Linking Registry Data: Technical and Legal Considerations
Draft White Paper for AHRQ Patient Registries Handbook II

Authors:

Stephen E. Fienberg, PhD, FRSC (lead author)
Carnegie Mellon University

Sara Rosenbaum, JD (lead author)
George Washington University

Susan Adams, JD (lead author)
Dartmouth College

Alan F. Karr, PhD
National Institute of Statistical Sciences

Bradley Malin, PhD
Vanderbilt University

Deven McGraw, JD, LLM
The Center for Democracy and Technology (CDT)

Maya A. Bernstein, JD
Office of the Assistant Secretary for Planning and Evaluation, DHHS

Melissa M. Goldstein, JD
George Washington University

Joy Pritts, JD
Georgetown University

Andy DeMayo, JD
Outcome

EXECUTIVE SUMMARY

The purpose of this paper is to identify important considerations and provide guidance to researchers and research sponsors who are interested in linking data held in a health information registry with additional data, such as from claims or other administrative files, or with another registry. There are two equally weighted and important sets of questions to address in the planning process: what is a feasible technical approach to linking the data, and

whether the linkage is legally feasible under the permissions, terms, and conditions that applied to the original compilations of each dataset. Legal feasibility depends on two factors: 1) the applicability of federal and state legal protections on health information and participation in human research to the specific purpose of the data linkage, and 2) any specific permission obtained from individual patients for the use of their health information.

There is a significant amount of analysis needed in the planning for such projects, as the technical approach may be influenced by permitted uses of the data under applicable regulations and the legal assessment may change depending on how the linkage needs to be performed and the nature and purpose of the resulting linked data set. The intention of the paper is to help investigators find an appropriate way to address their critical research questions, remain faithful to the conditions under which the data was originally collected, and protect individual patients by safeguarding their identities and maintaining the confidentiality of the data under applicable law.

The paper presumes the investigators have identified an explicit purpose for the data linkage in the form of a scientific question they are trying to answer. The nature of this objective is critical to assessing the applicable regulatory requirements for uses of the data. Investigators should assign the goal of the data linkage project to one of the following categories:

- Health care operations as defined by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, which includes health care quality-related activities,
- Public health practice,
- FDA requirements
- Research, or
- Some combination of these purposes.

If research is one purpose of the project, then the Common Rule (regulations for human subjects protection) is likely to apply to the project.

The application of the HIPAA Privacy Rule depends on the origins of the data sets being linked, and such origins may also influence the feasibility of making the data linkage. Investigators should know the source for the original data, the conditions under which it was compiled, and what kinds of permissions apply to it, from both individual patients and the custodial institutions. Health information is most often data that have two sources: individual and institutional; these sources may have legal rights and continuing interests in the use of the data.

This paper has eight sections divided into three major parts: Technical Aspects of Data Linkage Projects, Legal Aspects of Data Linkage Projects, and Risk Mitigation for Data Linkage Projects. First, sections 1, 2, 3, and 4 discuss the reasons and technical methods for linking datasets containing health information, including data held in registries. Section 1 reviews the general goals of health information registries and the scientific potential of linking diverse sources of data to discover relationships that might improve health care. This section also notes ethical

concerns about the use of health information other than in the clinical setting. As a result of these concerns, a tension exists between the goals of advancing biomedical knowledge and legal protections for the confidentiality of health information. The principal technical focus of this paper is to explain the approaches and formal methodologies that help mediate this tension.

Section 2 defines important concepts, including the different definitions of "disclosure" as used by statisticians and in the HIPAA Privacy Rule. Sections 3 and 4 discuss in some detail the probabilistic matching of common data elements to link datasets, as well as the procedural issues that are involved. These sections describe both the standard approach to linking datasets used by statisticians and some of the techniques used to address recognized problems.

In the second part of the paper -- addressing legal aspects of data linkage projects -- section 5 introduces the risks of identifying individuals that dataset linkage may entail and discusses the protection conferred by de-identification. The relative risk levels for identifying data elements plus the special problems associated with biospecimens and genotyping are also briefly described. Section 6 explains the legal standards of the HIPAA Privacy Rule, with a focus on how the regulation may protect individuals from these risks of harm. Such explanation emphasizes the importance of establishing a clear goal for a proposed dataset linkage project in determining the legal standards applicable the project. Unless the datasets for linkage are de-identified as defined by the HIPAA Privacy Rule, the regulatory standards impose specific conditions on permitted uses of health information. The relation of the HIPAA Privacy Rule to the HIPAA Security Rule is also noted.

Finally, sections 7 and 8 summarize both recognized and developing technical methods for mitigating the risks of re-identification for the individuals whose data resides in linked datasets. These methods, which are discussed in section 7, all contain a trade-off between the risk of re-identification and the utility of the resulting data. Privacy-preserving computation methodologies are briefly analyzed and the application of cryptographic approaches to privacy protection is introduced. Section 8 explores the role of security practices in preserving the confidentiality of health information in linked datasets. Two basic strategies for secure data maintenance exist: one strategy involves restrictions on information; the other involves restrictions on access. The consequences of these restrictions for projects using confidential data are noted.

At the end of the paper are two appendices. Appendix I consists of a hypothetical data linkage project scenario that is intended to provide context for the technical and legal information presented in the paper. Appendix II is a list of questions for the consideration of data linkage project leaders during the project-planning phase. The questions are intended to assist in organizing the resources needed to implement the project, including the statistical, regulatory, and collegial advice that might prove helpful in navigating the complexities of data linkage projects that are presented in the paper.

Not every relevant topic could be addressed in this paper. Some limitations of the discussion are the exclusion of the following issues: 1) considerations about linking data from public and private sectors, where different ethical and legal restrictions may apply, and 2) detailed information about the risks involved with identifying the health care providers that collect and provide data.

Dataset linkage entails the risks of loss of reliable confidential data management and identification or re-identification of individuals and institutions. Recognized and developing statistical methods and secure computation may limit these risks and may allow the public health benefits that registries linked to other datasets have the potential to contribute.

KEY POINTS

- Two equally weighted and important sets of questions must be addressed in the data linkage planning process: what is a feasible technical approach to linking the data, and is linkage legally feasible under the permissions, terms, and conditions that applied to the original compilations of each data set.
- Many statistical techniques (e.g., deterministic matching, probabilistic matching) for linking records exist; the choice of a technique should be guided by the types of data that are available.
- Linkage projects should include plans for managing common issues (e.g., records that only exist in one database, variations in units of measure).
- Linkage of de-identified data may result in accidental re-identification. Risks of re-identification vary depending on the variables used and should be managed with guidance from legal and statistical experts to minimize risk and ensure compliance with HIPAA, the Common Rule, and other legal and regulatory requirements.

INTRODUCTION

The purpose of this paper is to identify important considerations and provide guidance to researchers and research sponsors who are interested in linking data held in a health information registry with additional data, such as from claims or other administrative files, or another registry. The paper is designed to help investigators find an appropriate way to address their critical research questions, remain faithful to the conditions under which the data was originally collected, and protect individual patients by safeguarding their privacy and maintaining the confidentiality of the data under applicable law.

There are two equally weighted and important sets of questions to address in the planning process: what is a feasible technical approach to linking the data, and whether the linkage is legally feasible under the permissions, terms, and conditions that applied to the original compilations of each data set. Legal feasibility depends on the applicability of federal and state

legal protections for the confidentiality of health information and participation in human research to the specific purpose of the data linkage, as well as any specific permissions obtained from individual patients for the use of their health information. Indeed, there is a great deal of analysis and planning for such projects, as the technical approach may be influenced by permitted uses of the data under applicable regulations and the legal assessment may change depending on how the linkage needs to be performed and the nature and purpose of the resulting linked data set. The following discussion presumes the investigators have identified an explicit purpose for the data linkage in the form of a scientific question they are trying to answer. The nature of this objective is critical to assessing the applicable regulatory requirements for uses of the data. Investigators should assign the goal of the data linkage project to one of the following categories: health care operations as defined by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, which includes health care quality-related activities, public health practice, research, or some combination of these purposes. If research is one purpose of the project, then the Common Rule (human subjects protection regulations) is likely to apply to the project.

The application of the HIPAA Privacy and Security Rules depends on the origins of the data sets being linked, and such origins may also influence the feasibility of making the data linkage. Investigators should know the source for the original data, the conditions under which it was compiled, and what kinds of permissions apply to it, from both individual patients and the custodial institutions. Health information is most often data that have two sources: individual and institutional; these sources may have legal rights and continuing interests in the use of the data.

It is important to be aware that the legal requirements may not remain stable and the protections for the research use of health information is likely to change in response to continued development of electronic health information technologies.

This paper has eight sections divided into three major parts Technical Aspects of Data Linkage Projects, Legal Aspects of Data Linkage Projects, and Risk Mitigation for Data Linkage Projects. First, sections 1, 2, 3, and 4 discuss the reasons and technical methods for linking datasets containing health information, including data held in registries. Section 2 defines important concepts, including the different definitions of “disclosure” as used by statisticians and in the HIPAA Privacy Rule. Secondly, section 5 discusses the risks of identification for individuals inherent in data linkage projects and section 6 describes the legal standards of the HIPAA Privacy Rule that pertain to these risks. Finally, sections 7 and 8 summarize both recognized and developing technical methods for mitigating the risks of identification.

At the end of the paper are two appendices. Appendix I consists of a hypothetical data linkage project that is intended to provide context for the technical and legal information presented in the paper. Appendix II comprises a list of questions for the consideration of data linkage project leaders during the planning of a project. The questions are intended to assist in organizing the resources needed to implement the project, including the statistical, regulatory, and collegial

advice that might prove helpful in navigating the complexities of data linkage projects that are presented in the paper.

TECHNICAL ASPECTS OF DATA LINKAGE PROJECTS

1. Linking Records for Research and Improving Public Health

Data about the health of individuals included in registries come in a wide variety of forms, most of which have been gathered originally for the delivery of clinical services or payment for those services, and under promises or legal guarantees of confidentiality, privacy, and security. These sources of data may include individual doctor's records, billing information, vital statistics on births and deaths, health surveys, and data associated with biospecimens, among other sources.

The broad goals of registries are to amass data from potentially diverse sources to allow researchers to explore and evaluate alternative health outcomes in a systematic fashion. These goals are usually accomplished by gathering data from multiple sources and linking the data across sources either with explicit identifiers designed for linking or in a probabilistic fashion via the characteristics of the individuals to whom the data correspond. From the research perspective, the more data included the better, both in the number of cases and in the details and the extent of the health information. The richer the database, the more likely it is that data analysts will be able to discover relationships that might improve health care. Notably, many discussions about privacy protection focus on limiting the level of detail available in data to which others have access.

There is an ethical obligation to protect patient interests when collecting, sharing, and studying person-specific biomedical information.[1] Many people fear that information derived from their medical or biological records will be used against them in employment decisions, limit access to health or life insurance, or cause social stigma.[2] These fears are not unfounded, and various cases have transpired in which it was found that an individual's genetic characteristics or clinical manifestations were used in a manner inconsistent with an individual's expectations of privacy and fair use.[3] If individuals fear that their health-related information could be associated with, or used against, them, they may be less likely to seek treatment in a clinical context or participate in research studies.[4]

A tension exists between the broad goals of registries and regulations protecting individually identifiable information. Approaches and formal methodologies that help mediate this tension are the principal technical focus of this paper. To understand the extent to which these tools can assist data linkages involving registry data one needs to understand the risks of identification in different types of data.

A large body of federal law applies to privacy. A recent comprehensive review of privacy law and its effects on biomedical research identified no fewer than 15 separate federal laws pertaining to health information privacy.[5] Special federal laws also apply to health information

related to substance abuse.[6] A full review of all laws related to privacy, confidentiality, and security of health information also would consider separate state privacy protections, as well as state laws pertaining to the confidentiality of data. Nevertheless, the legal aspects of this paper only focus on the federal regulations commonly referred to as the HIPAA Privacy Rule.

2. What Do Privacy, Disclosure, and Confidentiality Mean?

Privacy is a term that has definition that varies with context.[7] As used in the HIPAA Privacy Rule, the term applies to protected health information (PHI). In this context, the concept of privacy relates to permitted uses and disclosures of individually identifiable health information. The HIPAA Privacy Rule addresses to whom the custodian of PHI, a "covered entity," may transmit it and under what conditions. The HIPAA Privacy Rule establishes three basic concepts of health information: data that are identifiable; data that lack certain direct identifiers, a "limited data set"; and data that are de-identified. Registries commonly acquire identifiable data and may create the last two categories of data. Along this spectrum of data, the HIPAA Privacy Rule applies different legal standards and protections.[8]

Disclosure has two different meanings: one is technical and the other is a HIPAA Privacy Rule definition.

2A. Technical definition

Disclosure relates to the attribution of information to the source of the data, regardless of whether the data source is an individual or organization. There are basically three types of disclosure of data that possess the capacity to make the identity of particular individuals known: identity, attribute, and inferential.

Identity disclosure occurs when the data source becomes known from the data release itself.[9] [10]

Attribute disclosure occurs when the released data make it possible to infer the characteristics of an individual data source more accurately than would have otherwise been possible.[11] [12] The usual way to achieve attribute disclosure is through identity disclosure. First, one identifies an individual through some combination of variables and then learns about the values of additional variables included in the released data. Attribute disclosure may occur, however, without identity disclosure, such as when all people from a population subgroup share a characteristic and this quantity becomes known for any individual in the subgroup.

Inferential disclosure relates to the probability of identifying a particular attribute of a data source. Because almost any data release can be expected to increase the likelihood of an attribute being associated with a data source, the only way to guarantee protection is to release no data at all. It is for this reason that researchers use certain methods not to prevent disclosure, but to limit or control the nature of the disclosure. These methods are known as *disclosure limitation* methods or *statistical disclosure control.*[13]

2B. HIPAA Privacy Rule definition

Disclosure means the release, transfer, provision of, access to, or divulging in any other manner of information outside of the entity holding the information.[14]

Confidentiality refers broadly to a quality or condition of protection accorded to statistical information as an obligation not to permit the transfer of that information to an unauthorized party.[15] Confidentiality can be owed to both individuals and health care organizations. A different notion of confidentiality arises from the special relationship between a clinician and patient. It relates to the ethical, legal, and professional obligation of those who receive information in the context of a clinical relationship to respect the privacy interests of their patients. Most often, the term is used in the former sense and not in the latter, but these two meanings inevitably overlap in a discussion of health information as data. The methods for disclosure limitation described here have been developed largely in the context of confidentiality protection as defined by laws, regulations, and especially by the practices of statistical agencies.

3. Linking Records and Probabilistic Matching

Computer-assisted record linkage goes back to the 1950s and was put on a firm statistical foundation by Fellegi and Sunter.[16] Most common techniques for record linkage either rely on the existence of unique identifiers or utilize a structure similar to the one Fellegi and Sunter described with the incorporation of formal statistical modeling and methods as well as new and efficient computational tools.[17, 18] The simplest way to match records from separate databases is to use a so-called deterministic method of linking the databases employing unique identifiers contained in each record. These identifiers might be names or social security numbers in the U.S., but these particular identifiers may not, in fact, be unique. As a result, some form of probabilistic approach is typically used to match the records. Thus, there is little distinction between methods using deterministic linkage as opposed to probabilistic linkage, except for the explicit representation of uncertainty in the matching process in the latter.

The now-standard approach is built on five key components for identifying matching pairs of records across two databases:[19]

1. Represent every pair of records using a vector of features (variables) that describe similarity between individual record fields. Features can be Boolean, discrete, or continuous.

2. Place feature vectors for record pairs into three classes: matches (M), non-matches (U), and possible matches *(P).* These correspond to “equivalent”, “nonequivalent,” and “possibly equivalent” (e.g., requiring human review) record pairs, respectively.

3. Perform record-pair classification by calculating the ratio $(P(\gamma \mid M)) / (P(\gamma \mid U))$ for each candidate record pair, where γ is a feature vector for the pair and $P(\gamma \mid M)$ and $P(\gamma \mid U)$ are the probabilities of observing that feature vector for a matched and non-matched pair, respectively. Two thresholds based on desired error levels—T_μ and T_λ—optimally separate the ratio values for equivalent, possibly equivalent, and nonequivalent record pairs.

4. When no training data in the form of duplicate and non-duplicate record pairs is available, matching can be unsupervised; that is, conditional probabilities for feature values are estimated using observed frequencies in the records to be linked.

5. Most record pairs are clearly non-matches, so one need not consider them for matching. This situation is managed by "blocking," or partitioning the databases, for example, based on geography or some other variable in both databases, so that only records in comparable blocks are compared. Such a strategy significantly improves efficiency.

The first four components lay the groundwork for accuracy of record-pair matching using statistical or machine learning prediction models such as logistic regression. The fewer near identifiers used in steps 1 and 2, the poorer the match is likely to be. Accuracy is well known to be high when there is a 1–1 match between records in the two databases and deteriorates as the overlap between the files decreases and the measurement error in the feature values increases. The fifth component provides for efficiently processing large databases, but to the extent that blocking is approximate and possibly inaccurate, its use decreases the accuracy of record-pair matching. The less accurate the matching, the more "error" there will be in the merged registry files. This error will impede quality analyses and findings from the resulting data.[20, 21]

This standard approach has problems when there are lists or files with little overlap, when there are undetected duplications within files, and when one needs to link three or more lists. In the latter case, one essentially matches all lists in pairs, and then resolves discrepancies. Unfortunately, there is no unique way to do this.

Record linkage methodology has been widely used by statistical agencies, especially in the U.S. Bureau of the Census. The methodology has been combined with disclosure limitation techniques, such as the addition of "noise" to variables in order to produce public use files that the agencies believe cannot necessarily be linked back to the original databases used for the record linkage. Another technique involves protecting individual databases by stripping out identifiers and then attempting record linkage. This procedure has two disadvantages. First, the quality of matches is likely to decrease markedly and the resulting merged records will still need to be protected by some form of disclosure limitation. Therefore, as long as there are not legal restrictions against the use of identifiers for record linkage purposes, it is preferable to use detailed identifiers to the extent possible and to remove them following matching procedure.

Currently there are not any special features of registry data that are known to enhance or inhibit matching. Registry data may be easier targets for re-identification because the specifics of diseases or conditions help to define the registries. In the United States, efforts are often made to match records using social security numbers. There are large numbers of entry errors for these numbers in many databases, as well as the problems associated with multiple people using one number and some people using multiple numbers.[22] Lyons et al. describe a very large-scale matching exercise in the United Kingdom linking multiple health care and social

services datasets using National Health Service numbers and various alternative sets of matching variables in the spirit of the record linkage methods described above. They report achieving accurate matching at rates of only about 95 percent. [23]

4. Procedural Issues in Linking Datasets

It is important to understand that neither "data" nor "link" can be defined unambiguously. For instance, a dataset may be altered by the application of tools for statistical disclosure limitation, in which case it is no longer the same dataset. Linkage need not mean, as it is customarily construed, "bring(ing) the two (or more) datasets together on a single computer." Many analyses of interest can be performed using technologies that do not require literal integration of the datasets. Even the relationship between datasets can vary. Two datasets can hold the same attributes for different individuals ("horizontal partitioning"), different attributes for the same individuals ("vertical partitioning"), or a complex combination of the two.

The process of linking horizontally partitioned datasets engenders little incremental risk of re-identification. There is, in almost all cases, no more information about a record on the combined dataset than was present in the individual dataset containing it. Moreover, any analysis requiring only data summaries (in technical terms, sufficient statistics) that are additive across the datasets can be performed using tools based on the computer science concept of secure summation.[24] Examples of analyses for which this approach works include creation of contingency tables, linear regression, and some forms of maximum likelihood estimation.

Only in a few cases have comparable techniques for vertically partitioned data been well enough understood to be employed in practice.[25] Instead, it is necessary to actually link individual subjects' records that are contained in two or more datasets. This process is inherently and unavoidably risky because the combined dataset contains more information about each subject than either of the components.

Discussed below is a preferred approach that is complex, but attenuates or even obviates other problems. Suppose that each of the two datasets to be linked contains the same unique identifiers (for individuals, an example is social security numbers) in all of the records. In this case, there exist techniques based on cryptography (homomorphic encryption[26] and hash functions) that enable secure determination of which individuals are common to both datasets and assignment of unique but uninformative identifiers to the shared records. Each dataset can then be purged of individual identifiers and altered to further limit re-identification, following which error-free and risk-free linkage can be performed.

Such techniques are computationally very complex, and may involve trusted third parties, who do not, however, have access to information in either dataset other than the common identifier. Therefore, in many cases the database custodian may prefer to remove identifiers and carry out statistical disclosure limitation prior to linkage. It is important to understand that this preferred approach compromises, perhaps irrevocably, the linkage process, and may introduce errors into the linked dataset that later—perhaps dramatically—alters the results of statistical analyses.

Many techniques for record linkage depend at some level on the presence of sets of attributes in both databases that are unique to individuals but do not lead to re-identification. For instance, the combination of date-of-birth, gender, and zip code of residence might be present in both databases. It is estimated that this combination of attributes uniquely characterizes a significant portion of U.S. citizens, so re-identification would only require access to a suitable external database. Other techniques such as the Fellegi-Sunter record linkage methods described above are more probabilistic in nature. They can be effective, but as we noted, they also introduce data quality effects that cannot readily be characterized.

No matter how linkage is performed, a number of other issues should be addressed. For instance, comparable attributes should be expressed in the same units of measure in both datasets, e.g., English or metric values for weight. Conflicting values of attributes for each individual common to both databases need reconciliation. Another issue involves the management of records that appear in only one database; the most common decision is to drop them. Data quality provides another example. It is one of the least understood statistical problems and has multiple manifestations.[27] Even assuming some limited capability to characterize data quality, the relationship between the quality of the linked dataset and the quality of each component should be considered. The linkage itself can produce quality degradation. The best way to address these issues is not clear, and intuition can be faulty. For example, there is reason to believe that the quality of a linked dataset is strictly less than that of either component and not, as might be supposed, somewhere between the two.

Finally, it is important to understand that there exist endemic risks to data linkage. Anyone with access to one of the original datasets and the linked dataset may learn, even if imperfectly, the values of attributes in the other. It may not be possible to determine what knowledge the linkage will create without actually doing the linkage. For these reasons, strong consideration should be given to forms of data protection such as licensing and restricted access in research data centers where both analyses and results can be controlled.

LEGAL ASPECTS OF DATA LINKAGE PROJECTS

5. Risks of Identification

The HIPAA Privacy Rule defines de-identification as the removal of names, addresses, telephone numbers, and other direct personal identifiers from health information; however, this removal process alone may not be sufficient. Residual data that are especially vulnerable to disclosure threats include (i) geographic detail, (ii) longitudinal information, and (iii) extreme values (e.g., income). Population health data are clearly more vulnerable than sample data, and variables that are available in other accessible databases pose special risks.

Statistical organizations such as the National Center for Health Statistics have traditionally focused on the issue of identity disclosure and thus refuse to report information in which individuals or institutions can be identified. This situation occurs, for example, when a data

source is unique in the population for the characteristics under study and is directly identifiable in the database to be released. But, such uniqueness and subsequent identity disclosure may not reveal any information other than the association of the source with the data collected in the study. In this sense, identity disclosure may only be a technical violation of a promise of confidentiality. Thus, uniqueness only raises the issue of possible confidentiality problems resulting from identification. A separate issue is whether the release of information is one that is permitted by the HIPAA Privacy Rule or is authorized by the data source.

The foregoing discussion implicitly introduces the notion of *harm*, which is not the same as a breach of confidentiality. For example, it is possible for a pledge of confidentiality to be technically violated, but produce no harm to the data source because the information is "generally known" to the public at large. In this case, some would argue that additional data protection is not required. Conversely, if one attempts to match records from one file to another file subject to a pledge of confidentiality and makes an "incorrect" match, there is no breach of confidentiality, but there is the possibility of harm if the match is assumed to be correct. Further, information on individuals or organizations in a release of sample statistical data may well increase the information about characteristics of individuals or organizations not in the sample. This information may produce an inferential disclosure for such individuals or organizations and cause them harm, even though there was no confidentiality obligation.

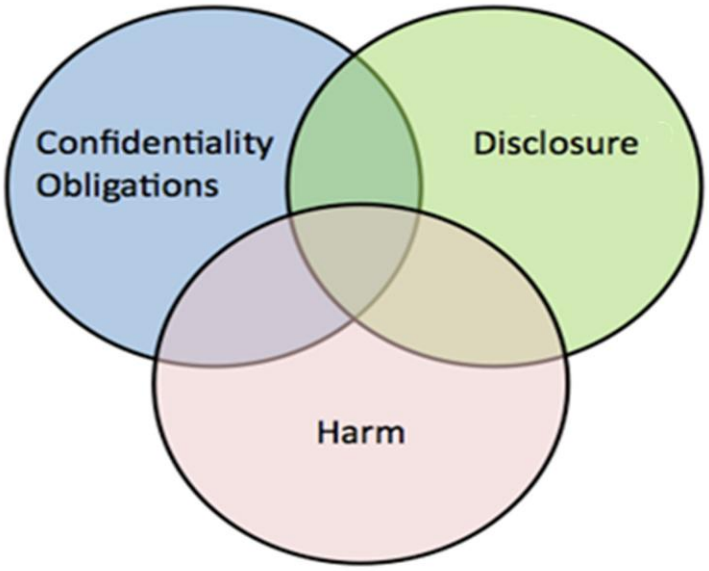

Figure 1. Relationships among confidentiality, disclosure, and harm

Some people believe that the way to assure confidentiality and prevent identity disclosure is to arrange for individuals to participate in a study anonymously. In many circumstances, such a belief is misguided, because there is a key distinction between collecting information

anonymously and ensuring that personal identifiers are not inappropriately made available. Moreover, we simply do not collect clinical health care data anonymously. Not only do patient records come with multiple identifiers crucial to ensuring patient safety for clinical care, but they also contain other information that may allow the identification of the patient, even if direct identifiers are stripped from the records.

Health or medical related data may also come from sample surveys in which the participants will have been promised that their data will not be released in ways that would allow them to be individually identified. Disclosure of such data can produce substantial harm to the personal reputations or financial interests of the participants, their families, and others with whom they have personal relationships. For example, in the pilot surveys for the National Household Seroprevalence Survey, the National Center for Health Statistics moved to make responses during the data collection phase of the study anonymous because of the harm that could potentially result from information that an individual had an HIV virus infection or engaged in high-risk behavior. But, such efforts still could not guarantee that one could not identify a participant in the survey database. This example also raises an interesting question about the confidentiality of registry data after an individual's death, in part because of the potential for harm to others. The health information of decedents is subject to the HIPAA Privacy Rule and several statistical agencies explicitly treat the identification of a deceased individual as a violation of their confidentiality obligations.

5a. Examples of how patients could be re-identified even if names are suppressed

For years, the confidentiality of health information has been protected through a process of "de-identification." This protection entails the removal of person-specific features, such as names, residential street addresses, phone numbers, and social security numbers. However, as discussed above, de-identification does not guarantee that individuals may not be identified from the resulting data. On multiple occasions, it has been shown that de-identified health information can be "re-identified" to a particular patient without hacking or breaking into a private health information system. For instance, in the mid-1990's Latanya Sweeney, then a graduate student at the Massachusetts Institute of Technology, showed that de-identified hospital discharge records, which were made publicly available at the state level, could be linked to identifiable public records in the form of voter registration lists. Her demonstration received notoriety, because it led to the re-identification of the medical status of the then-governor in the Commonwealth of Massachusetts.[28] This result was achieved by linking the data resources on their common fields of patient's date of birth, gender, and zip code. Various estimates indicate that the uniqueness of this combination of attributes in the U.S. population is somewhere between 65 and 87 percent, with certain subpopulations even more unique.[29]

5b: High-risk identifiers

One response to the Sweeney demonstration was the HIPAA Privacy Rule de-identification standard, which requires the removal of explicit identifiers (e.g., names), dates, geocodes (for populations of less than 20,000 inhabitants), and other data elements that in combination could

be used to ascertain an individual's identity. In all, the de-identification standard enumerates eighteen features that should be removed from patient information prior to data sharing.

Nonetheless, even the removal of these data elements may fail to prevent re-identification. In many instances, there are residual features that can lead to identification. The extent to which residual features can be used for re-identification depends on the availability of relevant data fields. Thus, one can roughly partition identifiers into "high" and relatively "low" risk features. The high-risk features are the sort that are documented in multiple environments and are publicly available. These are features that could be exploited by any recipient of such records. For instance, patient demographics are high-risk identifiers. Even de-identified health information permitted under the HIPAA Privacy Rule may leave certain individuals in a unique status and thus at high risk for identification through public data resources that contain similar features, such as public records containing birth, death, marriage, voter registration and property assessment information.

5c: Relatively low risk identifiers

In contrast, features that are lower risk do not appear in public records and are less available. For instance, clinical features, such as an individual's diagnosis and treatments, are relatively static because they are often mapped to standard codes for billing purposes. These features might appear in de-identified information, such as hospital discharge databases, as well as in identified resources such as electronic medical records. While combinations of diagnostic and treatment codes might uniquely describe an individual patient in a population, the identifiable records are available to a much smaller group than the general public. Moreover, this select group is ordinarily considered to be trustworthy, such as the clinicians and business associates of the custodial organization for the records, because they owe independent ethical, professional, and legal duties of confidentiality to their patients.

5d: Special issues with linkages to biospecimens

Medicine is increasingly moving towards evidence-based and personalized health care systems. In support of this trend, there is a growing focus on associations between clinical and biological phenomena. In particular, the decreasing cost of genome sequencing technology has facilitated a rapid growth in the volume of biospecimens and derived DNA sequence data. As much of this research is sponsored through federal funding, it is subject to data sharing requirements. However, biospecimens and DNA, in particular, are inherently unique and there are a number of routes by which DNA information can be identified to an individual.[30] For instance, there are over a million single nucleotide polymorphisms (SNPs) in the human genome; these little snippets of DNA are often used to make genetic correlations with clinical conditions. Yet, it is estimated that fewer than one hundred SNPs can uniquely represent an individual.[31] Thus, if de-identified biological information is tied to sensitive clinical information, it may provide a match to the identified biological information itself, e.g., in a forensic setting.[32]

Biospecimens and derivative information are of particular concern because they can convey knowledge not only about the individual from whom they are derived, but also other related

individuals. For instance, it is possible to derive estimates about the DNA sequence of relatives.[33] If the genetic information is predictive or diagnostic, it can adversely affect the ability of family members to obtain insurance and employment, or it may cause social stigmatization.[34] The Genetic Information Nondiscrimination Act of 2008 (GINA) prohibits health insurers from using genetic information about individuals or their family members, whether collected intentionally or incidentally, in determining eligibility and coverage, or in underwriting and premium setting. Insurers may, however, in collaboration with external research entities, request that policy holders undergo genetic testing, but a refusal to do so cannot be permitted to affect the premium or result in medical underwriting.[35]

6. The HIPAA Privacy Rule

The HIPAA Privacy Rule implements the Health Insurance Portability and Accountability Act of 1996,[36] whose purposes were to simplify the use of electronic health insurance claims and to improve the stability of and ability to maintain health insurance coverage as individuals changed employment. The United States Department of Health and Human Services (HHS) promulgated the HIPAA Privacy Rule to implement the statute's administrative simplification provisions. The HIPAA Privacy Rule expresses the value that Americans place on privacy, reflecting a strongly shared belief in personal autonomy, individuality, respect, and dignity.[37] Privacy is viewed as valuable not only to individuals but to society as well, and respect for privacy is conceived of as a social good. [38]

Certain experts have recommended strengthening the HIPAA Privacy Rule[39] and Congress took steps in this direction with passage of the American Recovery and Reinvestment Act of 2009 (ARRA), whose implementation, as of the date of this paper, is only in its earliest stages.[40]

The HIPAA Privacy Rule is recognized by some as a "landmark in health privacy protection" and a regulatory embodiment of "fair information practices."[41] The concept of fair information practice originated with a 1973 federal advisory committee report released by the United States Department of Health Education and Welfare. This broad concept exhibits a range of dimensions, such as limitations on information collection, data quality standards, purpose specification, use limitations, security safeguards, openness, individual participation, and accountability.[42] The federal Privacy Act of 1974[43] represented the first major foray by the federal government into the regulation of health information held by the federal government and certain contractors. The HIPAA Privacy Rule reaches a far broader group and regulates the use and disclosure of individually identifiable health information.

The HIPAA Privacy Rule creates a legal framework for health information that is designed to accommodate the need for information exchange with protections for privacy. In doing so, the HIPAA Privacy Rule creates a federal regulatory "floor" of legal protections, [44] while preserving "more stringent" state laws directed at health information. In general, only certain uses and disclosures of protected health information (PHI), defined as individually identifiable health information with certain limited exceptions, are permitted by the HIPAA Privacy Rule without patient authorization.[45] The HIPAA Privacy Rule preserves other federal laws and thus must be

considered in relation to separate and more protective federal privacy standards, such as special rules applicable to information about substance abuse.[46, 47]

The HIPAA Privacy Rule applies to "covered entities": health care providers, health care insurance plans, and health care clearinghouses that transmit insurance claims electronically.[48] The HIPAA Privacy Rule affords protections to individually identifiable health information held by covered entities and their "business associates," which are entities that perform work on behalf of covered entities involving the use of PHI. With very limited exceptions, the HIPAA Privacy Rule does not distinguish among types of health information, although psychotherapy notes receive greater protections.[49]

In general, the HIPAA Privacy Rule permits a covered entity to use and disclose PHI without an individual's written permission for certain core purposes that include treatment, payment, and health care operations. The HIPAA Privacy Rule defines the term "health care operations" to include the quality assurance and improvement activities conducted by health care providers. As a result, registries containing PHI that are created solely to conduct quality activities may be operated by covered entities or their business associates without the knowledge or authorization of patients. ARRA redefined health care operations to exclude certain types of communications related to the sale of products or services, which, as of the date of this paper, is not expected to significantly affect research or quality registries.[50]

The HIPAA Privacy Rule regulations permit covered entities to establish privacy policies that provide for consent by patients to the use and disclosure of their PHI, even if the policies cover the core purposes that these regulations exempt from such requirements. In other words, the HIPAA Privacy Rule permits disclosure of PHI for treatment purposes, but health care providers may be more stringent. They may elect to share patient information with another treating provider without getting written authorization or alternatively, to follow their own practice of obtaining specific consent.

For payment purposes and health care operations, the HIPAA Privacy Rule uses a "minimum necessary" standard to measure the appropriate use and disclosure of PHI. As a result, the HIPAA Privacy Rule limits uses and disclosures of PHI by covered entities to the minimum amount necessary to accomplish the intended purpose of the use or disclosure.[51] In contrast, providers do not have to limit disclosures to the minimum necessary for treatment purposes, but they are subject to safeguards within their organization restricting access to PHI only to those who appropriately rely on it to perform their duties.[52]

Covered entities also may disclose PHI without patient authorization for health care oversight, public health, disaster relief, and law enforcement activities; when required by other laws; and research, if certain regulatory criteria are satisfied.[53] Outside of treatment, payment, health care operations, and explicitly permitted disclosures, the HIPAA Privacy Rule requires written authorizations from individual patients for the use or disclosure of PHI and specifies the content and format of such authorizations. For the use or disclosure of PHI for research purposes, an

institutional review board (IRB) or Privacy Board may grant a waiver of authorization or approve the alteration of an authorization. Requests for a waiver or alteration of authorization are subject to specific regulatory criteria, which require that the research involve no more than minimal risk to the privacy of individuals.[54] As noted above, if the Common Rule applies to the research activities, then its separate regulatory criteria for waiver or alternation of consent to research participation will also apply.

For routine disclosures of protected health information to a third party who performs business and certain other functions on behalf of a covered entity, the third party "business associate" must enter into a formal agreement with the covered entity. A registry operator that received PHI in its performance of services to a covered entity may be a business associate, especially if the registry supports quality activities and analyses for the health care provider. Business associate agreements include assurances that the business associate will appropriately safeguard the PHI that it receives.[55]

ARRA broadens the definition of "business associate" to include organizations that provide data transmission of PHI to covered entities and routinely have access to PHI. Examples of organizations so affected by ARRA are Regional Health Information Organizations (RHIO), Health Information Exchanges (HIE), e-prescribing gateways, and vendors contracted by a covered entity to offer a personal health record to patients as part an electronic health record system.[56] In addition, ARRA expands the scope of the HIPAA Privacy Rule regulations and penalties for violations of the HIPAA Privacy Rule to business associates that obtain or create protected health information under a written contract executed in accordance with the terms of the HIPAA Privacy Rule.[57]

ARRA also expands the level and range of civil monetary penalties that may be enforced by the U.S. Department of Health and Human Services by creating a four-tiered system of increasingly severe penalties for a various categories of violations that are not timely corrected.[58] Covered entities and business associates weigh the risks of such penalties, even as the HIPAA Privacy Rule did not create a private right of action for individuals to recover damages for injuries arising from improper uses or disclosures of their PHI.

The HIPAA Privacy Rule addresses, but does not apply to, de-identified health information, which is no longer considered PHI. The standard for de-identification under the HIPAA Privacy Rule is when the health information "does not identify an individual and if the covered entity has no reasonable basis to believe that the information can be used to identify an individual".[59] The HIPAA Privacy Rule provides two methods to achieve de-identification. The first is a "safe harbor" that involves the removal of specified data elements plus any other unique identifying number, characteristic, or code. The second method is certification by an individual with appropriate scientific and statistical knowledge that there is a very small risk that information is identifiable.[60] The safe harbor has been criticized as providing insufficient protections, [61] and some experts view re-identification as relatively easy to achieve.[62, 63] Recent advances in

technology may provide additional ways to assure de-identification while preserving the utility of data.[64, 65]

A re-identification standard for use by covered entities is also specified in the HIPAA Privacy Rule. The standard allows a covered entity to assign a code or other means of record identification to permit de-identified information to be re-identified by the covered entity. The code may not be derived from or related to information about an individual data source and may not be otherwise capable of being translated to identify the data source. The covered entity is prohibited from using or disclosing the code for any other purpose and from disclosing the algorithm for re-identification.[66] By applying a re-identification code to disclosed de-identified health information, a covered entity can provide data verification and validation services for the disclosed data.

The HIPAA Privacy Rule also provides a standard for partially de-identified data as a "limited data set"[67] that may be used and disclosed for research, as well as public health purposes and health care operations, without the need to obtain individual patient authorization. The use or disclosure of the limited data set for the latter purposes does not require review by a Privacy Board or an institutional review board (IRB), provided there is no research component to the project. Use or disclosure of a limited data set involves a data use agreement between the data recipient and the covered entity holding the PHI that specifies the purpose for the limited data set (i.e., for public health, research, or health care operations), identifies who may use it, contains safeguards for subsequent use or disclosure of the data, and contains a notice that the data recipient may not identify the limited data set or contact the data subjects.[68]

The creation of a limited data set removes certain specified direct identifiers. At the same time, other data elements may be retained, such as city, state, zip code, and most importantly for research purposes, dates, and a code linked to the direct identifiers.[69] A covered entity may create a limited data set or arrange for a third party business associate to do so.[70] The covered entity or its business associate may serve as an "honest broker" for a data linkage project, providing limited data sets for the linkage plus data verification and validation services as its health care operations, and without being directly involved in the research project. Under these circumstances, the confidentiality of the health information is preserved through a data use agreement with the researcher limiting the purposes for which the data may be used and the absence of direct identifiers in the analytical dataset.

The HIPAA Privacy Rule is closely tied to the related HIPAA Security Rule, in that both regulations promote standards for reasonable safeguards to protect the privacy of PHI. The HIPAA Security Rule standard requires the implementation of appropriate administrative, technical, and physical safeguards.[71] The HIPAA Security Rule is designed to achieve several major goals: (1) to ensure that only authorized individuals use electronic PHI (ePHI); (2) that ePHI is used by authorized individuals only for an authorized purpose; and (3) that the ePHI is accurate and available.[72, 73] Although the HIPAA Security Rule sets forth requirements related to the protection of ePHI, many covered entities have implemented safeguards for all PHI—whether

electronic or otherwise. Prior to ARRA, the Security Rule applied only to covered entities. The new legislation extends the reach of the Security Rule to business associates,[74] while also imposing new obligations in the case of a security breach and increasing penalties for violations. ARRA creates duties for covered entities, their business associates, and certain vendors of personal health records to provide certain notices in the event of data breaches, including a requirement for covered entities to notify affected individuals if a breach of "unsecured protected health information" occurs that is "not secured through the use of a technology or methodology specified by the Secretary in (future) guidance."[75]

Application of the HIPAA Privacy Rule to research, including research involving the creation and use of registry data, has been broadly critiqued as unduly burdensome for researchers, incomplete and ineffective for protecting privacy, and significantly inconsistent with the Common Rule. Comprehensive and controversial proposals have been made to replace the current regulations. Critics propose a fundamentally different approach that would largely eliminate the need for patient authorization for research use and disclosures, while emphasizing stewardship and security.[76]

RISK MITIGATION FOR DATA LINKAGE PROJECTS

7. Methodology for Mitigating the Risk of Re-Identification

The disclosure limitation methods briefly described here are designed to protect against identification of individuals in statistical databases and are among the techniques that data linkage projects involving registries are most likely to use. One problem these methods do not address is the simultaneous protection of individual and institutional data sources. The discussion here also relates to the problems addressed by secure computation methodologies, which are explored in the next section.

7a. Basic methodology for statistical disclosure limitation

Duncan[77] categorizes the methodologies used for disclosure limitation in terms of disclosure limiting masks, i.e., transformations of the data where there is a specific functional relationship (possibly stochastic) between the masked values and the original data. The basic idea of masking involves data transformations. The goal is to transform an $n \times p$ data matrix Z through pre- and post-multiplication and the possible addition of noise, such as depicted in Equation (1):

$$Z \rightarrow AZB+C \qquad (1)$$

where A is a matrix that operates on cases, B is a matrix that operates on variables, and C is a matrix that adds perturbations or noise to the original information. Matrix masking includes a wide variety of standard approaches to disclosure limitation:

- Adding noise,
- Releasing a subset of observations (delete rows from *Z*),
- Cell suppression for cross-classifications,
- Including simulated data (add rows to *Z*),
- Releasing a subset of variables (delete columns from *Z*), and
- Switching selected column values for pairs of rows (data swapping).

This list also omits some methods, such as micro-aggregation and doubly random swapping, but it provides a general idea of the types of techniques being developed and applied in a variety of contexts, including to medical and public health data.

The possibilities of both identity and attribute disclosure remain even when a mask is applied to a dataset, although the risks may be substantially diminished.

Duncan suggests that we can categorize most disclosure limiting masks as suppressions (e.g., cell suppression), re-codings (e.g., collapsing rows or columns, or swapping), or samplings (e.g., releasing subsets), although he also allows for simulations as discussed below. Further, some masking methods alter the data in systematic ways, e.g., through aggregation or through cell suppression, whereas others do it through random perturbations, often subject to constraints for aggregates. Examples of perturbation methods are controlled random rounding, data swapping, and the post-randomization method (PRAM) of Gouweleeuw [78] that has been generalized by Duncan and others. One way to think about random perturbation methods is as a restricted simulation tool that connects them to other types of simulation approaches.

Various authors pursue simulation strategies and present general approaches to "simulating" from a constrained version of the cumulative empirical distribution function of the data. In 1993, Rubin asserted that the risk of identity disclosure could be eliminated by the use of synthetic data (in his case using Bayesian methodology and multiple imputation techniques) because there is no direct function link between the original data and the released data.[79] On the other hand, said another way, the data remain confidential because simulated individuals have replaced all of the real ones. Raghunathan, Reiter, and Rubin [80] provide details on the implementation of this approach. Abowd and Woodcock (for their chapter in Doyle et al., 2001) [81] describe a detailed application of multiple imputation and related simulation technology for a longitudinally linked individual and work history dataset. With both simulation or multiple-imputation methodology, however, it is still possible that the data values of some simulated individuals remain virtually identical to those in the original sample, or at least close enough that the possibility of both identity and attribute disclosure remain. As a result, checks for the possibility of unacceptable disclosure risk should be made.

Another important feature of the statistical simulation approach is information on the variability is directly accessible to the user. For example, in the Fienberg, Makov, and Steele [82] approach for categorical data, the data user can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis-Sturmfels Monte Carlo Markov chain algorithm to regenerate the full distribution of all possible tables with those margins. This technique allows the user to make inferences about the added variability in a modeling context that is similar to the approach to inference in Gouweleeuw, et al.[83] Similarly, Raghunathan and colleagues proposed the use of multiple imputations to measure directly the variability associated with the posterior distribution of the quantities of interest.[84] As a consequence, Rubin showed that simulation and perturbation methods represent a major improvement in access to data without sacrificing confidentiality over cell suppression and data swapping. These methods also conform to the statistical principle allowing the user of released data to apply standard statistical operations without being misled.

There has been considerable research on disclosure limitation methods for tabular data, especially in the form of multi-way tables of counts (contingency tables). The most popular methods include a process known as cell suppression, which systematically deletes the values in selected cells in the table and collapses categories. This process is a form of aggregation. While cell suppression methods have been very popular among the U.S. government statistical agencies and are useful for tables with non-negative entries rather than simply counts, they also have major drawbacks. First, good algorithms do not yet exist for the methodology when it is associated with high dimensional tables. More importantly, the methodology systematically distorts the information about the cells in the table for users and, as a consequence, makes it difficult for secondary users to draw correct statistical inferences about the relationships among the variables in the table. For further discussion on cell suppression and extensive references, see the various chapters in Doyle et al.,[85] especially the one by Duncan and his collaborators.

A special example of collapsing categories involves summing over variables to produce marginal tables. Instead of reporting the full multi-way contingency table, one or more collapsed versions of it might be reported. The release of multiple sets of marginal totals has the virtue of allowing statistical inferences about the relationships among the variables in the original table using log-linear model methods, e.g., see Bishop, Fienberg, and Holland.[86] With multiple collapsed versions, statistical theory makes it clear that one may have highly accurate information about the actual cell entries in the original table. As a result, the possibility of disclosures still requires investigation. In part to address this problem, a number of researchers have recently worked on the problem of determining upper and lower bounds for the cells of a multi-way table given a set of margins, although other measures of risk may clearly be of interest. The problem of computing bounds is in one sense an old one, at least for two-way tables, but it is also deeply linked to recent mathematical developments in statistics and has generated a flurry of new research.[87] [88]

7b. The risk-utility tradeoff

Common to virtually all of the methodologies discussed in the preceding section is the notion of a risk-utility tradeoff, in which the risk of disclosure is balanced with the utility of the released data, e.g., see Duncan[89], Fienberg[90], and their chapter with others in Doyle et al.[91] To keep this risk at a low level requires applying more extensive data masking that limits the utility of what is released. Advocates for the use of simulated data often claim that this use eliminates the risk of disclosure, but others dispute this claim.

7c. Privacy-preserving data mining methodologies

With the rise in the data mining and machine learning literature over the past two decades, there have been a large number of methods introduced under the banner of privacy-preserving computation. The methodologies vary, and many of them focus on standard tools such as the addition of noise or data swapping of one sort or another. But, the claims of identity protection in this literature are often exaggerated or unverifiable. For a discussion of some of these ideas and methods, see Fienberg and Slavkovic.[92] For two recent interesting examples explicitly set in the context of medical data, see Malin and Sweeney[93] and Boyens, Krishnan, and Padman.[94]

The common message of this literature is privacy protection has costs measured in the lack of availability of research data. To increase the utility of released data for research, some measure of privacy protection, however small, needs to be sacrificed. It is nonetheless still possible to optimize utility, subject to predefined upper bounds on what is considered to be acceptable risk of identification. See a related discussion in Fienberg.[95]

7d. Cryptographic approaches to privacy protection

While the current risks of identification in modern databases is similar for statistical agencies and biomedical researchers, there are also new challenges: contemporary information repositories that store social network data (e.g., cell phone, MySpace and Facebook data), product preferences (e.g., Amazon), web search data, and other sources of information not previously archived in a digital format. A recent literature emanating from cryptography focuses on algorithmic aspects of the problem with an emphasis on automation and scalability of a process for conferring anonymity. Automation, in turn, presents a fundamentally different perspective on how privacy is defined and provides for both a formal definition of privacy and proofs for how it can be protected. By focusing on the properties of the algorithm for anonymity, it is possible to formally guarantee the degree of privacy protection and the quality of the outputs in advance of data collection and publication.

This new approach, known as differential privacy, limits the incremental information a data user might learn beyond that known before exposure to the released statistics. No matter what external information is available, the differential privacy approach guarantees the same information is learned about an individual, whether or not information about the individual is present in the database. The papers by Dwork et al.[96, 97] provide an entry point to this literature. Differential privacy, as these authors describe it, works primarily through the addition of specific forms of noise to all data elements and the summary information reported, but it does not

address issues of sampling or access to individual-level micro-data. While these methods are intriguing, their utility for data linkages with registry data remains an open issue.

8. Security Practices, Standards, and Technologies

In general, people adopt two different philosophical positions about the preservation of confidentiality associated with individual-level data: (1) *restricted* or *limited information*, with restrictions on the amount or format of the data released, and (2) *restricted* or *limited access*, with restrictions on the access to the information itself.

If registry data are a public health good, then restricted access is justifiable only in situations where the confidentiality of data in the possession of a researcher cannot be protected through some form of restriction on the information released. Restricted access is intended to allow use of unaltered data by imposing certain conditions on users, analyses, and results that limit disclosure risk. There are two primary forms of restricted access. Licensing legally binds users to conditions such as agreeing not to use data for re-identification and advance review of publications, but allows users to transfer data to their sites, and to use the software of their choice. Research data centers, discussed in more detail below, and remote analysis servers are similar conceptually: users, and sometimes analyses, are evaluated in advance. The results are reviewed, and often limited, in order to limit risk of disclosure. The data remain at the holder's site and computers; the difference is whether access is in person or using the World Wide Web.

8a: Accountability

To limit the possibility of re-identification, data can be manipulated by the above techniques to mitigate risk. At the same time, it is important to ensure that researchers are accountable for the use of the datasets that are made available to them. Best practices in data security should be adopted with specific emphasis on authentication, authorization, access control, and auditing. In particular, each data recipient should be assigned a unique login ID or, if the data is made available online, through a query-response server. Prior to each session of data access, data custodians should authenticate users. Access to information should be controlled either in a role-based or information-based manner. Each user access and query to the data should be logged to enable auditing functions. If there is a breach in data protection, the data custodian can investigate the potential cause and make any required notifications.

8b: Registries as data enclaves

To limit access to data, many statistical agencies have built enclaves, often referred to as research data centers, where users can access and use data in a regulated environment. In such settings, the security of computer systems is controlled and managed by the agency providing the data. Such environments may maximize data security. For a more extensive discussion of the benefits of restricted access, see the chapter by Dunne in Doyle, et al.[98]

These enclaves incur considerable costs associated with their establishment and upkeep. A further limitation is that the enclave may require the physical presence of the data user, which also increases the overall cost to researchers working with the data. Moreover, such

environments often prevent users from executing specialized data analyses, which may require programming and other software development beyond the scope of traditional statistical software packages made available in the enclave.

The process for access to data in enclaves or restricted centers involves an examination of the research credentials of those wishing to do so. In addition, these centers control the physical access to confidential data files and review the materials that data users wish to take from the centers and to publish. Researchers who are accustomed to reporting residual plots and other information that allows for a partial reconstruction of the original data, at least for some variables, will encounter difficulties, because restricted data centers typically do not allow users to remove such information.

It is important to recognize, however, that query-response is not applicable to every data-sharing scenario, and data is often shared beyond the enclaves. When access logging and auditing is not possible, then a review of the security procedures and technologies employed by the recipient institution should occur. This process may be outsourced to an independent auditing agency.

8c: Layered restricted access to databases

In many countries, the traditional arrangement for data use involves restrictions on both information and access, with only highly aggregated data and summary statistics released for public use.

One potential strategy for privacy protection for the linkage of registries to other confidential data is a form of layered restrictions that combines two approaches with differing levels of access at different levels of detail in the data. The registry might function as an enclave similar to those described in the preceding section and public access might be limited to only aggregate data. Between these two extremes there might be several layers of restricted access. An example is licensing that includes privacy protection requiring greater protection as the potential for disclosure risk increases.

Such a layered approach might require a broader interpretation of the HIPAA Privacy Rule restrictions for certain kinds of medical records[99] or different forms of releases for patient records. The HIPAA Privacy Rule's detailed approach to releasing data can be shown not to fully protect individual data and at the same time unnecessarily restricts access to medical record data for research purposes. As a result, there is a need to develop a clearer sense of how health information subject to the HIPAA Privacy Rule might be linked with registry data and then subsequently protected. Such clarifications could allow for more complete research data while offering protection against the risks of identity disclosure to individuals and health care providers.

SUMMARY

This paper describes technical and current legal considerations for researchers interested in creating data linkage projects involving registry data. The discussion of the HIPAA Privacy Rule

provides a basis for understanding the conditions under which the use and disclosure of protected health information (PHI) is permitted for research and other purposes relevant to registries. These conditions determine whether and how the linkage of certain datasets may be legally feasible. In addition, the paper presents typical methods for record linkage that are likely to form the basis for the construction of data linkage projects. It also discusses both the hazards for re-identification created by data linkage projects, and the statistical methods used to minimize the risk of re-identification. Some limitations of this discussion are the exclusion of: 1) considerations about linking data from public and private sectors, where different ethical and legal restrictions may apply, and 2) detailed information about the risks involved with identifying the health care providers that collect and provide data.

Dataset linkage entails the risks of loss of reliable confidential data management and identification or re-identification of individuals and institutions. Recognized and developing statistical methods and secure computation may limit these risks and may allow the public health benefits that registries linked to other datasets have the potential to contribute.

Appendix I. Data-Linkage Scenario: Linking Clinical Registry Data with Insurance Claims Files

A research project is being designed to compare the effectiveness of one class, or specific generic medication within the class, to another for treating diabetes. The results should provide scientific evidence for patients, physicians, and policy makers on which to base decisions about the use of these drugs.

Registry developer A will collect limited data sets of information on patients discharged with a diagnosis of diabetes from hospitals in three states. These limited data sets do not include patient names or direct identifiers, so are not considered individually identifiable health information under the HIPAA Privacy Rule. The registry developer has institutional review board (IRB) approval to use this data for research purposes. The hospitals will provide the data sets to group B under a data use agreement that complies with the HIPAA Privacy Rule.

Group B intends to perform probabilistic matching of the registry data to a health insurance claims database to determine diabetes treatment outcomes. Registry developer A and the research group B entered into a formal collaboration for this research project.

The health insurance database will be derived from the claims data of multiple health plans operating in the same three states. The insurers' original data sets include direct beneficiary identifiers and constitute protected health information under the HIPAA Privacy Rule. Because the registry contains only limited data sets, the claims data collected in the insurance database will have to be linked to the registry data using probabilistic matching techniques.

Consequently, the research project will only use a limited data set of health insurance claims data to create the link with the registry data. The health insurance companies will provide the limited data sets of claims information to group B under data use agreements that comply with the HIPAA Privacy Rule.

The common data elements for the insurance database and the registry that will be used for linkage are date of birth, gender, race, hospital ID, state of hospital, date of admission, date of discharge, date of death (if patients died), ICD-9 code for primary diagnosis for the index hospitalization, primary procedure code for the index hospitalization, and the ZIP codes for patient addresses.

In order to protect the identity of the hospitals, the researchers were asked to sign a confidentiality agreement, which specifically defined the registry operator's proprietary information. Such proprietary information included the names or other identifiers of hospitals or other healthcare facilities participating in the registry. The researchers were precluded from using the names or other distinguishing characteristics of the hospitals in any public document, including publications or marketing materials. The confidentiality agreement did allow the

researchers to retain an identifier number for each hospital, so long as that number only identified generic characteristics and excluded any information about the hospital that would enable anyone to identify the specific hospital. For example, the researchers could not retain information that classified a particular hospital with a number that identified it as an academic teaching hospital based in a particular state with a certain number of beds, since in many instances, the identity of the hospital could be derived from such information.

Due to the potential contractual liability that may arise, the possibility of identifying participating hospitals is a critical issue to address.

Appendix II. Planning for Data Linkage Projects

The following questions are intended to assist the planning of data linkage projects that involve using registry data plus other files. Registry operators should use the answers to these questions to assemble necessary information and other resources to guide planning for their data linkage project. The questions, like the preceding discussion, are divided into regulatory and technical questions.

REGULATORY ISSUES:

The assumptions listed below apply to the regulatory questions that follow. The application of these assumptions to the proposed data linkage project should be confirmed or determined.

- The HIPAA Privacy Rule applies to the initial data sources
- Other law may restrict access or use of the initial data sources
- The Common Rule or FDA regulations may or may not apply to data linkage
- The Common Rule or FDA regulations may or may not apply to original datasets

Different regulatory concerns arise depending on the answers to each category of the following questions. Consultation as necessary with experienced health services, social science, or statistician colleagues, and regulatory personnel (Privacy Officer) or legal counsel is recommended to clarify answers for specific data linkage projects.

1. Purpose for data linkage:

- Research?
- Public health?
- Quality improvement?
- Required for post-marketing safety studies?
- Determining effectiveness of a product or service?
- Other purpose?
- Combination of purposes?

2. Conditions under which data (plus or minus biospecimens) were originally collected:

- Collected by law (e.g., regulatory purpose, public health purpose)?
- For treatment, payment, or health care operations?
- With documented consent to research participation and authorization for research use of protected health information from each individual?
- With an IRB alteration or waiver of consent and authorization?
- With permission of health care provider or plan?
- Contractual conditions or limitations on future use or disclosure (release)?
- What are the reasonable expectations of privacy or confidentiality for future uses of the data held by the original data sources and the data custodians?

3. Data:

- Is sensitive information involved (e.g., about children, infectious disease, mental health conditions)?
- Does the data contain direct identifiers? Indirect identifiers?
- Is protected health information involved (PHI)?
- Is a limited data set (LDS) and thus a data use agreement (DUA) involved?
- Is the data de-identified in accordance with the HIPAA Privacy Rule
- Does the data contain a code to identifiers?
- Who holds the key to the code?
- Is an honest broker involved?
- Does the code to identifiers conform to the re-identification standard in the HIPAA Privacy Rule?
- Is re-identification needed prior to performing the data linkage?
- After the data linkage, will the risk increase that the data may be identifiable?
- What is the minimally acceptable cell size to avoid identifying individuals?

4. Is the person or institution holding the data for the linkage:

- A covered entity under the HIPAA Privacy Rule or the American Recovery and Reinvestment Act of 2009?
- Not a covered entity?

5. Is the person or institution performing data linkage:

- A covered entity under the HIPAA Privacy Rule or the American Recovery and Reinvestment Act of 2009?
- Not a covered entity?

6. What other law or policies may apply to data use or disclosure (release)?

- Is governmental data involved?
- Are NIH data sharing policies involved?
- Does state law apply? Which state?

7. What are the terms and conditions that apply to data disclosure (release) and use under any agreement with the original source of the data.

- For individuals as the data source, the consent and authorization documents may contain limitations on data use, unless the data has been sufficiently de-identified.
- For data custodians as the data source, there may be a data use agreement or other contract that applies to data use by any subsequent holder of the data.

8. What are anticipated needs for data validation and verification?

- Initially for the data linkage processes?
- In the future?

9. What are future needs for privacy protection of the data source or maintenance of data confidentiality?

10. Anticipated future uses of the linked data:

- Will the data resulting from the linkage be maintained for multiple analyses? For the same or different purposes?
- Will the data resulting from the linkage be used for other linkages in the future?
- What permissions are necessary for or restrictions apply to planned future uses of the data?
- Are there requirements currently for tracking uses and disclosures of the data?

For information about the applicability of the HIPAA Privacy Rule, the Common Rule, and FDA regulations (human subjects protection) to registries, refer to chapter 6 of the manual, *Registries for Evaluating Patient Outcomes: A User's Guide* at http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=nr&ProcessID=21.

TECHNICAL ISSUES:

- Who is performing the linkage? Are the individuals performing the linkage permitted access to identifiers or restricted set of identifiers? Are they an honest broker or the source of one of the datasets to be linked?
- How easy will it be to know whether a given person is in the registry? Censuses are riskier than surveys.
- How will the error induced by the linkage itself be managed?
- Is there a common feature or pseudonym available across the datasets being linked?
- Is the registry a flat file or relational database? The latter is more difficult to manage, unless a primary key is applied.
- Is the registry relatively static or dynamic? The latter is harder to manage if data is being added over time, because the risk of identification increases.
- How many attributes are in the registry? The more attributes, the more difficult the risk of identification associated with the registry will be to manage.
- How will conflicting values of attributes that are common to both databases be resolved? Comparable attributes (e.g., weight) should be converted to the same units of measurement in datasets that will be linked.
- Does the registry contain information that makes the risk identification intrinsic to the registry? Other than direct identifiers such as names or social security numbers, fine-scale geography is also problematic.
- Is there a sound data dictionary?
- How many external databases will be linked to the registry data? How readily available and costly is each external database?
- How will records that appear in only one database be managed?
- How will the accuracy of the linked dataset relate to the accuracy of its components? The accuracy is only as good as that of the least accurate component.

REFERENCES

[1] Clayton E. Ethical, legal, and social implications of genomic medicine (Review). *New England Journal of Medicine* 2003; 349:562-569.

[2]Harris Louis Harris and Associates. Health Care Information Privacy: A Survey of the Public and Leaders, Conducted for EQUIFAX Inc; 1993.

[3] Gottlieb S. US employer agrees to stop genetic testing – Burlington Northern Santa Fe - News. *British Medical Journal* 2001*;* 322:449.

[4] Sterling R, Henderson G, Corbie-Smith G. Public willingness to participate in and public opinions about genetic variation research: a review of the literature. *American Journal of Public Health* 2006;96:1971-1978.

[5] Institute of Medicine, National Academy of Science, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Committee on Health Research and the Privacy of Health Information. (Nass, S.J., et al., eds.) Washington, DC: National Academies Press, 2009.

[6] Beckerman JZ, Pritts J, Goplerud E, et al. Health Information Privacy, Patient Safety, and Health Care Quality: Issues and Challenges in the Context of Treatment for Mental Health and Substance Use. *BNA's Health Care Policy Report* 2008 Jan 14;16(2).

[7] Solove D. A taxonomy of privacy. *University of Pennsylvania Law Review* 2006;154:477-560.

[8] Institute of Medicine, National Academy of Science, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Committee on Health Research and the Privacy of Health Information. (Nass, S.J., et al., eds.) Washington, DC: National Academies Press, 2009.

[9] Duncan GT, Jabine TB, de Wolf VA, editors. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Panel on Confidentiality and Data Access*. Committee on National Statistics. Washington, DC: National Research Council and the Social Science Research Council, National Academy Press;1993.

[10] Fienberg SE. *Confidentiality and disclosure limitation.* Encyclopedia of Social Measurement, Academic Press, Vol. 2, 2005;p.463–469.

[11] Duncan GT, Jabine TB, de Wolf VA, editors. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Panel on Confidentiality and Data Access*. Committee on National Statistics. Washington, DC: National Research Council and the Social Science Research Council, National Academy Press;1993.

[12] Fienberg SE. *Confidentiality and disclosure limitation.* Encyclopedia of Social Measurement, Academic Press, Vol. 2, 2005;p.463–469.

[13] Federal Committee on Statistical Methodology: Report on statistical disclosure limitation methodology. Statistical Policy Working paper 22; 2005. Publication No. NTIS PB94-165305. Available at: http://www.fcsm.gov/working-papers/spwp22.html

[14] 45 C.F.R. 160.103

[15] Institute of Medicine, National Academy of Science, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Committee on Health Research and the Privacy of Health Information. (Nass, S.J., et al., eds.) Washington, DC: National Academies Press, 2009.

[16] Fellegi IP, Sunter AB. A Theory for Record Linkage, *Journal of the American Statistical Association* 1969;40:1183-1210.

[17] BIlenko M, Mooney R, Cohen WW, et al. Adaptive name matching in information integration. *IEEE Intelligent Systems* 2003;18(5):16–23.

[18] Herzog TN, Schuren FJ, Winkler WE. Data Quality and Record Linkage Techniques. New York: Springer-Verlag; 2007.

[19] BIlenko M, Mooney R, Cohen WW, et al. Adaptive name matching in information integration. *IEEE Intelligent Systems* 2003;18(5):16–23.

[20] Winkler WE. *Overview of record linkage and current research directions.* US Census Bureau. Pulication No. RR 2006/02.

[21] Christen P, Churches T, Hegland M. A parallel open source data linkage system*. 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*;2004 May; Sydney, AUS.

[22] Abowd J, Vilhuber L. The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers (with discussion). *Journal of Business and Economics Statistics*, 2005;23(2):133-165.

[23] Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making* 2009; 9(3).

[24] Karr AF, Fulp WJ, Lin X, et al. Secure, privacy-preserving analysis of distributed databases. *Technometrics* 2007; 49(3):335-345.

[25] Karr AF, Lin X, Reiter, JP, et al. Privacy preserving analysis of vertically partitioned data using secure matrix products. *To appear in Journal of Official Statistics*; 2009.

[26] Rivest, RL, Adleman, L, Dertouzos, M. L. On data banks and privacy homomorphisms. *Foundations of Secure Computation;* R. DeMillo, ed. New York: Academic Press; 1978.

[27] Karr AF, Banks DL, Sanil AP. Data quality: A statistical perspective. *Statistical Methodology* 2006; 3(2):137-173.

[28] Sweeney L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, and Ethics* 1997;25: 98-110.

[29] Golle P. Revisiting the uniqueness of simple demographics in the U.S. population. ACM Workshop on Privacy in the Electronic Society, 2006. 77-80.; Sweeney L. Uniqueness of simple demographics in the US population. Carnegie Mellon University Data Privacy Laboratory, Pittsburgh, PA; 2000. Report Number LIDAP-WP04.

[30] Malin B. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *Journal of the American Medical Informatics Association* 2005; 12:28-34.

[31] Lin Z, Owen A, Altman, R. Genetics: genomic research and human subject privacy. *Science* 2004; 305:183.

[32] Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics* 2008; 4:e1000167.

[33] Cassa C, Schmidt B, Kohane I, et al. My sister's keeper?: genomic research and the identifiability of siblings. *BMC Medical Genomics* 2008 (32).

[34] Rothstein, MA. Genetic secrets: promoting privacy and confidentiality in the genetic era. New Haven: Yale University; 1997; Kass N, Medley A. Genetic screening and disability insurance: what can we learn from the health insurance experience. *Journal of Law, Medicine, and Ethics* 2007;35:66-73.; Phelan, JC. Geneticization of deviant behavior and consequences for stigma: the case of mental illness. *Journal of Health and Social Behavior* 2005;**46**:307-322.

[35] Pub. L. 110-233

[36] Pub. L. 104-191 §§ 262 and 264

[37] Pritts J. The importance and value of protecting the privacy of health information: the roles of the HIPAA privacy rule and the common rule in health research. 2009. Available at http://www.iom.edu/Object.File/Master/53/160/Pritts%20Privacy%20Final%20Draft%20web.pdf. Accessed on May 20, 2009.

[38] McGraw et. al., Privacy as an Enabler, Health Affairs 28:2 416-427.

[39] McGraw et. al., Privacy as an Enabler, Health Affairs 28:2 416-427.

[40] P. L. 111-5 §13400 et seq.

[41] Pritts J. The importance and value of protecting the privacy of health information: the roles of the HIPAA privacy rule and the common rule in health research. 2009. Available at http://www.iom.edu/Object.File/Master/53/160/Pritts%20Privacy%20Final%20Draft%20web.pdf. Accessed on May 20, 2009.

[42] Gelman R. Fair information practices: A basic history. 2008. Available at http://bobgellman.com/rg-docs/rg-FIPshistory.pdf. Accessed on June 1, 2009.

[43] Pub. L. 93-579, 5 U.S.C. Section 552a.

[44] Standards for Privacy of Individually Identifiable Health Information, Final Rule, 65 Fed. Reg. 82,462, 82,464 (Dec. 28, 2000).

[45] 45 C.F.R. 160.103

[46] Beckerman JZ, Pritts J, Goplerud E, et al. Health Information Privacy, Patient Safety, and Health Care Quality: Issues and Challenges in the Context of Treatment for Mental Health and Substance Use. *BNA's Health Care Policy Report* 2008 Jan14:16(2).

[47] 42 CFR Parts 2 and 2a

[48] 45 C.F.R. 160.102(a), 164.500

[49] 45 C.F.R. 164.501

[50] Pub. L. 111-5 §13406

[51] 45 C.F.R. 164.502(b), 164.508

[52]. 45 C.F.R. 164.514(d)(2)

[53] 45 C.F.R. 164.512

[54] 45 C.F.R. 164.512(i)(2)(ii) and (iii)

[55] 45 C.F.R. 164.502(e)

[56] Pub. L. 111-5 §13408

[57] Pub. L. 111-5 §13404

[58] Pub. L. 111-5 §13410

[59] 45 C.F.R. 165.514

[60] 45 C.F.R. 164.514(a) and (b)

[61] McGraw et. al. Privacy as an Enabler. *Health Affairs* 28(2):416-427.

[62] Sweeney L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, and Ethics* 1997;25: 98-110.

[63] Pritts J. The importance and value of protecting the privacy of health information: the roles of the HIPAA privacy rule and the common rule in health research. 2009. Available at http://www.iom.edu/Object.File/Master/53/160/Pritts%20Privacy%20Final%20Draft%20web.pdf. Accessed on May 20, 2009.

[64] Sweeney, L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, and Ethics* 1997;25:98-110.

[65] Aggarwal CC, Yu PS, editors. *Privacy-preserving data mining: models and algorithms*. Boston: Kluwer Academic Publishers; 2008.

[66] 45 C.F.R. 164.54(c)

[67] 45 C.F.R. 164.514(e)(3)(i)

[68] 45 C.F.R. 165.514(e)(3)(ii)

[69] Pritts J. The importance and value of protecting the privacy of health information: the roles of the HIPAA privacy rule and the common rule in health research. 2009. Available at http://www.iom.edu/Object.File/Master/53/160/Pritts%20Privacy%20Final%20Draft%20web.pdf. Accessed

on May 20, 2009.

[70] 45 C.F.R. 164.514(e)(3)

[71] 45 C.F.R. 164.530(c)

[72] 45 C.F.R. Parts164, Subpart C

[73] Institute of Medicine, National Academy of Science, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Committee on Health Research and the Privacy of Health Information. (Nass, S.J., et al., eds.) Washington, DC: National Academies Press, 2009.

[74] Pub. L. 111-5 §13401

[75] Pub. L. 111-5 §13402

[76] Institute of Medicine, National Academy of Science, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Committee on Health Research and the Privacy of Health Information. (Nass, S.J., et al., eds.) Washington, DC: National Academies Press, 2009.

[77] Duncan GT. Confidentiality and statistical disclosure limitation. *International Encyclopedia of the Social and Behavioral Sciences*. In Smelser N and Baltes P, editors, Vol. 4. New York: Elsevier; 2001. p.2521-2525.

[78] Gouweleeuw JM, Kooiman P, Willenborg LCRJ, et al. Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics* 1998;14:463-478.

[79] Rubin, Donald B. Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics* 1993, 9(2), 461-468.

[80] Raghunathan TE, Reiter J, Rubin DB. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 2003;19:1-16.

[81] Abowd J, Vilhuber L. The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers (with discussion). *Journal of Business and Economics Statistics*, 2005;23(2):133-165.

[82] Stephen E. Fienberg, Udi E. Makov, and Russell J. Steele. Disclosure Limitation Using Perturbation and Related Methods for Categorical Data (with discussion). *Journal of Official Statistics* 1998, 14(4), 485-511.

[83] Gouweleeuw JM, Kooiman P, Willenborg LCRJ, et al. Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics* 1998;14:463-478.

[84] Raghunathan TE, J.P. Reiter and D.B. Rubin. Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics* 2003, 19(1), 1-16.

[85] Doyle P, Lane J, Theeuwes J, et al, editors. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. New York: Elsevier; 2001.

[86] Yvonne M M Bishop YMM, SE Fienberg and PW Holland. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge. MA: MIT Press, New York: Springer-Verlag, 1995, Reprinted 2007.

[87] Dobra, Adrian and Fienberg, Stephen E. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, *Proceedings of the National Academy of Sciences*, 2000, 97, No. 22, 11885-11892.

[88] Fienberg SE, Slavkovic AB. Preserving the Confidentiality of Categorical Data Bases When Releasing Information for Association Rules." *Data Mining and Knowledge Discovery* 2005;11:155-180.

[89] Duncan GT. Confidentiality and statistical disclosure limitation. *International Encyclopedia of the Social and Behavioral Sciences*. In Smelser N and Baltes P, editors, Vol. 4. New York: Elsevier; 2001. p.2521-2525.

[90] Fienberg SE. Statistical perspectives on confidentiality and data access in public health. *Statistics in Medicine 2001;*20*:*1347-1356*.*

[91]Doyle P, Lane J, Theeuwes J, et al, editors. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. New York: Elsevier; 2001.

[92] Fienberg SE, Slavkovic AB. Preserving the Confidentiality of Categorical Data Bases When Releasing Information for Association Rules." *Data Mining and Knowledge Discovery* 2005;11:155-180.

[93] Malin B, Sweeney L. A secure protocol to distribute unlinkable health data. Proceedings of the American Medical Informatics Association Annual Meeting Symposium. Washington, DC: American Medical Informatics Association; 2005.

[94] Boyens C, Krishnan R, Padman R. On Privacy-Preserving Access to Distributed Heterogeneous Healthcare Information. *Proceedings of 37th Hawaii International Conference on System Sciences*. Publication No. HICSS-37 2004). 2009

[95] Fienberg SE. Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation. *Statistical Science* 2006;21:143-154.

[96] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors., TCC, Lecture Notes in Computer Science. Berlin: Springer-Verlag; 2006a;3876: 265–284.

[97] Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: Privacy via distributed noise generation. In EUROCRYPT; 2006:486–503.

[98] Doyle P, Lane J, Theeuwes J et al, editors. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. New York: Elsevier; 2001.

[99] Institute of Medicine, National Academy of Science, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Committee on Health Research and the Privacy of Health Information. (Nass, S.J., et al., eds.) Washington, DC: National Academies Press, 2009.