Patient Confidentiality in the Research Use of Clinical Medical Databases

Rajeev Krishna, PhD, Kelly Kelleher, MD, MPH, and Eric Stahlberg, PhD

Electronic medical record keeping has led to increased interest in analyzing historical patient data to improve care delivery. Such research use of patient data, however, raises concerns about confidentiality and institutional liability. Institutional review boards must balance patient data security with a researcher's ability to explore potentially important clinical relationships.

We considered the issues involved when patient records from health care institutions are used in medical research. We also explored current regulations on patient confidentiality, the need for identifying information in research, and the effectiveness of deidentification and data security. We will present an algorithm for researchers to use to think about the data security needs of their research, and we will introduce a vocabulary for documenting these techniques in proposals and publications. (*Am J Public Health.* 2007;97: 654–658. doi:10.2105/AJPH.2006.090902)

THE RAPID DIGITIZATION OF medical records and administrative databases coupled with advances in statistics and computing capabilities promise to make epidemiological studies for improving health care more fruitful than ever. Modern computing power provides quantitative researchers with numerous new techniques for exploring and identifying correlations in large data warehouses.[1] Common to such efforts is the need for access to large quantities of potentially sensitive patient health information (protected health information, such as names, record numbers, addresses, and so on).

Interest in maintaining—and legal sanctions for violating—patient confidentiality are of particular concern to researchers who use medical data (administrative, diagnostic, etc.) in analytic studies. Balancing the conflicting interests of ensuring patient confidentiality with providing access to sufficiently detailed information for adequate research is a serious challenge to health care organizations and data providers and their respective institutional review boards (IRBs). Although existing legal restrictions in the United States attempt to strike such a balance, no computing system is entirely secure, and there is understandable concern about unintended or inappropriate releases of information.

Fortunately, there are numerous concepts and techniques from the domains of data security and statistical disclosure limitation that may be brought to bear on this problem. Application of these techniques allows tradeoffs between data usability and data security, giving researchers access to relevant data while at the same time minimizing the potential damage of a breech in data security. We reviewed privacy issues surrounding the use of electronic data collected in routine medical care, and we considered advanced approaches to minimizing potential privacy violations when data is used for medical research. Because of the complexity of this problem and the array of techniques available for improving data security, we did not delve into specific technologies or security algorithms. Rather, this discussion is intended to be an introduction for researchers and their human participant oversight structures and a starting point for conversations with information service departments about the best security solutions for a given situation.

DEFICIENCIES OF CURRENT REGULATION

In the United States, current regulations on the use of protected health information for research purposes under the Health Insurance Portability and

Accountability Act (HIPAA) divide medical record sets into 3 categories: identified data, deidentified data, and limited data. *Identified data* include any data that could be used by a recipient to uniquely identify the person from an individual patient record. Access to such data requires explicit consent by study participants or a waiver of the consent requirement by an IRB. Furthermore, use of identified data incurs numerous restrictions that primarily involve the tracking of protected health information disclosures. By contrast, *deidentified data* is data with all such identity information removed (HIPAA provides a specific list of 18 data elements that must minimally be removed), and this data may be used freely.

Sets containing *limited data* are available only to research, public health, and health care organizations. Unlike the other categories of data sets, limited data sets attempt to provide high-quality (of sufficient detail as to be useful for research purposes) and accessible (able to be acquired and used) data for research, public health, and other health care–related tasks. Through a limited data set, researchers may access data elements, such as date and geographic information, without some of the restrictions for using fully identified data.

Considerable research in privacy-preserving data mining,[2,3] disclosure risk assessment[4,5] and data deidentification, obfuscation, and protection[6,7] can be found in computing and database management literature and is often directly applicable to these medical privacy issues. More generally, groups such as government organizations commonly encounter confidentiality issues in the release of statistical data, resulting in extensive discussions about disclosure limitation techniques.[8–12] There is little evidence in the medical literature, however, to suggest that researchers exploit this flexibility in practice; instead, they depend on explicit removal of specific identifying data elements (deidentification) or the physical security of the data infrastructure (with an identified data set).

The problem with this tendency toward use of either identified or deidentified data is complex. On the one hand, it gives decisionmakers the impression that deidentified data is inherently safe for public consumption. The open accessibility of large demographic databases across a variety of topics, however, may invalidate this assumption. For example, students at the Massachusetts Institute of Technology were able to re-identify 35% of the records in a 30-year span of the Chicago homicide victims database by correlating data elements with records in the Social Security Death Index, even though both sets were public and were considered to be deidentified.[13] Thus, the goal of deidentification may not be upheld when multiple deidentified data sets are available.

On the other hand, it is easy to believe that the data security restrictions on the use of identified data sets will ensure confidentiality. Unfortunately, the risk of inadvertent disclosure rises with the number of authorized users and with the number of duplicate data sets required, regardless of the perceived level of security at each access point. A single individual who writes down a password could compromise an entire data infrastructure. Indeed, the recent string of data security breeches (e.g., lost and stolen tape backups and laptops, and credit card and online banking database intrusions) shows the vulnerability of presumptively secure systems.

Owing to the increasing legal and ethical implications with the use of medical record data, perhaps the greatest concern is that little effort is applied to the documentation of data security efforts when the results of an analysis are published. For example, a brief review of the 2005 editions of the *American Journal of Public Health* revealed at least 35 Research and Practice articles that used potentially protected health information (not including studies that used publicly available government data). Of these, only 1 article clearly identified the security measures employed. The majority (n=21) either indicated IRB approval or exemption or explained why IRB approval was not sought (usually because deidentified data was used). Thirteen articles simply stated that IRB approval was not necessary. Because of the potential for disclosure even with deidentified data, this lack of documentation is itself a concern. It is understandable, however, because there is a lack of common vocabulary for succinctly describing such efforts. It is unfortunate because an expectation of such disclosure on the part of publications could substantially improve the practice of data security as a whole. The remainder of this discussion will provide a framework for thinking about maximizing data security. We also will introduce a vocabulary for describing data security efforts.

METHODS OF DATA SECURITY

Although establishing the confidentiality of a given piece of data can often be challenging, the concept of disclosure control is founded on a number of fairly straightforward principles and techniques. Literature in the statistical disclosure control domain generally divides this task into limitation of access (eliminating certain data elements from view) and statistical approaches (modifying or structuring the data to destroy uniquely identifiable characteristics).[8,9] Because we are discussing disclosure control as it pertains to research needs, and to facilitate communication with institutional information services departments and data providers, we have adopted a framework that draws heavily from the computer science domain. Thus, in this section, we introduce a vocabulary of methods for reducing the identifiability of data. The section “Maximizing Data Security in Research” will use this framework to present a high-level algorithmic approach to acquiring useful research data in a form that minimizes the damage of unintended disclosure.

Data Exclusion

Exclusion of specific data elements is the basis of most general restrictions on data use. In this realm, carefully constructed aggregate data or removal of entire records provide the highest level of confidentiality. Second to this is individual record deidentification in which specific identifying fields (such as those specified by HIPAA) are removed. The goal is to verify that the deidentification process maximizes data a particular researcher needs while ensuring sufficient commonality between records for anonymity.

A number of existing systems can aid in this. For example, the concept of *K*-anonymity[14] and the use of systems such as Datafly[15] (Data Privacy Lab, Carnegie Mellon University, Pittsburgh, Pa) ensure that at least *k* records in any given data set are indistinguishable along any parameter of interest. Field masking can maintain specific aspects of the data set that are of research interest. Along a similar vein, Concept Match[16] (National Cancer Institute, National Institutes of Health, Rockville, Md) provides a system for deidentifying free text fields by removing words that do not match a predetermined set of interest words for a domain. The resulting anonymous text consists of standard medical terms and connector words, with most of its research usefulness intact. Regardless of the methodology, however, data exclusion invariably destroys aspects of the original data that may be useful when making inferences or conclusions.

Data Transformation

Moving one step from the absolute confidentiality provided by exclusion of data, we found a number of data transformation techniques that provide a statistical guarantee of confidentiality. Most secure among these are methods of data transformation. The common theme in these techniques is to make an irreversible modification to the data that destroys the original values or correlations (this method is termed *lossy*, because some information is irrecoverably lost in the process) while preserving the relationships of interest. As with data exclusion, techniques exist to modify data globally (as aggregation does for data exclusion) or at the level of individual elements.

Data perturbation is an example of global data transformation. The idea is to preserve aggregate trends in the original data while removing or altering the actual data. For example, data may be intelligently swapped between records, preserving the overall set of values in a field but eliminating the precise mapping between fields of a given record, or random "noise" may be added to the data, maintaining the statistical properties of a field while randomly altering exact values in any given record within some threshold amount. Bakken et al. present a more rigorous exploration of such techniques and many of the other concepts discussed in this section.[7]

Hashing of individual data elements involves a lossy 1-way transformation or mapping of data. A simple hash of 20 unique zip code values (protected under HIPAA) may randomly replace each unique zip code with a value between 1 and 1000 at each entry in the data set. This transformation probabilistically maintains the uniqueness of zip code values and thus preserves much of the research value. However, the finite probability of a "collision"—2 zip codes mapping to the same new value—greatly complicates confident recovery of original values by reversing the transformation. Many standard hashing algorithms exist, including the Message-Digest Algorithm 5 (MD5),[17]developed at MIT, and the Secure Hash Algorithm1 (SHA1) developed by the National Institute for Standards and Technology.[18]

Data Encryption

A further step from absolute confidentiality leads to reversible data transformations, such as data element encryption. The idea of encryption is to take input data (plaintext) and output new data (cyphertext) from which the original cannot be practically recovered without the use of specialized information external to the encrypted data (the key). A simple example would be to create a 1-to-1 mapping of a replacement letter for each letter in the alphabet or a code. With the mappings in hand, recovering plaintext from cyphertext is a simple matter. Without the mapping, the problem becomes far more complicated. A key point here is that the strength of an encryption scheme is often measured on the *impracticality* of inappropriate recovery, not the *impossibility.* The ability to break the code is tied to the quantity of data available under the same key, the quality of the key, and the encryption algorithm itself. Details about cryptographic techniques have been published elsewhere.[19]

A good cryptographic technique will hide all relationships between the original text and cyphertext. Although valuable to protecting privacy, this creates a problem for researchers, particularly in situations of semifree text fields. Consider an analysis of health trends by employer (also a restricted field under HIPAA). The employer name may be encrypted with the understanding that identical names will lead to identical cyphertext, allowing comparison of potential employer effects without access to the actual name. Unfortunately, if the name is entered as free text, small variations in the entry (e.g., Wendys vs Wendy's) could lead to substantial variations in the cyphertext, making it impossible to use the field data in an analysis with any degree of confidence. Fixing these variations in letters used (syntax) for words with the same meaning is a process called *normalization.*

Cryptographic technique also carries some lessons for use and dissemination of protected data. Perhaps the most important lesson is that good encryption is not a substitute for good data access security. Techniques like encryption can, at best, provide an added safeguard by increasing the level of sophistication necessary on the part of an intruder and thereby decreasing the practicality of attempting a breech. Given time, nearly every reversible cryptographic technique can be compromised.

Another important lesson from cryptography is the value

of variability in data. The first instinct of many institutions when constructing research data sets is to establish 1 uniform deidentified data set for all researchers to access as needed. Although this is the easiest and sometimes the only practical solution, it also increases the risk of exposure. Consider again the employer example. One could imagine using knowledge of major employers in the area and an understanding of patient demographics to begin recovering employer information from the full list and thereby begin breaking the coding scheme used to protect the information. This danger is compounded if individual researchers maintain local duplicates of some or all of the master data set for their work, and the danger is further compounded if some of those duplicates are of identified data because of the needs of a particular project.

This risk can be reduced by a number of simple steps. Ideally, individual data sets should be constructed for each research effort, providing only the subset of records relevant to that effort. Furthermore, each data set should be encoded independently. Although the actual algorithm may be the same, unique keys should be used in the encryption or coding of data for each research effort, and the ordering of individual records should be randomized whenever possible. This ensures that data from one research project cannot be compared against data from another, reduces the potential of a security breech, and limits the damage should a breech occur.

Data Obfuscation

In the context of this discussion, we use the term *data obfuscation* to denote any approach to masking data that is weaker than cryptography and is employed primarily to preserve relationships within a data set that would be destroyed by more rigorous masking techniques. It should be noted that the term is often used more generically in the literature, although it generally relates to the tradeoff between anonymity and usability of data.

Practical use of such techniques may be most evident in interrelated numeric data, such as dates or addresses. For example, epidemiology researchers may be interested in accessing highly specific location data to correlate health patterns with neighborhoods, cities, or regions. However, finding clusters of poor health outcomes does not require knowledge of actual patient addresses. It merely requires relationships between patient addresses. Thus, data extraction for the study may translate addresses into some other metric that preserves relative locations without revealing the actual physical location. Although this complicates recovery of the original information, it does not provide the level of structured security that encryption or hashing systems do. In this example, sufficient quantities of data and a general knowledge of population trends may allow an intruder to approximate the original locations with relative ease. However, data obfuscation is not intended to eliminate the need for data access security; it simply increases the complexity of recovery and reduces the pool of would-be intruders.

MAXIMIZING DATA SECURITY IN RESEARCH

The previous section provides a good foundation for discussing the confidentiality of medical data used in research studies. The actual techniques used will depend on both the needs of an individual research effort and careful consultations with institutional information services departments and data providers. Initially, it will take time and effort on the part of data managers and researchers; however, a set of standard, reusable practices should develop in short order, making the process very straightforward. Such standardization also will facilitate communication of the security infrastructure to IRBs. The next 5 paragraphs provide an approach for guiding these discussions.

What Data Is Needed?

The obvious first step in any data extraction is careful specification of the data requirement. This is standard practice in most research efforts, and consideration should be give of what records are necessary (i.e., only records that meet certain criteria) and what fields of a record are necessary (if patient names, license numbers, and so on, are not needed, or if free text fields and images will not be evaluated, they should not be provided). Furthermore, as research progresses, access to any subsets of data deemed unnecessary upon inspection also should be removed. This provides a cleaner operating environment for research and minimizes damage should a security breech occur.

What Data Can Be Encrypted?

Any relevant relationships discovered in transformed or obscured data would be useless without the ability to recover original values. In the general case, this will require that at least 1 field be masked in a recoverable fashion. This field provides a reference by which the original record may be discovered if necessary. This recoverable field should be selected to continue to maximize patient confidentiality in the event of unauthorized access. For example, an encrypted study-specific patient identifier would be more appropriate than an encrypted social security number. Should an intruder recover the encrypted data, exploiting the information would still require breaking the security of the main database to gain access to the full record.

What Data Should Be Transformed or Obfuscated?

Researchers should now determine confidentiality and the level of acceptable data loss for each field in the desired records. Those fields that only require aggregate properties or probabilistic uniqueness should be masked by lossy transformation techniques, leaving fields with confidentiality concerns addressed but also with important relationships preserved. It may be that any attempt to obscure this information would excessively complicate the study. However, some effort should be dedicated to considering how such data elements may be obfuscated without destroying relevant relationships.

Establishing the Confidentiality of Remaining Data

It is clearly impractical and often detrimental to mask or obfuscate every field in a data set. Thus, after systematically hiding identifiable data, we are still left with a number of fields in their original form. As a final step in the construction of a research data set, it may be valuable to assess, if not further manipulate or eliminate, any remaining unique records. Application of techniques such as *K*-anonymity can ensure that, although each record may be uniquely identifiable by use of obscured fields, no record will provide a starting point for breaking obfuscation techniques by standing out as unique in the unobscured fields.

Physical Data Security and Auditing

The steps discussed thus far may provide some confidence that an unintended data release will not lead to significant exposure of protected health information, but this is not a reason to be casual about the security of the data itself. Although not explored in depth in this discussion, it is important to recognize that the best defense is good physical data security. To that end, standard data security practices should be used to ensure that the data remain in a secure access-restricted storage area and that separate credentials (usernames and passwords) are given to each authorized user. This not only prevents unauthorized access but also provides an audit trail should an incident occur. Recognizing that security breeches are often a product of social engineering (breaking system security by manipulating legitimate users—for example, by claiming that your password is not working and asking a legitimate user to log you in) rather than the hacking of the physical security infrastructure, a short training session on basic data security (protecting passwords, locking workstations when not in use, and so on) may be warranted for study staff.

CONCLUSIONS

We have presented a brief overview of data security techniques and the application of these techniques to medical research databases. Data security is of particular relevance with the proliferation of electronic medical and administrative records and the ease with which such data can be exported outside of the secure institutional infrastructure. We have introduced a vocabulary for discussing these issues and have introduced an approach that researchers, information services departments, and IRB committees can use to begin applying security techniques. Indeed, coordination among these groups and the incorporation of security considerations into IRB and journal approval procedures are the keys to ensuring continued patient protection in an increasingly digital and interconnected world. ■

About the Authors

Rajeev Krishna and Kelly Kelleher are with the Columbus Children's Research Institute, Columbus, Ohio, and the Ohio State University Medical Center, Columbus. Eric Stahlberg is with the Ohio Supercomputer Center, Columbus.

Requests for reprints should be sent to Kelly Kelleher, MD, MPH, Center for Innovation in Pediatric Practice, 700 Children's Dr, Columbus, OH 43205 (e-mail: kellehek@ccri.net).

This article was accepted July 11, 2006.

Contributors

K. Kelleher and R. Krishna originated and designed this article. All authors drafted, revised, and reviewed the article.

Acknowledgments

This work was funded by the Samuel J. Roessler Memorial Medical Scholarship at the Ohio State University College of Medicine.

Human Participant Protection

No human subjects were used in this work, and IRB approval was not required.

References

1. Castellani B, Castellani J. Data mining: qualitative analysis with health informatics data. *Qual Health Res.* 2003;13:1005–1018.

2. Agrawal R, Srikant R. Privacy-preserving data mining. In: Proceedings of 2000 ACM SIGMOD Conference on Management of Data; May 16–18, 2000; Dallas, Tex.

3. Verykios V, Bertino E, Fovino I, Provenza L, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record.* 2004;33:50–57.

4. Steel P. Disclosure risk assessment for microdata. Available at: http://www.census.org/srd/sdc/steel.disclosure\%20risk\%20assessment\%20for\%20microdata.pdf. Accessed June 2005.

5. Domingo-Ferrer J, Torra V. Disclosure risk assessment in statistical data protection. *J Computational Appl Math.* 2004;164:285–293.

6. Sweeney L. *Computational Disclosure Control: A Primer on Data Privacy Protection* [PhD thesis]. Cambridge, Mass: Massachusetts Institute of Technology; 2001.

7. Bakken DE, Rarameswaran R, Blough DM, Franz AA, Palmer TJ. Data obfuscation: anonymity and desensitization of usable data sets. *IEEE Secur Privacy.* 2004;2:34–41.

8. Gonzalez M. *Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22.* Washington, DC: Office of Management and Budget; 1994.

9. Willenborg L, de Waal T. *Statistical Disclosure Control in Practice.* New York, NY: Springer-Verlag New York Inc; 1996.

10. Helmpecht B, Schackis D. *Manual on Disclosure Control Methods.* Luxembourg, Belgium: Office for Official Publications of the European Communities; 1993.

11. Domingo-Ferrer J, Mateo-Sanz J. Current directions in statistical data protection. *Res Official Stat.* 1998;2:105–112.

12. MacNeil D, Pursey S. Disclosure control methods in the public release of microdata files of small business. Available at: http://www.amstat.org/sections/srms/proceedings/papers/1999_044.pdf. Accessed December 22, 2006.

13. Ochoa S, Rasmussen J, Robson C, Salib M. Re-identification of individuals in Chicago's homicide database: a technical and legal study. Available at: http://citeseer.ist.psu.edu/ochoa01reidentification.html. Accessed June 2005.

14. Sweeney L. K-anonymity: a model for protecting privacy. *Int J Uncertainty, Fuzziness Knowledge-Based Syst.* 2002;10:557–570.

15. Sweeney L. Datafly: a system for providing anonymity in medical data. In: Lin TY, Qian S, eds. *Database Security XI: Status and Prospects.* New York, NY: Chapman & Hall; 1998:356–381.

16. Berman J. Concept–match medical data scrubbing: how pathology text can be used in research. *Arch Pathol Lab Med.* 2003;127:680–686.

17. Rivest R. The MD5 message digest algorithm. Available at: http://www.faqs.org/rfcs/rfc1321.html. Accessed December 22, 2006.

18. Schneier B. *Applied Cryptography: Protocols, Algorithms, and Source Code in C.* 2nd ed. New York, NY: John Wiley & Sons; 1995.