Chapter 11

Computational Disclosure Control for Medical Microdata: The Datafly System

Latanya Sweeney, Massachusetts Institute of Technology

Abstract

We present a computer program named Datafly that uses computational disclosure techniques to maintain anonymity in medical data by automatically generalizing, substituting and removing information as appropriate without losing many of the details found within the data. Decisions are made at the field and record level at the time of database access, so the approach can be used on the fly in role-based security within an institution, and in batch mode for exporting data from an institution. Often organizations release and receive medical data with all explicit identifiers, such as name, address, phone number, and social security number, removed in the incorrect belief that patient confidentiality is maintained because the resulting data look anonymous; however, we show that in most of these cases, the remaining data can be used to re-identify individuals by linking or matching the data to other databases or by looking at unique characteristics found in the fields and records of the database itself. When these less apparent aspects are taken into account, each released record can be made to ambiguously map to many possible people, providing a level of anonymity which the user determines.

Introduction

Sharing and disseminating electronic medical records while maintaining a commitment to patient confidentiality is one of the biggest challenges facing medical informatics and society at large. To the public, patient confidentiality implies that only people directly involved in their care will have access to their medical records and that these people will be bound by strict ethical and legal standards that prohibit further disclosure (Woodward, 1996). The public is not likely to accept that their records are kept "confidential" if large numbers of people have access to their contents.

On the other hand, analysis of the detailed information contained within electronic medical records promises many advantages to society, including improvements in medical care, reduced institution costs, the development of predictive and diagnostic support systems, and the integration of applicable data from multiple sources into a unified display for clinicians; but these benefits require sharing the contents of medical records with secondary viewers, such as researchers, economists, statisticians, administrators, consultants, and computer scientists, to name a few. The public would probably agree these secondary parties should know some of the information buried in the record, but such disclosure should not risk identifying patients.

In 1996, the National Association of Health Data Organizations (NAHDO) reported that 37 states had legislative mandates to gather hospital-level data. Last year, 17 of these states reported they had started collecting ambulatory care (outpatient) data from hospitals, physician offices, clinics, and so on. Table 1 contains a list of the fields of information which NAHDO recommends these states accumulate. Many of these states have subsequently given copies of collected data to researchers and sold copies to industry. Since the information has no explicit identifiers, such as name, address, phone number or social security number, confidentiality is incorrectly believed to be maintained.

Table 1. -- Data Fields Recommended by NAHDO for State Collection of Ambulatory Data

Patient Number
Patient ZIP Code
Patient Racial Background
Patient Birth Date
Patient Gender
Visit Date
Principal Diagnosis Code (ICD9)
Procedure Codes (up to 14)
Physician ID#
Physician ZIP code
Total Charges

In fairness, there are many sources of administrative billing records with fields of information similar to those listed in Table 1. Hospital administrators often pass medical records along in part to independent consultants and outside agencies. There are the records maintained by the insurance companies. Pharmaceutical companies run longitudinal studies on identified patients and providers. Local drug stores maintain individualized prescription records. The list is quite extensive. Clearly, we see the possible benefits from sharing information found within the medical record and within records of secondary sources; but on the other hand, we appreciate the need for doctor-patient confidentiality. The goal of this work is to provide tools for extracting needed information from medical records while maintaining a commitment to patient confidentiality. These same techniques are equally applicable to financial, demographic and educational microdata releases, as well.

Background

We begin by first stating our definitions of de-identified and anonymous data. In de-identified data, all explicit identifiers, such as social security number, name, address and phone number, are removed, generalized or replaced with a made-up alternative. De-identifying data does not guarantee that the result is anonymous however. The term anonymous implies that the data cannot be manipulated or linked to identify any individual. Even when information shared with secondary parties is de-identified, we will show it is often far from anonymous.

There are three major difficulties in providing anonymous data. One of the problems is that anonymity is in the eye of the beholder. For example, consider Table 2. If the contents of this table are a subset of an extremely large and diverse database then the three records listed in Table 2 may appear anonymous. Suppose the ZIP code 33171 primarily consists of a retirement community; then there are very few people of such a young age living there. Likewise, 02657 is the ZIP code for Provincetown, Massachusetts, in which we found about 5 black women living there year-round. The ZIP code 20612 may have only one Asian family. In these cases, information outside the data identifies the individuals.

Table 2. -- De-identified Data that Are not Anonymous

ZIP Code	Birthdate	Gender	Ethnicity
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Most towns and cities sell locally collected census data or voter registration lists that include the date of birth, name and address of each resident. This information can be linked to medical microdata that include a date of birth and ZIP code, even if the names, social security numbers and addresses of the patients are not present. Of course, local census data are usually not very accurate in college towns and areas that have a large transient community, but for much of the adult population in the United States, local census information can be used to re-identify de-identified microdata since other personal characteristics, such as gender, date of birth, and ZIP code, often combine uniquely to identify individuals.

The 1997 voting list for Cambridge, Massachusetts contains demographics on 54,805 voters. Of these, birth date alone can uniquely identify the name and address of 12% of the voters. We can identify 29% by just birth date and gender, 69% with only a birth date and a 5-digit ZIP code, and 97% (53,033 voters) when the full postal code and birth date are used. These values are listed in Table 3. Clearly, the risks of re-identifying data depend both on the content of the released data and on related information available to the recipient.

Table 3. -- Uniqueness of Demographic Fields in Cambridge Voter List

Birth date alone	12%
birth date and gender	29%
birth date and 5-digit ZIP	69%
birth date and full postal code	97%

A second problem with producing anonymous data concerns unique and unusual information appearing within the data themselves. Instances of uniquely occurring characteristics found within the original data can be used by reporters, private investigators and others to discredit the anonymity of the released data even when these instances are not unique in the general population. Also, unusual cases are often unusual in other sources of data as well making them easier to identify. Consider the database shown in Table 4. It is not surprising that the social security number is uniquely identifying, or given the size of the database, that the birth date is also unique. To a lesser degree the ZIP codes in Table 4 identify individuals since they are almost unique for each record. Importantly, what may not have been known without close examination of the particulars of this database is that the designation of Asian as a race is uniquely identifying. During an interview, we could imagine that the janitor, for example, might recall an Asian patient whose last name was Chan and who worked as a stockbroker for ABC Investment since the patient had given the janitor some good investing tips.

Table 4. -- Sample Database in which Asian is

A Uniquely Identifying Characteristic

SSN	Ethnicity	Birth	Sex	ZIP
819491049	Caucasian	10/23/64	m	02138
749201844	Caucasian	03/15/65	m	02139
819181496	Black	09/20/65	m	02141
859205893	Asian	10/23/65	m	02157
985820581	Black	08/24/64	m	02138

Any single uniquely occurring value or group of values can be used to identify an individual. Consider the medical records of a pediatric hospital in which only one patient is older than 45 years of age. Or, suppose a hospital's maternity records contained only one patient who gave birth to triplets. Knowledge of the uniqueness of this patient's record may appear in many places including insurance claims, personal financial records, local census information, and insurance enrollment forms. Remember that the unique characteristic may be based on diagnosis, treatment, birth year, visit date, or some other little detail or combination of details available to the memory of a patient or a doctor, or knowledge about the database from some other source.

Measuring the degree of anonymity in released data poses a third problem when producing anonymous data for practical use. The Social Security Administration (SSA) releases public-use files based on national samples with small sampling fractions (usually less than 1 in 1,000); the files contain no geographic codes, or at most regional or size of place designators (Alexander et al., 1978). The SSA recognizes that data containing individuals with unique combinations of characteristics can be linked or matched with other data sources. So, the SSA's general rule is that any subset of the data that can be defined in terms of combinations of characteristics must contain at least 5 individuals. This notion of a minimal bin size, which reflects the smallest number of individuals matching the characteristics, is quite useful in providing a degree of anonymity within data. The larger the bin size, the more anonymous the data. As the bin size increases, the number of people to whom a record may refer also increases, thereby masking the identity of the actual person.

In medical databases, the minimum bin size should be much larger than the SSA guidelines suggest. Consider these three reasons: most medical databases are geographically located and so one can presume, for example, the ZIP codes of a hospital's patients; the fields in a medical database provide a tremendous amount of detail and any field can be a candidate for linking to other databases in an attempt to re-identify patients; and, most releases of medical data are not randomly sampled with small sampling fractions, but instead include most if not all of the database.

Determining the optimal bin size to ensure anonymity is tricky. It certainly depends on the frequencies of characteristics found within the data as well as within other sources for re-identification. In addition, the motivation and effort required to re-identify released data in cases where virtually all possible candidates can be identified must be considered. For example, if we release data that maps each record to 10 possible people and the 10 people can be identified, then all 10 candidates may even be contacted or visited in an effort to locate the actual person. Likewise, if the mapping is 1 in 100, all 100 could be phoned since visits may then be impractical, and in a mapping of 1 in 1000, a direct mail campaign could be employed. The amount of effort the recipient is willing to spend depends on their motivation. Some medical files are quite valuable, and valuable data will merit more effort. In these cases, the minimum bin size must be further increased or the sampling fraction reduced to render these efforts useless.

Of course, the expression of anonymity most semantically consistent with our intention is simply the probability of identifying a person given the released data and other possible sources. This conditional probability depends on frequencies of characteristics (bin sizes) found within the data and the outside world. Unfortu-

nately, this probability is very difficult to compute without omniscience. In extremely large databases like that of SSA, the database itself can be used to compute frequencies of characteristics found in the general population since it contains almost all the general population; small, specialized databases, however, must estimate these values. In the next section, we will present a computer program that generalizes data based on bin sizes and estimates. Following that, we will report results using the program and discuss its limitations.

Methods

Earlier this year, Sweeney presented the Datafly System (1997) whose goal is to provide the most general information useful to the recipient. Datafly maintains anonymity in medical data by automatically aggregating, substituting and removing information as appropriate. Decisions are made at the field and record level at the time of database access, so the approach can be incorporated into role-based security within an institution as well as in exporting schemes for data leaving an institution. The end result is a subset of the original database that provides minimal linking and matching of data since each record matches as many people as the user had specified.

Diagram 1 provides a user-level overview of the Datafly System. The original database is shown on the left. A user requests specific fields and records, provides a profile of the person who is to receive the data, and requests a minimum level of anonymity. Datafly produces a resulting database whose information matches the anonymity level set by the user with respect to the recipient profile. Notice how the record containing the Asian entry was removed; social security numbers were automatically replaced with made-up alternatives; and birth dates were generalized to the year, and ZIP codes to the first three digits. In the next three paragraphs we examine the overall anonymity level and the profile of the recipient, both of which the user provides when requesting data.

Diagram 1. -- The Input to the Datafly System is the Original Database and Some User Specifications, and the Output is a Database Whose Fields and Records Correspond to the Anonymity Level Specified by the User, in this Example, 0.7.

User
-fields & records
-recipient profile
-anonymity 0.7

Original Medical Database

SSN	Race	Birth	Sex	ZIP
819491049	Caucasian	10/23/64	m	02138
749201844	Caucasian	03/15/65	m	02139
819181496	Black	09/20/65	m	02141
859205893	Asian	10/23/65	m	02157
985820581	Black	08/24/64	m	02138

Datafly

Resulting Database, anonymity 0.7

SSN	Race	Birth	Sex	ZIP
444444444	Caucasian	1964	m	02100
555555555	Caucasian	1965	m	02100
333333333	Black	1965	m	02100
222222222	Black	1964	m	02100

The overall anonymity level is a number between 0 and 1 that specifies the minimum bin size for every field. An anonymity level of 0 provides the original data, and a level of 1 forces Datafly to produce the most general data possible given the profile of the recipient. All other values of the overall anonymity level between 0 and 1 determine the minimum bin size *b* for each field. (The institution is responsible for mapping the ano-

nymity level to actual bin sizes though Sweeney provides some guidelines.) Information within each field is generalized as needed to attain the minimum bin size; outliers, which are extreme values not typical of the rest of the data, may be removed. When we examine the resulting data, every value in each field will occur at least *b* times with the exception of one-to-one replacement values, as is the case with social security numbers.

Table 5 shows the relationship between bin sizes and selected anonymity levels using the Cambridge voters database. As A increased, the minimum bin size increased, and in order to achieve the minimal bin size requirement, values within the birth date field, for example, were re-coded as shown. Outliers were excluded from the released data and their corresponding percentages of N are noted. An anonymity level of 0.7, for example, required at least 383 occurrences of every value in each field. To accomplish this in the birth date field, dates were re-coded to reflect only the birth year. Even after generalizing over a 12 month window, the values of 8% of the voters still did not meet the requirement so these voters were dropped from the released data.

Table 5. -- Anonymity Generalizations for Cambridge Voters Data with Corresponding Bin Sizes *

Anonymity	BinSize	BirthDate	Drop%
1.0			
.9	493	24	4%
.8	438	24	2%
.7	383	12	8%
.6	328	12	5%
.5	274	12	4%
.4	219	12	3%
.3	164	6	5%
.2	109	4	5%
.1	54	2	5%
0.0			

* The birth date generalizations (in months) required to satisfy the minimum bin size are shown and the percentages of the total database dropped due to outliers is displayed. The user sets the anonymity level as depicted above by the slide bar at the 0.7 selection. The mappings of anonymity levels to bin sizes is determined by the institution.

In addition to an overall anonymity level, the user also provides a profile of the person who receives the data by specifying for each field in the database whether the recipient could have or would use information external to the database that includes data within that field. That is, the user estimates on which fields the recipient might link outside knowledge. Thus each field has associated with it a profile value between 0 and 1, where 0 represents full trust of the recipient or no concern over the sensitivity of the information within the field, and 1 represents full distrust of the recipient or maximum concern over the sensitivity of the field's contents. The role of these profile values is to restore the effective bin size by forcing these fields to adhere to bin sizes larger than the overall anonymity level warranted. Semantically related sensitive fields, with the exception of one-to-one replacement fields, are treated as a single concatenated field which must meet the minimum bin size, thereby thwarting linking attempts that use combinations of fields.

Consider the profiles of a doctor caring for a patient, a clinical researcher studying risk factors for heart disease and a health economist assessing the admitting patterns of physicians. Clearly, these profiles are all different. Their selection and specificity of fields are different; their sources of outside information on which they could link are different; and, their uses for the data are different. From publicly available birth certificate,

driver license, and local census databases, the birth dates, ZIP codes and gender of individuals are commonly available along with their corresponding names and addresses; so these fields could easily be used for re-identification. Depending on the recipient, other fields may be even more useful, but we will limit our example to profiling these fields. If the recipient is the patient's caretaker within the institution, the patient has agreed to release this information to the care-taker, so the profile for these fields should be set to 0 to give the patient's caretaker full access to the original information. When researchers and administrators make requests that do not require the most specific form of the information as found originally within sensitive fields, the corresponding profile values for these fields should warrant a number as close to 1 as possible but not so much so that the resulting generalizations do not provide useful data to the recipient. But researchers or administrators bound by contractual and legal constraints that prohibit their linking of the data are trusted, so if they make a request that includes sensitive fields, the profile values would ensure that each sensitive field adheres only to the minimum bin size requirement. The goal is to provide the most general data that are acceptably specific to the recipient. Since the profile values are set independently for each field, particular fields that are important to the recipient can result in smaller bin sizes than other requested fields in an attempt to limit generalizing the data in those fields; a profile for data being released for public use, however, should be 1 for all sensitive fields to ensure maximum protection. The purpose of the profile is to quantify the specificity required in each field and to identify fields that are candidates for linking; and in so doing, the profile identifies the associated risk to patient confidentiality for each release of data.

Results

Numerous tests were conducted using the Datafly System to access a pediatric medical record system (Sweeney, 1997). Datafly processed all queries to the database over a spectrum of recipient profiles and anonymity levels to show that all fields in medical records can be meaningfully generalized as needed since any field can be a candidate for linking. Of course, which fields are most important to protect depends on the recipient. Diagnosis codes have generalizations using the International Classification of Disease (ICD-9) hierarchy. Geographic replacements for states or ZIP codes generalize to use regions and population size. Continuous variables, such as dollar amounts and clinical measurements, can be treated as categorical values; however, their replacements must be based on meaningful ranges in which to classify the values; of course this is only done in cases where generalizing these fields is necessary.

The Group Insurance Commission in Massachusetts (GIC) is responsible for purchasing insurance for state employees. They collected encounter-level de-identified data with more than 100 fields of information per encounter, including the fields in Table 1 for approximately 135,000 patients consisting of state employees and their families (Lasalandra, 1997). In a public hearing, GIC reported giving a copy of the data to a researcher, who in turn stated she did not need the full date of birth, just the birth year. The average bin size based only on birth date and gender for that population is 3, but had the researcher received only the year of birth in the birth date field, the average bin size based on birth year and gender would have increased to 1125 people. It is estimated that most of this data could be re-identified since collected fields also included residential ZIP codes and city, occupational department or agency, and provider information. Furnishing the most general information the recipient can use minimizes unnecessary risk to patient confidentiality.

Comparison to μ-ARGUS

In 1996, The European Union began funding an effort that involves statistical offices and universities from the Netherlands, Italy and the United Kingdom. The main objective of this project is to develop specialized software for disclosing public-use data such that the identity of any individual contained in the released data cannot be recognized. Statistics Netherlands has already produced, but has not yet released, a first version of a program named μ-Argus that seeks to accomplish this goal (Hundepool, et al., 1996). The μ-Argus program is considered by many as the official confidentiality software of the European community even though Statistics Netherlands admittedly considers this first version a rough draft. A presentation of the concepts on which μ-Argus is based can be found in Willenborg and De Waal (1996).

The program μ-Argus, like the Datafly System, makes decisions based on bin sizes, generalizes values within fields as needed, and removes extreme outlier information from the released data. The user provides an overall bin size and specifies which fields are sensitive by assigning a value between 0 and 3 to each field. The program then identifies rare and therefore unsafe combinations by testing 2- or 3-combinations across the fields noted by the user as being identifying. Unsafe combinations are eliminated by generalizing fields within the combination and by local cell suppression. Rather than removing entire records when one or more fields contain outlier information, as is done in the Datafly System, the μ-Argus System simply suppresses or blanks out the outlier values at the cell-level. The resulting data typically contain all the rows and columns of the original data though values may be missing in some cell locations.

In Table 6a there are many Caucasians and many females, but only one female Caucasian in the database. Tables 6b and 6c show the resulting databases when the Datafly System and the μ-Argus System were applied to this data. We will now step through how the μ -Argus program produced the results in Table 6c.

Table 6a. -- There is Only One Caucasian Female, Even Though There are Many Females and Caucasians

SSN	Ethnicity	Birth	Sex	ZIP	Problem
819181496	Black	09/20/65	m	02141	shortness of breath
195925972	Black	02/14/65	m	02141	chest pain
902750852	Black	10/23/65	f	02138	hypertension
985820581	Black	08/24/65	f	02138	hypertension
209559459	Black	11/07/64	f	02138	obesity
679392975	Black	12/01/64	f	02138	chest pain
819491049	Caucasian	10/23/64	m	02138	chest pain
749201844	Caucasian	03/15/65	f	02139	hypertension
985302952	Caucasian	08/13/64	m	02139	obesity
874593560	Caucasian	05/05/64	m	02139	shortness of breath
703872052	Caucasian	02/13/67	m	02138	chest pain
963963603	Caucasian	03/21/67	m	02138	chest pain

Table 6b. -- Results from Applying the Datafly System to the

Data in Table 6a *

SSN	Ethnicity	Birth	Sex	ZIP	Problem
902387250	Black	1965	m	02140	shortness of breath
197150725	Black	1965	m	02140	chest pain
486062381	Black	1965	f	02130	hypertension
235978021	Black	1965	f	02130	hypertension
214684616	Black	1964	f	02130	obesity
135434342	Black	1964	f	02130	chest pain
458762056	Caucasian	1964	m	02130	chest pain
860424429	Caucasian	1964	m	02130	obesity
259003630	Caucasian	1964	m	02130	shortness of breath
410968224	Caucasian	1967	m	02130	chest pain
664545451	Caucasian	1967	m	02130	chest pain

*The minimum bin size is 2. The given profile identifies only the demographic fields as being likely for linking. The data are being made available for semi-public use so the Caucasian female record was dropped as an outlier.

Table 6c. -- Results from Applying the Approach of the m-Argus System to the Data in Table 6a*

SSN	Ethnicity	Birth	Sex	ZIP	Problem
	Black	1965	m	02141	shortness of breath
	Black	1965	m	02141	chest pain
	Black	1965	f	02138	hypertension
	Black	1965	f	02138	hypertension
	Black	1964	f	02138	obesity
	Black	1964	f	02138	chest pain
	Caucasian	1964	m	02138	chest pain
			f	*02139*	hypertension
	Caucasian	1964	m	02139	obesity
	Caucasian	1964	m	02139	shortness of breath
	Caucasian	1967	m	02138	chest pain
	Caucasian	1967	m	02138	chest pain

*The minimum bin size is 2. SSN was marked as being most identifying, the birth, sex, and ZIP fields were marked as being more identifying, and the ethnicity field was simply marked as identifying. Combinations across these were examined; the resulting suppressions are shown. The uniqueness of the Caucasian female is suppressed; but, there still remains a unique record for the Caucasian male born in 1964 that lives in the 02138 ZIP code.

The first step is to check that each identifying field adheres to the minimum bin size. Then, pairwise

combinations are examined for each pair that contains the "most identifying" field (in this case, SSN) and those that contain the "more identifying" fields (in this case, birth date, sex and ZIP). Finally, 3-combinations are examined that include the "most" and "more" identifying fields. Obviously, there are many possible ways to rate these identifying fields, and unfortunately different identification ratings yield different results. The ratings presented in this example produced the most secure result using the μ-Argus program though admittedly one may argue that too many specifics remain in the data for it to be released for public use.

The value of each combination is basically a bin, and the bins with occurrences less than the minimum required bin size are considered unique and termed outliers. Clearly for all combinations that include the SSN, all such combinations are unique. One value of each outlier combination must be suppressed. For optimal results, the μ-Argus program suppresses values which occur in multiple outliers where precedence is given to the value occurring most often. The final result is shown in Table 6c. The responsibility of when to generalize and when to suppress lies with the user. For this reason, the μ-Argus program operates in an interactive mode so the user can see the effect of generalizing and can then select to undo the step.

We will briefly compare the results of these two systems, but for a more in-depth discussion, see Sweeney (1997). The μ-Argus program checks at most 2- or 3-combinations of identifying fields, but not all 2- or 3-combinations are necessarily tested. Even if they were, there may exist unique combinations across 4 or more fields that would not be detected. For example, Table 6c still contains a unique record for a Caucasian male born in 1964 that lives in the 02138 ZIP code, since there are 4 characteristics that combine to make this record unique, not 2. Treating a subset of identifying fields as a single field that must adhere to the minimum bin size, as done in the Datafly System, appears to provide more secure releases of microdata.

Discussion

The Datafly and μ-Argus systems illustrate that medical information can be generalized so that fields and combinations of fields adhere to a minimal bin size, and by so doing, confidentiality can be maintained. Using such schemes we can even provide anonymous data for public use. There are two drawbacks to these systems but these shortcomings may be counteracted by policy.

One concern with both μ-Argus and Datafly is the determination of the proper bin size and its corresponding measure of disclosure risk. There is no standard which can be applied to assure that the final results are adequate. What is customary is to measure risk against a specific compromising technique, such as linking to known databases, that we assume the recipient is using. Several researchers have proposed mathematical measures of the risk which compute the conditional probability of the linker's success (Duncan, et al., 1987).

A policy could be mandated that would require the producer of data released for public use to guarantee with a high degree of confidence that no individual within the data can be identified using demographic or semi-public information. Of course, guaranteeing anonymity in data requires a criterion against which to check resulting data and to locate sensitive values. If this is based only on the database itself, the minimum bin sizes and sampling fractions may be far from optimal and may not reflect the general population. Researchers have developed and tested several methods for estimating the percentage of unique values in the general population based on a smaller database (Skinner, et al., 1992). These methods are based on subsampling techniques and equivalence class structure. In the absence of these techniques, uniqueness in the population based on demographic fields can be determined using population registers that include patients from the database, such as local census data, voter registration lists, city directories, as well as information from motor vehicle agencies, tax assessors and real estate agencies. To produce an anonymous database, a producer could use population registers to identify sensitive demographic values within a database, and thereby obtain a measure of risk for the release of the data.

The second drawback with the µ-Argus and Datafly systems concerns the dichotomy between researcher needs and disclosure risk. If data are explicitly identifiable, the public would expect patient consent to be required. If data are released for public use, then the producer should guarantee, with a high degree of confidence, that the identity of any individual cannot be determined using standard and predictable methods and reasonably available data. But when sensitive de-identified, but not necessarily anonymous, data are to be released, the likelihood that an effort will be made to re-identify an individual increases based on the needs of the recipient, so any such recipient has a trust relationship with society and the producer of the data. The recipient should therefore be held accountable.

The Datafly and µ-Argus systems quantify this trust by profiling the fields requested by the recipient. But recall that profiling requires guesswork in identifying fields on which the recipient could link. Suppose a profile is incorrect; that is, the producer misjudges which fields are sensitive for linking. In this case, these systems might release data that are less anonymous than what was required by the recipient, and as a result, individuals may be more easily identified. This risk cannot be perfectly resolved by the producer of the data since the producer cannot always know what resources the recipient holds. The obvious demographic fields, physician identifiers, and billing information fields can be consistently and reliably protected. However, there are too many sources of semi-public and private information such as pharmacy records, longitudinal studies, financial records, survey responses, occupational lists, and membership lists, to account a priori for all linking possibilities.

Table 7. -- Contractual Requirements for Restricted-Use of Data Based on Federal Guidelines and the Datafly System

There must be a legitimate and important research or administrative purpose served by the release of the data. The recipient must identify and explain which fields in the database are needed for this purpose.

1. The recipient must be strictly and legally accountable to the producer for the security of the data and must demonstrate adequate security protection.
2. The data must be de-identified. It must contain no explicit individual identifiers nor should it contain data that would be easily associated with an individual.
3. Of the fields the recipient requests, the recipient must identify which of these fields, during the specified lifetime of the data, the recipient could link to other data the recipient will have access to, whether the recipient intends to link to such data or not. The recipient must identify those fields for which the recipient will link the data.
4. The provider should have the opportunity to review any publication of information from the data to insure that no potential disclosures are published.
5. At the conclusion of the project, and no later than some specified date, the recipient must destroy all copies of the data.
6. The recipient must not give, sell, loan, show, or disseminate the data to any other parties.

What is needed is a contractual arrangement between the recipient and the producer to make the trust explicit and share the risk. Table 7 contains some guidelines that make it clear which fields need to be protected against linking since the recipient is required to provide such a list. Using this additional knowledge and the techniques presented in the Datafly System, the producer can best protect the anonymity of patients in data even when the data are more detailed than data for public-use. Since the harm to individuals can be extreme and irreparable and can occur without the individual's knowledge, the penalties for abuses must be stringent. Significant sanctions or penalties for improper use or conduct should apply since remedy against abuse lies outside the Datafly System and resides in contracts, laws and policies.

Acknowledgments

The author acknowledges Beverly Woodward, Ph.D., for many discussions, and thanks Patrick Thompson, for editorial suggestions. The author also acknowledges the continued support of Henry Leitner and Harvard University DCE. This work has been supported by a Medical Informatics Training Grant (1 T15 LM07092) from the National Library of Medicine.

References

Alexander, L. and Jabine, T. (1978). Access to Social Security Microdata Files for Research and Statistical Purposes, *Social Security Bulletin,* (41), 8.

Duncan, G. and Lambert, D. (1987). The Risk of Disclosure for Microdata, *Proceedings of the Bureau of the Census Third Annual Research Conference*, Washington, D.C.: Bureau of the Census.

Hundepool, A. and Willenborg, L. (1996). μ - and τ-ARGUS: Software for Statistical Disclosure Control, *Third International Seminar on Statistical Confidentiality*, Bled.

Lasalandra, M. (1997). Panel Told Releases of Medical Records Hurt Privacy, *Boston Herald*, Boston, (35).

National Association of Health Data Organizations. (1996). A Guide to State-Level Ambulatory Care Data Collection Activities, Falls Church, VA.

Skinner, C. and Holmes, D. (1992). Modeling Population Uniqueness, *Proceedings of the International Seminar on Statistical Confidentiality*, International Statistical Institute, 175-199.

Sweeney, L. (1997). *Guaranteeing Anonymity When Sharing Medical Data, The Datafly System*, MIT Artificial Intelligence Laboratory Working Paper, Cambridge, 344.

Willenborg, L. and De Waal, T. (1996). *Statistical Disclosure Control in Practice*, New York: Springer-Verlag.

Woodward, B. (1996). Patient Privacy in a Computerized World, *1997 Medical and Health Annual 1997*, Chicago: Encyclopedia Britannica, Inc., 256-259.