remember is that each computer system with its host organization has its own security weaknesses or vulnerabilities. To minimize the likelihood of harm from threats, organizations must perform an information security risk assessment which serves as the foundation for an information assurance plan. Because computer security is relative, i.e. absolute security does not exist, an information assurance plan seeks to apply cost-effective control measures to reduce to acceptable levels the likelihood of loss to an organization from likely threats. In other words, the information assurance plan is designed to manage risk. Control measures include policies, procedures, and technology. Risk assessments should be repeated periodically because both threats and vulnerabilities change over time and used to update information assurance plans.

The HIPAA Security Standard reflects good practice in the information security industry and, thus, provides guidance to medical dataminers about how to proceed. Thanks to HIPAA many resources have emerged in the last several years to help, including *The CPRI Toolkit: managing information security in healthcare* (see http://www.himss.org/resource) and *Managing Information Security Risks: The OCTAVE^{sm} Approach* (Alberts and Dorofee, 2003). The website of the National Institute of Standards and Technology contains a wealth of guidance on computer information security in general as well as specific topics (see http://crst.nist.gov, particularly the Special Publications section).

2.2.4 Achieving Information Security: Administrative, Physical, and Technical Safeguards

The measures to control threats and vulnerabilities can be organized into three categorizes of safeguards: administrative, physical and technical. The HIPAA Security Standard describes "Administrative Safeguards" as administrative actions, policies and procedures "to manage the selection, development, implementation, and maintenance of security measures to protect electronic protected health information and to manage the conduct of the covered entity's workforce in relation to the protection of that information" (Department of Health and Human Services, 2003, pg. 261). Administrative safeguards include policies and procedures such as risk assessment and management, assigning responsibility for information security, developing rules and procedures for assigning access to information, sanctioning misbehavior, responding to security incidents and implementing a security training and awareness program. Physical safeguards include policies, procedures and measures to control physical

access to information assets such as computer sites, servers, networks, and buildings. HIPAA focuses special attention on workstations processing patient information requiring hospitals to identify their uses as well as controls on physical access. Technical controls include the various devices typically associated with "information security" such as passwords, firewalls, and encryption as well as technical measures for assuring health information integrity. Virtual private networks, tokens for user access, audit logs and public/private key infrastructure (PKI) are examples of technical safeguards.

2.2.5 Laws and Regulations

The following regulations found in the Code of Federal Regulations (CFR) are likely to apply to the use of health care data in data mining in the United States:

1. Standards for Privacy of Individually Identifiable Health Information; Final Rule Title 45 CFR Parts160 and 164, known as the HIPAA Privacy Standard (Department of Health and Human Services, 2002),
2. Security Standards Final Rule Title 45 CFR Parts 160, 162, and 164, known as the HIPAA Security Rule, (Department of Health and Human Services, 2003), and
3. Department of Health and Human Services (HHS) or the Food and Drug Administration (FDA) Protection of Human Subjects Regulations, known as the Common Rule Title 45 CFR part 46 (Department of Health and Human Services, 2001) or Title 21 CFR parts 50 and 56, respectively (Food and Drug Administration, 2002).

The European Union, Canada, and Australia have instituted their own laws and regulations in this area (see list of websites cited in Section 9. resources).

A full explanation of the HIPAA regulations is beyond the scope of this chapter, however understanding some of its basic requirements are essential for those engaging in healthcare data mining. These regulations set the national floor for the use and disclosure of most personally identifiable health information in the health care delivery system in the United States. While they supersede contrary state laws, they do not supersede state laws and regulations that are more stringent. Many states have more stringent laws and regulations. A discussion of common questions follows (see also http://www.hhs.gov/ocr/hipaa/)

Who must comply with HIPAA?

Most healthcare providers, health plans and healthcare clearinghouses must comply with HIPAA. Excluded from HIPAA are healthcare providers that do not transmit electronic information, records covered by the Family Educational Rights and Privacy Act and employment health records held by a covered entity in its role as employer. Healthcare information collected by entities not covered by HIPAA is not subject to these regulations.

What information is protected by HIPAA?

The HIPAA Privacy Standard applies to individually identifiable health information including oral, written and electronic information used by covered entities to make decisions. For providers this includes medical records and billing records. For health plans this includes enrollment, payment, claims adjudication, and case or medical management record systems records. This protected health information is known as PHI. The HIPAA Security Standard only applies to electronic information and does not cover oral or paper information.

What rights does HIPAA grant patients?

Patients have a right to:

1. a notice of information practices from providers and health plans that states how PHI is used and protected,
2. obtain copies of their healthcare records,
3. amend their healthcare records, and
4. an accounting of disclosures made for purposes other than treatment, payment and healthcare operations.

What must entities covered by HIPAA do?

Covered entities must:

1. provide a notice of information practices and abide by the notice,
2. designate an individual to be responsible for privacy,
3. provide appropriate administrative, physical and technical safeguards for PHI.
4. only use and disclose PHI in accordance with the HIPAA Privacy Standard, and
5. have written agreements with business associates with whom they share PHI requiring the business associate to protect the PHI.

What are the key rules for use and disclosure of PHI?

1. Except for specific exclusions, an authorization from the patient is required for covered entities to use or disclose PHI for purposes other than treatment, payment and healthcare operations.
2. Only the minimum necessary amount of PHI may be used or disclosed to satisfy the purpose of the use or disclosure, with the exception that physicians may disclose the entire record to other providers for treatment purposes.

What is meant by healthcare operations?

Healthcare operations are the usual business operations of healthcare providers and health plans. Specifically included are: quality assessment and improvement activities, outcomes evaluation, development of clinical guidelines, population-based activities relating to improving health or reducing health care costs, protocol development, case management and care coordination, contacting of health care providers and patients with information about treatment alternatives; and related functions that do not include treatment, reviewing the competence or qualifications of health care professionals, evaluating practitioner, provider performance and health plan performance, conducting training programs in which students, trainees, or practitioners in areas of health care learn under supervision to practice or improve their skills as health care providers, training of non-health care professionals, accreditation, certification, licensing, or credentialing activities, underwriting, premium rating, and other activities relating to the creation, renewal or replacement of contracts, conducting or arranging for medical review, legal services, and auditing functions, including fraud and abuse detection and compliance programs, business planning and development, such as conducting cost-management and planning-related analyses for managing and operating the entity, including formulary development and administration, development or improvement of methods of payment or coverage policies; customer service, resolution of internal grievances, sale, transfer, merger, or consolidation.

What is the difference between health care operations and research?

For HIPAA, research means a systematic investigation, including research development, testing, and evaluation, that is designed to develop or contribute to generalizable knowledge. If the same query is used on the same data in one case to improve efficiency, and in the second case to contribute generalizable knowledge, it is not research in the first case but is in the second case. Additional protections must be in place for research.

What does HIPAA require for research?

HIPAA research requirements only apply to HIPAA covered entities (providers, health plans and health plan clearinghouses).

PHI used or disclosed for research must be authorized by the patient unless:

1. A waiver has been granted by a privacy review board or institutional review board (IRB):
   - The review board or IRB must document that they believe the PHI used or disclosed involves no more than minimal risk to the privacy of individuals based on:
     - An adequate plan to protect PHI identifiers from improper use and disclosure;
     - An adequate plan to destroy those identifiers at the earliest legal and practical opportunity consistent with the research, and
     - Adequate written assurances that the PHI will not be reused or disclosed to any other person or entity except as required by law, for authorized oversight of the research study, or for other research for which the use or disclosure of the PHI is permitted by the Privacy Rule.
   - The research could not practicably be conducted without the requested waiver.
   - The research could not practicably be conducted without access to and use of the PHI.
2. For reviews preparatory to research and with the researcher making the following written assertions:
   - The use or disclosure is sought solely to review PHI as necessary to prepare the research protocol or other similar preparatory purposes;
   - No PHI will be removed from the covered entity during the review; and
   - The PHI that the researcher seeks to use or access is necessary for the research purposes.
3. For research on decedent's information, the covered entity is assured by the researcher that the use or disclosure is solely for research on the PHI, and is necessary for research purposes.
4. If the PHI has been de-identified in accordance with the standards of the Privacy Rule and therefore is no longer PHI. HIPAA describes two approaches for de-identification, including 1) a person with appropriate knowledge and experience applies and documents generally accepted statistical and scientific methods for de-identifying

information, or 2) remove 18 specific identifiers listed in section 164.514 of the rule.

5. If the information is released as a limited data set as prescribed by the Privacy Standard (with 18 specific identifiers removed) and a data usage agreement with the researchers stating that they will not attempt to re-identify the information.

What are the exceptions to the requirement for authorizations prior to disclosure of PHI?

HIPAA permits disclosures without authorizations as required by law for public health, health oversight activities, victims of abuse, neglect or domestic violence, judicial and administrative proceeding, law enforcement, for specialized government functions (military and veteran activities, national security and intelligence, medical suitability, correctional institutions, public benefit programs) and research. It should be noted that for each of these exceptions there are additional provisions that govern the details of the disclosures. It should also be noted that HIPAA permits but does not require any disclosures.

In addition to the HIPAA Privacy Rule researchers must comply with the HHS and FDA Protection of Human Subjects Regulations.

What is the relationship of these regulations?

"There are two main differences. First, the HHS and FDA Protection of Human Subjects Regulations are concerned with the risks associated with participation in research. These may include, but are not limited to, the risks associated with investigational products and the risks of experimental procedures or procedures performed for research purposes, and the confidentiality risks associated with the research. The Privacy Rule is concerned with the risk to the subject's privacy associated with the use and disclosure of the subject's PHI.

Second, the scope of the HHS and FDA Protection of Human Subjects Regulations differs from that of the Privacy Rule. The FDA regulations apply only to research over which the FDA has jurisdiction, primarily research involving investigational products. The HHS Protection of Human Subjects Regulations apply only to research that is conducted or supported by HHS, or conducted under an applicable Office for Human Research Protections (OHRP)-approved assurance where a research institution, through their Multiple Project Assurance (MPA) or Federal-Wide Assurance (FWA), has agreed voluntarily to follow the HHS Protection of Human

Subjects Regulations for all human subjects research conducted by that institution regardless of the source of support. By contrast, the Privacy Rule applies to a covered entity's use or disclosure of PHI, including for any research purposes, regardless of funding or whether the research is regulated by the FDA" (National Institutes of Health, pg. 5, February 5, 2004).

What are the differences between the HIPAA Privacy Rule's requirements for authorization and the Common Rule's requirements for informed consent?

"Under the Privacy Rule, a patient's authorization is for the use and disclosure of protected health information for research purposes. In contrast, an individual's informed consent, as required by the Common Rule and the Food and Drug Administration's (FDA) human subjects regulations, is a consent to participate in the research study as a whole, not simply a consent for the research use or disclosure of protected health information. For this reason, there are important differences between the Privacy Rule's requirements for individual authorization, and the Common Rule's and FDA's requirements for informed consent. However, the Privacy Rule's authorization elements are compatible with the Common Rule's informed consent elements. Thus, both sets of requirements can be met by use of a single, combined form, which is permitted by the Privacy Rule. For example, the Privacy Rule allows the research authorization to state that the authorization will be valid until the conclusion of the research study, or to state that the authorization will not have an expiration date or event. This is compatible with the Common Rule's requirement for an explanation of the expected duration of the research subject's participation in the study. It should be noted that where the Privacy Rule, the Common Rule, and/or FDA's human subjects regulations are applicable, each of the applicable regulations will need to be followed (National Institutes of Health, pg. 10, February 5, 2004)

Under the Common Rule, when may individually identifiable information be used for research without authorization or consent?

"Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects" (Department of Health and Human Services, pg 5, August 10, 2004).

3. REVIEW OF THE LITERATURE: DATA MINING AND PRIVACY AND SECURITY

In the previous sections managing health information privacy and security has been described as required by organizations involved in the industry of delivering healthcare; e.g. healthcare providers, health plans, payors, and clearinghouses. In this section we will explore the additional issues that large scale data mining presents for managing health information privacy and security. Data mining offers many possible benefits to the medical community, including administrators as well as researchers. One example of the value that can be derived from large data collections is demonstrated by Kaiser Permanente's Northern California Region reduction of the risk of their members dying from cardiovascular causes so that it is no longer their number one cause of death. According to the 2002 Annual Report of the National Committee for Quality Assurance (2002, pg. 23), "Since 1996, appropriate cholesterol control (as defined by HEDIS, an LDL level of less than 130) among the CAD population has improved from 22 percent to 81 percent. Among eligible patients discharged after a heart attack, 97 percent were on beta-blockers. The mortality rate from heart attacks at KPNC hospitals are up to 50 percent lower than at similar hospitals across the state." This was made possible by the development of a clinical data repository to support real-time direct healthcare delivery to its membership (over three million individuals), evidence-based medical knowledge and use of this data to guide their healthcare delivery processes (Levin et al., 2001) (Pheatt et al., (2003).

As information technology has become commonly used to support the core processes of healthcare, enormous volumes of data have been produced. Numerous organizations desire access to this data to apply techniques of knowledge discovery. Privacy concerns exist for information disclosed without illegal intrusion or theft. A person's identity can be derived from what appears to be innocent information by linking it to other available data. Concerns also exist that such information may be used in ways other than promised at the time of collection. Ways to share person specific data while providing anonymity of the individual are needed. Stated another way, controls are needed to manage the inferences about individual identity that can be made from shared person specific data. The Federal Office of Management and Budget (1994) has developed an approach to limit disclosure from government data so that the risk that the information could be used to identify an individual, either by itself or in combination with other information, is very small. This Report on Statistical Disclosure Limitation Methodology, Statistical Policy, discusses both tables and microdata. The report includes a tutorial, guidelines, and recommendations for good

practice; recommendations for further research; and an annotated bibliography. Techniques, rules and procedures for tables (magnitude versus frequency, counts, suppression, random versus controlled rounding, confidentiality editing) and microdata (sampling, removing identifiers, demographic detail, high visibility variables, adding random noise, rank swapping, blank and imputation for randomly selected records and blurring) are documented.

3.1 General Approaches to Assuring Appropriate Use

Past experience has shown three approaches to be common for using personal data in research and secondary analysis: using personal data only with the subject's consent, using personal data without explicit consent with a public interest mandate, and making the data anonymous before use (Lowrance, 2002). The discussion above of HIPAA and the Common Rule address the first two of these techniques, obtaining subject consent and authorized public interest use such as public health. Developing methods for assuring data anonymity offers promise for the future.

> As Sweeney (2003, pg. 15) says: "The goal of pioneering work in data anonymity is to construct technology such that person-specific information can be shared for many useful purposes with scientific assurances that the subjects of the data cannot be re-identified."

A discussion of specific methods for making data anonymous follows below. Each specific approach embodies one or more of four general approaches to the problem of assuring against disclosure of confidential information when querying statistical databases containing individually identifiable information including: conceptual, query restriction, data perturbation, and output perturbation approaches (Adam and Wortmann, 1989). Unfortunately none of these approaches offers a completely satisfactory solution. The conceptual model has not been implemented in an on-line environment and the others involve considerable complexity and cost and may obscure medical knowledge.

3.1.1 Conceptual approach

In the conceptual model, a user can access only the properties of population (i.e. a collection of entities that have common attributes and its statistics such as patients of certain ages, genders and ZIP codes) and tables that aggregate information. The user thus knows the attributes of the population and its origin, but may not manipulate the data or launch queries that merge and intersect subpopulations from the collection. The user may

only see statistical tables that contain either zero or at least two individuals never information on a single individual. With no access to the data and tables with only data on more than one individual, disclosure of information about a single individual is prevented. Tables 4-1 and 4-2 illustrate acceptable and unacceptable data tables using the conceptual approach.

Table 4-1 displays attributes only about types of persons by age, sex and ZIP and includes cells with numbers larger than one. Table 4-2 displays the same types of data but includes one cell with information about only a single individual (Female, Age 31-40), a presentation not permitted in the conceptual approach. While this model is thought to provide anonymity, it has never been implemented at a practical level with a production software system.

Table 4-1. Population Attributes Acceptable Table

Attributes	# Male	# Female	ZIP
Age 10-20	2031	2301	94027
Age 21-30	231	243	94027
Age 31-40	24	27	94027

Table 4-2. Population Attributes Unacceptable Table

Attributes	Male	Female	ZIP
Age 10-20	231	241	94027
Age 21-30	23	24	94027
Age 31-40	2	1	94027

3.1.2 Query restriction approach

Five methods have been developed to restrict queries:

1. query-set-size control - a method that returns a result only if its size is sufficient to reduce the chances of identification,
2. query-set-overlap control - a method that limits the number of overlapping entities among successive queries of a given user,
3. auditing – a method that creates up-to-date logs of all queries made by each user and constantly checks for possible compromise when a new query is issued,
4. cell suppression – a method that suppresses cells that might cause confidential information to be disclosed from being released, and
5. partitioning – a method that clusters individual entities into a number of mutually exclusive subsets thus preventing any subset from containing precisely one individual.

3.1.3 Data perturbation

This approach alters the data before permitting access to users. For example the source data is replaced with data having the same probability distribution (Islan and Brankovic, 2004). In other words, noise is inserted in the data that seeks to achieve anonymity and at the same time not change the statistical significance of query results. Users do not have access to the original data.

3.1.4 Output perturbation

This approach permits use of the original data, but modifies or renders the output incomplete. Techniques of output perturbation include processing only a random sample of the data in the query, adding or subtracting a random value that will not alter the statistical difference from the result, and rounding up or down to the nearest multiple of a certain base. Murphy and Chueh have published a successful implementation of query output perturbation to determine if a research database contains a set of patients with specific characteristics of sufficient size for statistical significance (Murphy and Chueh, 2002). In this example, the query result alters the number of patients with the specific characteristics by adding or subtracting a small random number. In addition, for values nearing zero, a result of less than three is presented.

3.2 Specific Approaches to Achieving Data Anonymity

Rendering data anonymous assures freedom from identification, surveillance or intrusion for the subjects of medical research or secondary data analysis while allowing data to be shared freely among investigators (Meany, 2001). Achieving complete data anonymity poses a considerable challenge. For example, 87% of individuals in the United States can be uniquely identified by their date of birth, gender and 5-digit ZIP code (Sweeney, 2002). True anonymity also poses ethical problems of its own, including loss of the possibility of benefit to the individual patient from knowledge discovered from the data, greatly increased complexity of maintaining an up-to-date database, and elimination of some checks against scientific fraud (Behlen and Johnson, 1999).

A number of techniques exemplifying or combining the general approaches described above have been advocated to help address this issue, including:

1. Data aggregation

2. Data de-identification
3. Binning
4. Pseudonymisation
5. Mediated access

3.2.1 Data Aggregation (an example of the Conceptual Approach)

Providing access only to aggregate data while prohibiting access to records containing data on an individual constitutes one approach commonly advocated to reduce risks to privacy. Although this approach does protect privacy, it critically limits medical research. Clinical research requires prospectively capturing and analyzing data elements associated with individual patients. Outliers are often a major focus of interest. Aggregate data does not support such efforts.

3.2.2 Data de-identification (an example of the Data Perturbation Approach)

The HIPAA Privacy Standard excludes de-identified health information from protected health information. De-identified health information may be used and disclosed without authorization. The HIPAA Privacy Standard considers information to have been de-identified by the use of either a statistical verification of de-identification or by removing 18 explicit elements of data. Such data may be used or disclosed without restriction. The details of these approaches are described in the pamphlet, *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule* (Department of Health and Human Services, July 13, 2004).

While these approaches to de-identification provide compliance with the HIPAA Privacy Standard, they do not guarantee anonymity. A number of reports have appeared recently that criticize these approaches for being too complicated and posing a threat to clinical research and care (Melton, 1997) (Galanddiuk, 2004). A variety of approaches for this issue have been published including successful implementation of policies, procedures, techniques and toolkits that meet academic medical center needs and comply with the Privacy Standard (UCLA) (Sweeney 1996) (Moore et al, 2000) (Ruch et al, 2000) (Moore et al 2001) (Thomas et al, 2002), (Lin et al, 2004) (Oiliveira and Zaïane, 2003) (Saul, 2004).

Goodwin and Prather performed a study that de-identified the data in the Duke TMR perinatal database in accordance with the HIPAA Privacy Standard and assigned a coded identifier to each patient to permit re-identification of patients under controlled circumstances. The database

contained data on 19,970 patients with approximately 4,000 potential variables per patient (Goodwin and Prather, 2002). They noted several issues:

1. To meet the requirement for removing all elements of date except year required the conversion of dates to days since conception to permit the data to be useful for pregnancy studies.
2. Clinician users were still able to identify one patient by her extremely young age.
3. The process was tedious, time-consuming and expensive.

They concluded that it is imperative to maintain the public's trust by doing everything possible to protect patient privacy in clinical research, and privacy protection will require careful stewardship of patient data.

3.2.3 Binning (Another example of the Data Perturbation Approach)

Binning deploys a technique for generalizing records in a database by grouping like records into a category and eliminating their unique characteristics (for example, grouping patients by age rather than date of birth). Elegant work has been done using this approach. One approach permits the level of anonymity to be controlled and matched to a user profile indicating the likelihood that data external to the database would be used permitting re-identification (Sweeny, 1997). Another report provides a measure of the information loss due to binning (Lin et al, 2002).

3.2.4 Pseudonymisation (Another example of the Data Perturbation Approach)

This technique involves replacing the true identities of the individuals and organizations while retaining a linkage for the data acquired over time that permits re-identification under controlled circumstances (Quatin et al., 1998). A trusted third party and process is involved. The trusted third party and process must be strictly independent, adhere to a code of conduct with principles of openness and transparency, have project-specific privacy and security policies and maintain documentation of operating, reporting and auditing systems (Claerhout et al., 2003).

3.2.5 Mediated access (A combination of Query Restriction and Output Perturbation Approaches)

Mediated access puts policy, procedure and technology between the user and the data and, thus, illustrates a general point that all medical investigators should bear in mind: sound health information privacy and security programs include a range of controls (Wiederhold and Bilello, 1998). "The system is best visualized as residing on a distinct workstation, operated by the security officer. Within the workstation is a rule-based system which investigates queries coming in and responses to be transmitted out. Any query and any response which cannot be vetted by the rule system is displayed to the security officer for manual handling. The security officer decides to approve, edit, or reject the information. An associated logging subsystem provides both an audit trail for all information that enters or leaves the domain, and provides input to the security officer to aid in evolving the rule set, and increasing the effectiveness of the system." (Wiederhold et al, 1996). The workstation, nonetheless, depends on and functions as a component of a broader security architecture that provides layered protection against unauthorized access by deploying sound practices such as encryption of transmissions, intrusion prevention with firewalls and a public/private key infrastructure. When functioning as a whole, the workstation and technical infrastructure provide several security controls, including:

1. authentication of users (optionally more extensive for external users),
2. authorization (determination of approved role),
3. processing of requests for data using policy-based rules,
4. initiating interaction of security officer oversight for requests that conflict with rules,
5. communication of requests that meet the rules to the internal databases,
6. communication from the internal databases of unfiltered results,
7. processing the unfiltered results to ensure that policy rules are met,
8. initiating interaction with security officer oversight when results do not meet the rules,
9. writing origin identification, query, action and results to a log file, and
10. transmission of data meeting rules to requestor (Wiederhold, 2002).

Ferris and colleagues report an approach that adds de-identification and re-identification to other security controls and supports the HIPAA

requirement for accounting for disclosures (Ferris et al, 2002). There are two modules in this approach:

1. Key Escrow Module

   This module consists of a privacy manager that uses key escrow to support de-identification and re-identification, user authentication, logging of user sessions, generation and storage of query-specific public/private keys and manages role-based access.

2. Biomedical Database Module

   This module associates the research database with the database manager and audit database. It is used for accessing the research data, generating an audit trail and de-identifying results when required.

3.3 Other Issues in Emerging "Privacy Technology"

Two kinds of privacy issues for computer science research have been identified: those inherent in applications of developing technology and those related to information practices needed in the development of technology. New efforts in "privacy technology" attempt to protect individual privacy while permitting the collection, sharing and uses of person-specific information. This research addresses two major concerns: disclosure of individually identifiable sensitive information by the linkage of information with other publicly available databases, and the use of information obtained for one purpose for another purpose. Threats to Homeland Security have made considerable funding available to investigate this topic in order to support bio-terrorism surveillance and protect individual privacy.

For example, Sweeney and colleagues at Carnegie-Mellon University have built "CertBox" to provide privacy protection in biosurveillance.

> "Emergency room visits and other healthcare encounters will be reported daily to the state's public health department under the authority of public health law. Collected health information will be filtered in real-time by a self-contained machine called a CertBox, which automatically edits combinations of fields (often demographics) so that released information relates to many people ambiguously. Settings are preset for a specific population and set of data fields and then sealed to prohibit tampering. CertBox technology de-identifies health information in accordance to the scientific standard of de-identification allowed under HIPAA. The resulting de-identified data is then shared with bio-terrorism surveillance systems. CertBox technology (more generally termed a "privacy appliance" by DARPA) allows us to certify that resulting data are properly de-identified and to warranty that resulting data remain practically useful for anomaly detection algorithms in bioterrorism surveillance" (Sweeney, L., 2003, p.15).

Other aspects of privacy technology include detecting and removing or replacing identifying information from information in text (e.g. medical reports, letters, notes, email) (Sweeney, L, 1996) (Ruch, et al., 2000) as well as facial images (Newton, et al., 2003). Techniques have been reported for embedding encrypted digital watermarking and patient identifiers in medical images (Tzelepi, 2002) to protect privacy during use and transmission.

Data mining investigators have begun encouraging their colleagues to take a research interest in issues related to protecting the privacy and security of personal information. For example, Berman argues that:

> "Human subjects issues are a legitimate area of research for the medical data miners. Novel protocols for achieving confidentiality and security while performing increasingly ambitious studies (distributed network queries across disparate databases, extending the patient's record to collect rich data from an expanding electronic medical record, linking patient records to the records of relatives or probands, peer-to-peer exchange of medical data) will be urgently needed by the data mining community" (Berman, 2002).

The techniques of data mining have been used to address the issue of auditing access and use of data as well as for testing devices for intrusion detection and access control. Commercial products exist that automatically correlate and compare suspicious information gathered from different points in computer systems, draw conclusions, and act on potential attacks and security violations (Dicker, 2003).

Berman's suggestion illustrates a general point: research into privacy and security technology necessarily entails the study of values and their embodiment in technological artifacts. Instead of assuming that ensuring privacy necessarily requires sacrificing research efficiency and efficacy, Berman's suggestion pushes researchers toward considering their relationship in specific instances and developing new approaches to both privacy and research design. In this respect, Berman echoes core concerns of a major body of research in the field of Human-Computer Interaction, known as "Value Sensitive Design" (Friedman, Kahn, and Borning, draft June 2003; Taipale, 2003).

3.4 "Value Sensitive Design": A Synthetic Approach to Technological Development

"Value Sensitive Design" attempts to incorporate relevant important considerations (values) into new technology throughout the entire lifecycle of design, development, deployment and retirement. Deriving inspiration from related ideas in computer ethics, social informatics, computer

supported cooperative work and participatory design, value sensitive design implements the basic proposition that all technology relates in design and use to important values and, therefore, cannot fundamentally emerge as "value-neutral" (Friedman, Kahn, and Borning, draft June 2003). Value sensitive design enables incorporating any value into the design process but places high priority on human values "with ethical import" including privacy as well as related values in health care such as human welfare, trust, autonomy, accountability, identity, and informed consent.

In practice, value sensitive design includes conceptual, empirical and technical investigations. In conceptual investigations, designers consider key questions about the context of technological implementation such as "Who are the direct and indirect stakeholders and how are they affected?" "What values are implicated and how do the designers evaluate their relative importance and priority?" Empirical investigations entail study of the actual human contexts of technological implementation. Technical investigations evaluate the relative support particular technological designs provide for the interests of specific stakeholders and realization of specific values. These investigations often identify conflicts among stakeholders and values that must be addressed in the design process. For example, in designing data mining technologies for medical investigations, stakeholders include investigators, study subjects and patients with the disease. Values include assuring integrity of research data as well as enhancing the welfare of patients and protecting subject privacy. The properties of specific technical designs may provide greater support for the interests and values of one group of stakeholders (for example, the subjects and their privacy) than for others. All design methodologies inevitably make choices of these kinds. Value sensitive design has developed means for making explicit the choices and their rationale (Friedman, Kahn, and Borning, draft June 2003).

In a spirited defense of data mining in bioterrorism surveillance, Taipale invokes the principles of value sensitive design in justifying privacy protections slated for development under the Terrorist Information Awareness (TIA) program (Tiapale, 2003) (see case study below for detailed review of TIA). TIA included programs for developing privacy appliances incorporating what Taipale calls rule-based processing, selective revelation, and strong credentialing and auditing. Rule-based processing entails research on intelligent query agents that negotiate access to specific data bases depending on the inquirer's authorization and meta-data labels about specific data items. Selective revelation technologies employ

> "an iterative, layered structure that reveals personal data partially and incrementally in order to maintain subject anonymity. Initial revelation would be based on statistical or categorical analysis …. This analysis would be applied to data that was sanitized or filtered

> in a way so that it did not reveal personally identifying information. Based on initial results, subsequent revelations may or may not be justified. At each step, legal and technical procedures can be built in to support particular privacy policies (or other policies, such as security clearances, etc.) (Taipale, 2003 pg. 79)".

Strong, encrypted tamper-proof auditing mechanisms that log access to distributed databases help protect against insider and outsider threats. Sweeney's "CertBox" constitutes an example of such a privacy appliance.

3.5 Responsibility of Medical Investigators

In addition to the usual security risks, medical research may add potential loss of life or health and requires special emphasis on privacy, confidentiality and data integrity (Berman, 2002). The Common Rule provides for subject safety and, with the HIPAA Privacy Standard, specifies accountability for the use and disclosure of protected health information. The medical data miner must conduct only valid research and in a way that protects human subjects. A privacy review board or institution review board is required to provide oversight for each project. Careful attention must be given to assure that there is proper justification and documentation of the process especially for waivers for individual consent or use of research conducted under the exemption of the common rule or the HIPAA Privacy Standard (Berman, 2002).

The Utah Resource for Genetic and Epidemiologic Research (RGE) (http://www.research.utah.edu/rge/) is an example of a well established medical data mining implementation where responsibilities have been made explicit (Wylie J.E., and Mineau, G.P., 2003). It has been functioning since 1982 when it was established on executive order of the Governor of Utah to be a data resource for the collection, storage, study and dissemination of medical and related information for the purpose of reducing morbidity or mortality. It does not perform research but maintains and improves data for research projects. The RGE contains over six million records. It includes genealogies on the founders of Utah and their descendants, cancer records, birth and death certificates, driver's license records, census records and follow-up information from the Health Care Financing Administration.

It has the following policies and procedures:

1. All data received by the RGE comes from contributors who have contracts that specify the conditions for use of the data and requires the data contributors to approve projects that use their data.

2. A committee composed of contributors and others familiar with the issues of medical data and research review all requests for access.
3. IRB approval is required for each project.
4. Access is project specific and may not be used for other purposes.
5. Data must be destroyed or returned at the end of the project.
6. Projects must justify reasons for access to information that identifies individuals.
7. Identifying information is removed from records and stored in separate tables in a relational database and requires the RGE staff to recombine the data for record linking and data management.
8. If projects wish to contact individuals, they must arrange for data contributors or their designees to contact the individual about interest in the proposed project. Identifying information is provided to the project only for those individuals wishing to participate. Other information or biospecimens are only collected after informed consent is obtained.
9. For-profit organizations may not have direct access to RGE data. They may participate with university or other non-profit entities. Commercial project sponsors may participate directly only in research activities that involve no individually identifying information.

Sweeney (2003, pg. 13) makes a plea to fellow computer scientists whose point applies as well to medical investigators:

> "Most computer scientists can no longer afford to do their work in an ivory tower and rely on the social scientists and lawyers to make decisions about limits of its use. First, policy makers and lawyers may not fully understand the technology. Second, decisions will often be made as a reaction to biased or sensationalized public opinion. Third, policy decisions are often crude and sub-optimal, and tend to legislate over simple technical remedies. Finally, there is a horrible temporal mismatch -- policy can be a function of years but new technology is a function of months, so policy enacted on today's technology may be totally inappropriate for tomorrow's and policy supporting technology today can prohibit it tomorrow. Computer scientists can and must insulate their creations from such risk."

In other words, medical investigators must proactively take responsibility for assuring adequate privacy controls in their projects. To ignore this responsibility risks potentially ceding control or completely losing their projects. The case study of the Terrorist Information Awareness Program illustrates just such a scenario.

4. CASE STUDY: THE TERRORIST INFORMATION AWARENESS PROGRAM (TIA)

4.1 The Relevance of TIA to Data Mining in Medical Research

This case study examines the controversy surrounding termination of the Terrorist Information Awareness (TIA) Program, a very large counterterrorism effort organized by the Defense Advanced Research Projects Agency (DARPA). At first glance, the TIA program would appear to have no relevance to data mining in medical data analysis and research because of its focus on crime prevention and law enforcement. This difference in mission, however, should not distract analysis from some core commonalities with respect to privacy and security as well as functionality. Data mining in counterterrorism and medical data analysis face the problem of developing legal, ethically appropriate and secure methods of managing individually identifiable information. Whether beginning with identified possible subjects for investigation and medical research, or discovering possible subjects in the course of analysis, organizations conducting terrorist investigations and medical data analysis must obey relevant privacy laws, establish appropriate policies and procedures, train their workforce, and implement risk-based administrative, physical and technical privacy and security safeguards.

As will be explained below, TIA lost funding and faced censure from journalists, the Inspector General of the Department of Defense, and Congress partially because DARPA paid insufficient attention to some of these core controls. Medical investigators similarly ignore the guidance of HIPAA at their peril and may take TIA's experience as an object lesson in what to avoid. As medical researchers study the TIA program, they will also find some of its proposed data mining capabilities very attractive, particularly the powerful data aggregation, analysis and linking tools as well as the virtual collaboration tools described below. Quite rightfully wanting to apply such capabilities to their research, medical investigators confront another shared characteristic with TIA, having to balance the rights and

welfare of individuals with possible benefits of their work for society as a whole. Patients and terrorists potentially face entirely different sets of consequences if identified by investigators as fit subjects for analysis. Patients may benefit or contribute to knowledge that benefits others in their situation. Terrorists potentially face punishment. But, in both cases, the investigators using data mining techniques of various kinds potentially gather data about unrelated persons risking invasion of privacy and potentially broader harms. The American public basically accepts as legitimate the aims of medical research and counterterrorism. As the fate of TIA demonstrates, however, individual programs must carefully assess, judiciously weigh and clearly explain the trade-offs between individual and societal welfare in specific instances, particularly in these times of struggle.

4.2 Understanding TIA

DARPA's "Report to Congress regarding the Terrorism Information Awareness Program" (DARPA, 2003) outlined the intended goals, responsible organizational structure, system components, expected efficacy, and relevant federal laws and regulations as well as privacy concerns of the program known as the Terrorism Information Awareness Program or TIA. Under the original name of the "Total Information Awareness Program", DARPA planned to create a system of systems "to integrate technologies into a prototype to provide tools to better detect, classify, and identify potential foreign terrorists." (DARPA, 2003 p. 3) The target technologies existed in a range of states of development from desirable but not yet acquired or developed to transitioning to operational use (DARPA, 2002 p 15.) Multiple universities, small and large defense contractors, and Federally-funded Research and Development Organizations worked on the various component technologies, many of which were funded as part of DARPA's ongoing research in counterterrorist technologies. DARPA created the Information Awareness Office (IAO) in January 2002 "to integrate advanced technologies and accelerate their transition to operational users." (DARPA, 2003, p. 1) The TIA prototype network included a main node in the Information Operations Center of the US Army Intelligence and Security Command (INSCOM) with additional nodes at subordinate INSCOM commands as well as other defense and intelligence agencies such as the National Security Agency, the Central Intelligence Agency and the Defense Intelligence Agency. In the post-September 11, 2001 world, TIA was supposed to help accelerate development and deployment of core tools in the fight against the "asymmetric threat" posed by terrorists "in full compliance with relevant policies, laws, and regulations, including those governing information about US persons." (DARPA, 2003, pg. 4)

TIA drew its component systems from three other categories of work coordinated by DARPA's IAO, including the Data Search, Pattern Recognition and Privacy Protection Programs, the Advanced Collaborative and Decision Support Programs, and the Language Translation Programs. The Data Search, Pattern Recognition and Privacy Protection Program coordinated development of technologies with the intention of seeking, analyzing and making available to decision-makers information about individually identifiable human beings potentially associated with terrorism. To find this data, TIA-sponsored data mining technologies would search disparate federal databases for transactions by persons – transactions such as applications for passports, visas, and airline ticket purchases – and attempt to link them with other events such as arrests or suspicious activities that, taken together, might indicate a terrorist act in the making. The Genisys program sponsored technologies to virtually aggregate data in support of effective analysis across heterogeneous databases and public sources. The Evidence Extraction and Link Discovery program enabled "connecting the dots" of suspicious activities beginning with a particular object such as a person, place or thing. Scalable Social Network Analysis sponsored technologies for separating terrorist groups from other groups using techniques of advanced social network analysis. (Taipale, 2003; DARPA, 2002; DARPA, 2003). In order to protect against abuses of the privacy of individually identified persons as well as protect sensitive data sources, the Genisys Privacy Protection Program intended to develop a "privacy appliance" for "providing critical data to analysts while controlling access to unauthorized information, enforcing laws and policies through software mechanisms, and ensuring that any misuse of data can be quickly detected and addressed." (DARPA, 2003, p. 6) The proposed privacy technology would also have sought to improve identity protection by limiting inference from aggregate sources. (DARPA, 2003) In other words, while data mining enables rapid identification of terrorist suspects and activities, the privacy technologies help prevent abuse of American law and regulations about individual citizen's privacy thus achieving security with privacy. (Taipale, 2003)

TIA intended to integrate the advanced data mining technologies with advanced collaboration tools sponsored under the Advanced Collaborative and Decision Support Programs and with advanced language translation and analysis tools sponsored under the Language Translation Programs. The advanced collaboration tools included sophisticated war-gaming and simulation capabilities as well as data sharing technology. The language translation tools were planned to enable rapid translation and preliminary analysis of foreign language materials from open and restricted source materials (DARPA, 2003). According to DARPA (2003), the integration of these powerful, computerized tools under TIA would support a series of

steps among distributed, collaborating experts attempting to discover the plans and intentions of potential terrorists, including:

1. Develop terrorist attack scenarios;
2. Initiate automated searches of data bases using terrorist attack scenarios and other intelligence information as starting points;
3. Identify individuals suspected of involvement in terrorist activities;
4. Identify associations among suspect individuals;
5. Link such associations with associations of other individuals;
6. Develop competing hypotheses about the plans of suspect associates in conjunction with other types of intelligence data;
7. Introduce the behavior and activities of suspect associates into models of known patterns of behavior and activity indicative of terrorist attack;
8. Generate range of plausible outcomes with options for action;
9. Analyze risks associated with each option for action;
10. Present complete analysis to decision-maker; and,
11. Record all steps of process in a corporate knowledge base for future review and use.

DARPA (2003, pg. 4) intended this entire process to yield four major benefits to the fight against counterterrorism, including:

1. Increase by an order-of-magnitude the information available for analysis by expanding access to and sharing of data;
2. Provide focused warnings within an hour after occurrence of a triggering event or passing of an articulated threshold;
3. Automatically cue analysts based on partial matches with patterns in a database containing at least 90 per cent of all known foreign terrorist attacks;
4. Support collaboration, analytical reasoning, and information sharing among analysts to hypothesize, test and propose theories and mitigating strategies about possible futures; and, thus
5. Enabling decision-makers effectively to evaluate the impact of current or future policies.

4.3 Controversy

Beginning in November 2002, columnists from major newspapers and magazines including *The Washington Post, The New York Times,* and *The National Review* as well as scholarly organizations publicly criticized the TIA program on multiple grounds. Concerns about governmental abuse of

personal information that sacrifices the privacy rights of individual American citizens in the name of national security in the post-9/11 world constituted the heart of their criticism. This specific controversy occurred in the context of an ongoing Congressional review of the privacy implications of multiple new security programs such as TIA and the Transportation Security Agency's Computer Assisted Passenger Prescreening System II (CAPPS II), to say nothing of HIPAA. Although other columnists and commentators attempted to defend TIA (Taipale, 2003; Taylor, 2002), the controversy produced an audit and an unfavorable report about the TIA program by the Inspector General of the Department of Defense. Congress ultimately withdrew funds for the TIA program (Office of the Inspector General, 2003). The controversy sheds important light on issues that any data mining project manipulating individually identifiable information must necessarily confront and manage well, especially projects that potentially harm or do not directly benefit the persons under surveillance.

Commentators argued that TIA posed multiple threats to the privacy of individual Americans, including:

1. Violates the Fourth Amendment of the Constitution by searching a data base containing detailed transaction information about all aspects of the lives of all Americans (Safire, 2002; Washington Post, 2002; Crews, 2002; Stanley and Steinhardt, 2003)
2. Undermines existing privacy controls embodied in the Code of Fair Information Practices, such as improper reuse of personal data collected for a specific purpose (Simons and Spafford, 2003; Safire, 2002; Crews, 2002)
3. Overcomes "privacy by obscurity" including inappropriate coordination of commercial and government surveillance (Safire, 2002; Washington Post, 2002; Stanley and Steinhardt, 2003)
4. Increases the risk of falsely identifying innocent people as terrorists (Crews, 2002; Simons and Spafford, 2003; Stanley and Steinhardt, 2003)
5. Increases the risk and cost of identity theft by collecting comprehensive archives of individually identifiable information in large, hard-to-protect archives (Simons and Spafford, 2003)
6. Accelerates development of the total surveillance society (Safire, 2002; Washington, 2002; Crews, 2002; Stanley and Steinhardt, 2003)

Other undesirable consequences in addition to invasion of privacy potentially flowed from TIA, including:

1. Undermining the trust necessary for the successful development of the information economy and electronic commerce (Crews, 2002; Simons and Spafford, 2003)
2. Undesirably altering the ordinary behavior of the American population including quelling healthy civil disobedience, "normalizing" terrorist behavior, and inhibiting lawful behavior (Crews, 2002; Simons and Spafford, 2003)
3. Creating new, rich targets for cyberterrorism and other forms of individual malicious abuse of computerized personal information (Crews, 2002; Simons and Spafford, 2003)

Some commentators also argued that TIA demonstrated important organizational shortcomings, including:

1. Poor choice of leadership with Admiral John Poindexter of Iran-Contra fame as program director (Safire, New York Times, November 14, 2002; Washington Post Editorial November 16, 2002)
2. Insufficient oversight (Safire, New York Times, November 14, 2002; Washington Post Editorial November 16, 2002; Simons and Spafford, January 2003)
3. Low likelihood of achieving its goal of "countering terrorism through prevention" (Crews, National Review November 26, 2002; Simons and Spafford, January 2003)

On December 12, 2003, the DOD Inspector General (DOD IG) issued a report on its audit of the TIA program entitled, "Terrorism Information Awareness Program (D-2004-033) (Office of the Inspector General, December 12, 2003). The DOD IG conducted the audit in response to questions from Senator Charles E. Grassley, Ranking Member of the Senate Finance Committee with supporting letters and questions from Senator Chuck Hagel and Senator Bill Nelson. The audit objectives included assessing "whether DARPA included the proper controls in developmental contracts for the TIA program that would ensure that the technology, when placed in operational environment, is properly managed and controlled." (DOD, Office of the Inspector General, 2003, p. 3). The audit focused particularly on DARPA's appreciation for the importance of protecting the privacy of individuals potentially subject to TIA surveillance. The DOD IG (2003, p. 4) summarized its conclusions as follows:

> "Although the DARPA development of TIA-type technologies could prove valuable in combating terrorism, DARPA could have better addressed the sensitivity of the technology to minimize the possibility for Governmental abuse of power and to help insure the successful transition of the technology into the operational environment."

While acknowledging the application of TIA-type technologies in foreign intelligence, the DOD IG expressed strong reservations about DARPA's inattention to the implications of TIA for potential governmental abuse in domestic intelligence and law enforcement purposes. The DOD IG particularly faulted DARPA program management for not having consulted experts in policy, privacy and legal matters to ensure successful transition to the operational environment. Four factors contributed to DARPA's inattention to these issues (DOD, Office of the Inspector General, 2003, pg. 4), including:

1. DARPA did not implement the best business practice of performing a privacy impact assessment (PIA);
2. Under Secretary of Defense for Acquisition, Technology and Logistics initially provided oversight of the TIA development and did not ensure that DARPA included in the effort the appropriate DOD policy, privacy and legal experts;
3. DARPA efforts historically focused on development of new technology rather than on the policies, procedures and legal implications associated with the operational use of technology; and,
4. The DARPA position was that planning for privacy in the operational environment was not its responsibility because TIA research and experiments used synthetic artificial data or information obtained through normal intelligence channels.

To have exercised due care, safeguarded taxpayers' money, and protected its program, DARPA should have taken several precautions, including:

1. Employed governmental best privacy practice by executing a PIA. In the words of the DOD IG (2003 p. 7), a PIA "consists of privacy training, gathering data on privacy issues, identifying and resolving the privacy risks, and approval by (the agency) privacy advocate";
2. Ensured adequate oversight by a responsible agency with experts in policy, privacy and legal matters;
3. Developed in advance policies and procedures as well as technology for protecting privacy; and,
4. Considered the ultimate use of the information in the operational environment not just the source of data used in research experiments.

Taking these precautions would have integrated privacy concerns into TIA's entire developmental and acquisition lifecycle instead of relegating that responsibility to end-users. DARPA could thus have avoided causing unnecessary alarm among the members of Congress and the American

public and, had the program continued, avoided wasting taxpayer's money on expensive retrofits or redesign of the TIA applications.

By the time the DOD IG released its report, Congress had terminated all funding for TIA and most of its component applications. Nonetheless, the DOD IG made two additional recommendations to guide development of future TIA-type programs. Before resuming TIA-type research, DARPA should take specific steps to integrate privacy management into its research and development management process, including:

1. Conduct Privacy Impact Assessments on potential research and development projects using models such as the PIA of the Internal Revenue Service, endorsed by the Federal Chief Information Officer's Council, as a best practice for evaluating privacy risks in information systems; and
2. Appoint a Privacy Ombudsman to oversee PIAs and thoroughly scrutinize TIA-type applications from a privacy perspective.

4.4 Lessons Learned from TIA's Experience for Medical Investigators Using "Datamining" Technologies

The TIA program imagined integrating many innovative technologies into an effort to preempt terrorist attacks by identifying and sharing information about suspicious activities among relevant Federal agencies. "Data mining" technologies of various types with the purpose of examining individually identifiable information in Federal and commercial databases constituted the program's core functionality. While not as comprehensive as TIA, medical data analysis and research employing datamining of patient records invites comparison as well as contrast with DARPA's counterterrorism research and development program. In particular, medical investigators should not take for granted the good will of their patients, their institutions or their funding agencies. Unlike the program management of DARPA and TIA, principle investigators must take personal responsibility for assuring proper identification and implementation of privacy controls, thorough training of their staff in privacy responsibilities and communication of their efforts to all relevant audiences. TIA teaches medical researchers some specific lessons when translated into the environment of healthcare research and data analysis:

1. Medical researchers should take full advantage of the privacy functions of the Institutional Review Board (IRB). From the perspective of the DOD IG's report on TIA, the IRB represents an oversight board that is fully equipped to advise and monitor

researchers on privacy policies, procedures and practices. In most academic research institutions, HIPAA has strengthened the IRB's awareness and competence to manage privacy issues.

2. Medical researchers should devote great care in preparing the privacy and security portions of their IRB forms, particularly the informed consent form. The IRB review forms can function for the individual research project like the Privacy Impact Assessment in Federal agencies in helping to identify and propose mitigation plans for project privacy risks. The informed consent form provides an ideal vehicle for explaining to patient-subjects a project's privacy protections.
3. Medical investigators should cultivate an effective relationship with the medical center's HIPAA Privacy and Security Officers. Like the privacy ombudsman in Federal agencies, the HIPAA Privacy and Security Officers function as points of articulation and communication when necessary between the researcher, the patient-subject, the institution, and external agencies such as the Office of Civil Rights, Department of Health and Human Services.
4. Medical investigators should consider the advisability of a project external advisory board when conducting research or using datamining methods that might provoke special privacy concerns. If properly composed and chartered, an external advisory board can provide useful expertise in policy, privacy and legal matters external to a medical researcher's own institution and lend extra credibility to a project's good faith efforts in the event of controversy.
5. Medical investigators should formally develop and document in writing privacy and security policies and procedures for the research project or its parent unit. As HIPAA and the DOD IG report emphasize, these policies and procedures must include administrative and physical as well as technical privacy and security controls. These written policies and procedures should inform the information about privacy protections included in the IRB and informed consent forms.

5. CONCLUSIONS AND DISCUSSION

A formal approach to managing the use and disclosure of personal health information is in the best interests of patients, individual researchers, organizations and society. The risks to those who do not adhere to good security and privacy practices are considerable. Future laws and regulations are likely to increase penalties for inappropriate use or disclosure. While much attention has been given to research, organizations should implement the same general processes to support analyses done for the purpose of healthcare operations as for research.

> "Researchers have no automatic right to review patient data. Besides developing strategies for minimizing patient risk, as described herein, investigators should take simple steps to characterized their compliance with human subjects requirements" (Berman, pg. 33, 2002).

A recent publication recommends:

> "First, sensitive raw data like identifiers, names, addresses and the like, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms, should also be excluded, because such a knowledge can equally well compromise data privacy, as we will indicate. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem." (Verykios et al, pg. 1, 2004).

While these are good recommendations, they are insufficient for medical data mining. As long as the original data is available, there is risk to confidentiality, integrity and availability of the data. Thus, an effective privacy program depends upon implementing robust security controls. Medical dataminers should be sure to employ several important security practices, including:

1. Mandatory oversight by a privacy board or institutional review board with approval for each project should be established.
2. The methods sections of research proposals and publication submissions should include a description of steps to minimize patient risks and that IRB approval has been obtained.
3. Good access control and authorization should be used for each session and query.
4. Where possible, the common identifiers (e.g. names, addresses) of the data subjects should be removed or hidden from the data user.
5. Robust audit practices should be instituted.
6. Training for all principle investigators that reinforces their responsibilities should be required.
7. Sanctions should be applied for violations of policy and/or procedures.
8. Trends in breaches and sanctions should be tracked and trended over time and used in the process of security awareness and training.

6. ACKNOWLEDGEMENTS: FUNDING SOURCES OR RESEARCH PARTNERS

Dr. Cooper and Dr. Collmann thank Kim Schwarz, Adam Robinson, Georganne Higgins and Jim Wilson for their assistance in various aspects of this paper. National Library of Medicine contract NO1-LM-3-3506, "Applications of advanced network infrastructure in health and disaster management: Project Sentinel Collaboratory" supported Dr. Collmann's work on this chapter. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the National Library of Medicine.

REFERENCES

Adam, N.R., Wortmann, J.C. (1989). "Security-control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys (CSUR)* 21(4) 515 - 556.

Alberts C, Doroffe A. (2003). *Managing Information Security Risks: The OCTAVEsm Approach.* Boston, MA, Addison-Wesley.

Behlen, F.M., Johnson, S.B. (1999). "Multicenter Patient Records Research: Security Policies and Tools," *J Am Med Inform Assoc*. 6(6) 435-43.

Berman, J.J. (2002). "Confidentiality Issues for Medical Data Miners," *Artif Intell Med.* 26(1-2):25-36.

California HealthCare Foundation (1999). Medical Privacy and Confidentiality Survey Summary and Overview, http://www.chcf.org/documents/ihealth/survey.pdf.

Cios, K.J., Moore, G.W. (2002). "Uniqueness of Medical Data Mining," *Artif Intell Med*. 26(1-2), 1-24.

Claerhout, B., De Moor, G.J., De Meyer, F. (2003). "Secure Communication and Management of Clinical and Genomic Data: The Use of Pseudonymisation as Privacy Enhancing Technique," *Stud Health Technol Inform*. 95:170-5.

Crews, Jr., C.W., November 26, 2002). "The Pentagon's Total Information Awareness Project: Americans Under the Microscope?", *Techknowledge*, Issue #45, originally in *National Review Online*, November 25, 2002.

Defense Advanced Research Project Agency (July 19, 2002). "Total Information Awareness Program (TIA) System Description Document (SDD)," Version 1.1.

Defense Advanced Research Project Agency (May 20, 2003). Information Awareness Office, "Report to Congress regarding the Terrorist Information Awareness Program: In response to Consolidated Appropriations Resolution, Pub.L. No. 108-7, Division M, § 111(b)", Detailed Information.

Department of Defense (December 12, 2003). Office of the Inspector General, Information Technology Management, "Terrorist Information Awareness Program" (D-2004-033).

Department of Health and Human Services (August 10, 2004). Office for Human Research Protections Guidance on Research Involving Coded Private Information or Biological Specimens, http:// www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf.

Department of Health and Human Services (July 13, 2004). Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule, (NIH Publication Number 03–5388), http://privacyruleandresearch.nih.gov/pr_02.asp.

Department of Health and Human Services (2002). Final Privacy Standard, Title 45 CFR Parts 160 and 164, http://www.hhs.gov/ocr/hipaa/privrulepd.

Department of Health and Human Services (2003). Final Security Standard, Title 45 CFR Parts 160, 162, and 164, www.cms.hhs.gov/hipaa/hipaa2/regulations/security/03-3877.pdf.

Department of Health and Human Services (2001). Human Subjects Regulations Common Rule Title 45 part 46, http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

Department of Health and Human Services (2001). Office for Human Research Protections, Code of Federal Regulations, Title 45, Part 46, Subpart A, 46.101 (b) (4); http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm#subparta.

Department of Health and Human Services (2004). Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule, (NIH Publication Number 03–5388), http://privacyruleandresearch.nih.gov/pr_02.asp .

Department of Health and Human Services (August 14, 2002). Office of the Secretary. 45 CFR Part 160, 162, and 164, Standards for Privacy of Individually Identifiable Health Information: Final Rule, Federal Register, Vol. 67, No. 157, 53181-53273.

Department of Health and Human Services (February 20, 2003). Office of the Secretary. 45 CFR Part 160, 162, and 164, Security Standards: Final Rule. Federal Register, Vol. 68, No. 34, 8333-8381.

Dicker, K.M. (2003). "The Evolution of Data Mining and Related Security Correlation Technology," *SANS Institute*, http://www.giac.org/practical/GSEC/Keith_Dickter_GSEC.pdf.

Federal Office of Management and Budget (1994). Statistical Policy Working Paper 22, *Report on Statistical Disclosure Limitation Methodology*, http://www.fcsm.gov/working-papers/wp22.html.

Ferris, T.A., Garrison, G.M., Lowe, H.J. (2002). "A Proposed Key Escrow System for Secure Patient Information Disclosure in Biomedical Research Databases," in *Proc AMIA Symp*. 245-9.

Food and Drug Administration (2002). Protection of Human Subjects Regulations Title 21 CFR parts 50 and 56, http://vm.cfsan.fda.gov/~lrd/cfr50.html.

Friedman, B., Kahn, JR., P.H. and Borning, A., et al. (Draft of June 2003). Value Sensitive Design: Theory and Methods, http://www.ischool.washington.edu/vsd/vsd-theory-methods-draft-june2003.pdf.

Galandiuk, S. (2004). Legislative Threat to Clinical Science: The Obfuscation and De-identification of Protected Health Information," *Br J Surg*. 91(3) 259-61

Goldman, J. and Hudson, Z. (2000). "Perspective Virtually Exposed: Privacy and E-Health," *Health Affairs,* 19(6), 140-8.

Goodwin, L.K. and Prather, J.C. (2002). "Protecting Patient Privacy in Clinical Data Mining," *J Healthc Inf Manag,* 16(4):62-7.

Health Privacy Project (2003). *Medical Privacy Stories*, http://www.healthprivacy.org/usr_doc/Privacy_storiesupd.pdf

International Information Security Foundation (1997). Generally-Accepted System Security Principles, http://web.mit.edu/security/www/GASSP/gassp021.html

Islan, M.Z., and Brankovic, L., A. (2004). "Framework for Privacy Preserving Classification in Data Mining, School of Electrical Engineering and Computer Science," *Australasian Computer Science Week*.

Levin, E.G., Arango, J., Steimle, A.E., Lee, P.C., Fireman, B. (2001). "Innovative Approach to Guidelines Implementation Is Associated with Declining Cardiovascular Mortality in a Population of Three Million [abstract]," in *American Heart Association's Scientific Se*ssions, Anaheim, California.

Lin, Z,. Hewett, M., Altman, R.B. (2002). "Using Binning to Maintain Confidentiality of

Medical Data," in *Proc AMIA Symp*. 454-8.
Lin, Z., Owen, A.B., Altman, R.B. (2004). "Genetics. Genomic Research and Human Subject Privacy," *Science*, 9:305(5681):183.
Lowrance, W. (2002). "Learning from Experience: Privacy and the Secondary Use of Data in Health Research," The Nuffield Trust; www.nuffield trust.org.uk
Malin B., Sweeney L. (2001). "Re-identification of DNA through an Automated Linkage Process," in *Proc AMIA Symp*. 423-7.
Malin, B., Sweeny, L., and Newton, E. (2003). "Trail Re-identification: Learning Who You Are from Where You Have Been," Carnegie Mellon University, School of Computer Science Data Privacy Laboratory, *Technical Report*, LIDAP-WP12 (Pittsburgh).
Meany, M.E. (2001). "Data Mining, Dataveillance, and Medical Information Privacy," in *Privacy in Health Care.* J, Humber, ed., Humana Press, pp. 145-164.
Melton, L.J. (1997). "The Threat to Medical-Records Research," *N Engl J Med*., 13;337(20) 1466-70.
Moore, G.W., Brown, L.A., Miller, R.E. (2001). "Gödelization of a Pathology Database: Re-Identification by Inference," *Johns Hopkins Autopsy Resource*, http://www.netautopsy.org
Moore, G.W., Brown, L.A., Miller, R.E. (2000). "Set Theory Definition and Algorithm for Medical De-identification," *Johns Hopkins Autopsy Resource*, http://www.netautopsy.org
Murphy, S.N., Chueh, H.C. (2002). "A Security Architecture for Query Tools Used to Access Large Biomedical Databases," in *Proc AMIA Symp*. 552-6.
National Committee for Quality Assurance (2002). *Annual Report*.
National Institute of Health (2004). HIPAA Privacy Rule, Frequently Asked Questions # 17; http://privacyruleandresearch.nih.gov/faq.asp#17
National Institute of Health (2004). HIPAA Privacy Rule, Clinical Research and the HIPAA Privacy Rule, http://privacyruleandresearch.nih.gov/clin_research.asp
Newton, E., Sweeney, L. and Malin, B. (2003). *Preserving Privacy by De-identifying Facial Images*, Carnegie Mellon University, School of Computer Science, *Technical Report*, CMU-CS-03-119 (Pittsburgh).
Oliveira, S.R.M., Zaïane, O.R. (2003). "Protecting Sensitive Knowledge by Data Sanitization," in *Proceedings of the Third IEEE International Conference on Data Mining,* Melbourne, Florida, USA, 613-616.
Pheatt, N., Brindis, R., Levin, E. (2003). "Putting Heart Disease Guidelines into Practice: Kaiser Permanente Leads the Way," *The Permanente Journal*, 7(1) 18-23, http://xnet.kp.org/permanentejournal/winter03/guides.html
Quantin, C., Bouzelat, H., Allaert, F.A., Benhamiche, A.M., Faivre, J., Dusserre, L. (1998). "Automatic Record Hash Coding and Linkage for Epidemiological Follow-up Data Confidentiality," *Methods Inf Med*, 37(3) 271-7.
Ruch, P., Baud, R. H ., Rassinoux A., Bouillon, P., Robert, G. (2000). "Medical Document Anonymization with a Semantic Lexicon," in *Proc AMIA Symp* 729-733.
Safire, W. (November 14, 2002). "You are a Suspect," *New York Times*.
Saul, M. (2004). "De-Identification Tool for Patient Records Used in Clinical Research," *Health Services Library System*, 9(3).
http://www.hsls.pitt.edu/about/news/hslsupdate/2004/june/iim_de_id/
Simons, B. Spafford, E.H. (2003). Co-chairs, US ACM Policy Committee, Association for Computing Machinery, Letter to Honorable John Warner, Chairman, Senate Committee on Armed Forces.
Stanley, J., Steinhardt, B., (January 2003). *Bigger Monster, Weaker Chains: The Growth of an American Surveillance Society*, American Civil Liberties Union, Technology and Liberty Program.
Sweeney, L. (1997). "Weaving Technology and Policy Together to Maintain Confidentiality,"

J Law Med Ethics, 25(2-3):98-110, 82.
Sweeney, L. (1997). "Guaranteeing Anonymity When Sharing Medical Data, The Datafly System," in *Proc AMIA Symp* 51-55.
Sweeney, L. (2002). "K-anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems,* 10(7) 557-570.
Sweeney, L. (2003). "Navigating Computer Science Research through Waves of Privacy Concerns: Discussions among Computer Scientists at Carnegie Mellon University," *ACM Computers and Society*, 34(1):1-18.
Sweeney, L. (1996). "Replacing Personally-Identifying Information in Medical Records, The Scrub System," in *Proc. AMIA,* 333-337.
Taipale, K.A. (2003). "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *The Columbia Science and Technology Law Review*, Vol. V, 5-83, http://www.stlr.org/cite.cgi?volume=5&article=2
Taylor, S., (December 2002). "Big Brother and Another Overblown Privacy Scare," *Atlantic Online*
Thomas, S.M., Mamlin, B., Schadow, G., McDonald, C. (2002). "A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method," in *Proc AMIA Symp*. 777-81.
Tzelepi, S., Pangalos, G. and Nikolacopoulou, G. (2002). "Security of Medical Multimedia," *Med. Inform,.* 27(3):169–184.
UCLA DataServer - An open source xml data gateway, UCLA medical imaging informatics, http://www.mii.ucla.edu/dataserver/docs/features/deidentification.html
Verykios, V.S., et al. (2004). "State-of-the-art in Privacy Preserving Data Mining," *SIGMOD Record,* 33(1):1-8.
Washington Post (November 16, 2002). "Total Information Awareness," Saturday.
Wiederhold, G., Bilello, M. (1998). "Protecting Inappropriate Release of Data from Realistic Databases," in DEXA *'98 Workshop on Security and Integrity of Data Intensive Applications*, http://www-db.stanford.edu/pub/gio/TIHI/DEXAgio.html
Wiederhold, G., Bilello, M., Sarathy, V., Qian, X. (1996). "A Security Mediator for Health Care Information," in *Proc AMIA Symp*. 120-4.
Wiederhold, G. (2002). "Future of Security and Privacy in Medical Information," *Stud Health Technol Inform*, 80:213-29.
Wylie J.E., and Mineau, G.P. (2003). "Biomedical Databases: Protecting Privacy and Promoting Research," Trends Biotechnol, 21(3):113-6.

SUGGESTED READINGS

Department of Defense, Office of the Inspector General, Information Technology Management, "Terrorist Information Awareness Program (D-2004-033), December 12, 2003.
The DOD IG's report describes core institutional issues in protecting privacy in data mining.

Department of Health and Human Services, *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule*, (NIH Publication Number 03–5388).
The NIH Guide to privacy considerations in human subject research provides a good introduction to these issues.

Berman, J.J., "Confidentiality Issues for Medical Data Miners." *Artif Intell Med.*, 26(1-2):25-

36 (2002).
This article describes some of the innovative computational remedies that will permit researchers to conduct research and share their data without risk to patient or institution.

Sweeney, L., "Navigating Computer Science Research Through Waves of Privacy Concerns: Discussions among Computer Scientists at Carnegie Mellon University." *ACM Computers and Society*. 34 (1) (2003).
This article introduces the nature of privacy concerns in computer science research and explains the potential benefits and risks.

ONLINE RESOURCES

Asia

Asia-Pacific Privacy Charter Initiative
http://www.bakercyberlawcentre.org/appcc/announce.htm

Australia

Federal Privacy Law
http://www.privacy.gov.au/act/

Data Matching
http://www.privacy.gov.au/act/datamatching/index.html

Complaint Case Notes and Determinations
http://www.privacy.gov.au/act/casenotes/index.html

Canada

Canada's Health Information Infostructure
http://www.privcom.gc.ca/information/02_03_02_e.asp#002

Personal Information Protection and Electronic Documents Act
http://www.medicalpost.com/mpcontent/article.jsp?content=20031211_144135_3716

Department of Health and Human Services

Office of Civil Rights - HIPAA Security and Privacy
http://www.hhs.gov/ocr/hipaa/

HIPAA Privacy Rule and Public Health Guidance from CDC and the U.S. Department of Health and Human Services
http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm

Common rule: Title 45 Part 46 Protection of Human Subjects
http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

HIPAA and research
http://privacyruleandresearch.nih.gov/

European Union

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
http://www.cdt.org/privacy/eudirective/EU_Directive_.html

Briefing Materials on the European Union Directive on Data Protection
http://www.cdt.org/privacy/eudirective/

US Department of Commerce – Safe Harbor
http://www.export.gov/safeharbor/

Health Information and Systems Society

CPRI Toolkit: Managing Information Security in Healthcare,
http://www.himss.org/asp/cpritoolkit_toolkit.asp

Institute of Medicine

For the Record: Protecting Electronic Health Information
http://books.nap.edu/html/for/

Health services research: Protecting Data Privacy in Health Services Research
http://www.nap.edu/books/0309071879/html/

Institutional Review Boards and Health Services Research Data Privacy: A Workshop Summary
http://books.nap.edu/books/NI000228/html/

National Institutes of Health

Research Repositories, Databases, and the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/research_repositories.asp

Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/pr_02.asp

Clinical Research and the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/clin_research.asp

Institutional Review Boards and the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/irb_default.asp

Privacy Boards and the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/privacy_boards_hipaa_privacy_rule.asp

Security Research Centers

Carnegie Mellon University, Privacy Technology Center
http://center.privacy.cs.cmu.edu/index.html

Johns Hopkins Autopsy Resource
Http://www.netautopsy.org

Purdue University, Center for Education and Research in Information Assurance and Security
http://www.cerias.purdue.edu/

Stanford University Security Laboratory
http://theory.stanford.edu/seclab/index.html

AT&T Labs- Research
http://www.research.att.com

IBM Privacy Research Institute
http://www.research.ibm.com/privacy

Microsoft Research
http://research.microsoft.com

QUESTIONS FOR DISCUSSION

1. How should medical investigators address the components of security (confidentiality, integrity, availability and accountability) for a new project?

2. What types of oversight should medical investigators establish when planning data mining projects on patient data.

3. What are the trade-offs that should be considered in developing a risk management plan for a medical data mining project?

4. Describe what a medical investigator must do to respect the rights granted to patients by the HIPAA Privacy Standard and the requirements of the Common Rule.

5. How is data mining used to enhance security and brainstorm potential avenues of research in this area?