Guaranteeing Anonymity when Sharing Medical Data, the Datafly System

Latanya Sweeney

Clinical Decision Making Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

We present a computer program named Datafly that maintains 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 and phone number, removed in the incorrect belief that patient confidentiality is maintained because the resulting data look anonymous; however, we show the remaining data can often 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 determined by the user.

INTRODUCTION

In 1996, TIME/CNN conducted a telephone poll of 406 adults in the United States[1] in which 88% replied that to the best of their knowledge, no medical information about themselves had ever been disclosed without their permission. In a second question, 87% said laws should prohibit organizations from giving out medical information without obtaining the patient's permission.

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[1]. The public is not likely to accept that their records are kept "confidential" if large numbers of people have access to their contents. As more HMO's and hospitals merge, the number of people with access to any particular record increases dramatically since most systems allow full access to all records by virtually all personnel. Beyond hospital systems there are records at insurance companies, consulting firms, pharmacies and elsewhere that include copies of all or part of this information.

As one would expect, there have been abuses, here are just a few. In 1995, Woodward[2] cited an alarming case of a Maryland banker who cross-referenced a list of patients with cancer against a list of people who had outstanding loans at his bank and then called in the loans. Linowes and Spencer[3] surveyed 87 Fortune 500 companies with a total of 3.2 million employees and found that 35% said they used medical records to make decisions about employees. The New York Times[4] reported cases of snooping by insiders in large hospital computer networks, even though the use of a simple audit trail, a record of each person who looked up a patient's record, could curtail such behavior.

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

Table 1. Data fields recommended by NAHDO.

In 1996, the National Association of Health Data Organizations (NAHDO) reported that 37 states had legislative mandates to gather hospital-level data[5]. Last year, 17 of these states reported they had started collecting ambulatory care data from physician offices, clinics, and so on. Table 1 contains a list of fields which NAHDO recommends these states accumulate. Many of these states have subsequently given copies of

1091—8280/97/$5.00

collected data to researchers and sold copies to industry since the data are incorrectly believed to be anonymous. The public would probably agree secondary parties should know some information buried in the record, but such disclosure should not risk identifying patients. The goal of this work is to provide tools for extracting needed information from medical records while maintaining patient confidentiality.

BACKGROUND

Last year, we presented the Scrub System[6] which locates and replaces personally-identifying information in unrestricted text. The Scrub System found 99-100% of identifying references, while the straightforward approach of global search-and-replace properly located no more than 30-60% of all such references. However, the Scrub System merely de-identifies information and cannot guarantee anonymity. In de-identified data, all explicit identifiers, such as name, address and phone number, are removed, generalized, or replaced with made-up alternatives. Anonymous, however, implies the data cannot be manipulated, matched or linked to identify any individual. Even when information shared with secondary parties is de-identified, it is 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.

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

Table 2. De-identified data that is not anonymous.

Most municipalities sell locally collected census data or voter lists that include the date of birth, name and address of each resident. This information can be linked to data 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 data 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. Clearly, the risks of re-identifying data depend both on the content of the released data and on related information available to the recipient.

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, especially since unusual cases are often unusual in other sources of data as well making them easier to identify.

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 the unique characteristic may be any little detail or combination of details available to the memory of a patient or a doctor, or knowledge about the data 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[7]. 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: (1) most medical databases are geographically located and so one can presume, for example, the ZIP codes of a hospital's patients; (2) 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, (3) most releases of medical data are not randomly sampled with small sampling fractions, but instead include most of the database.

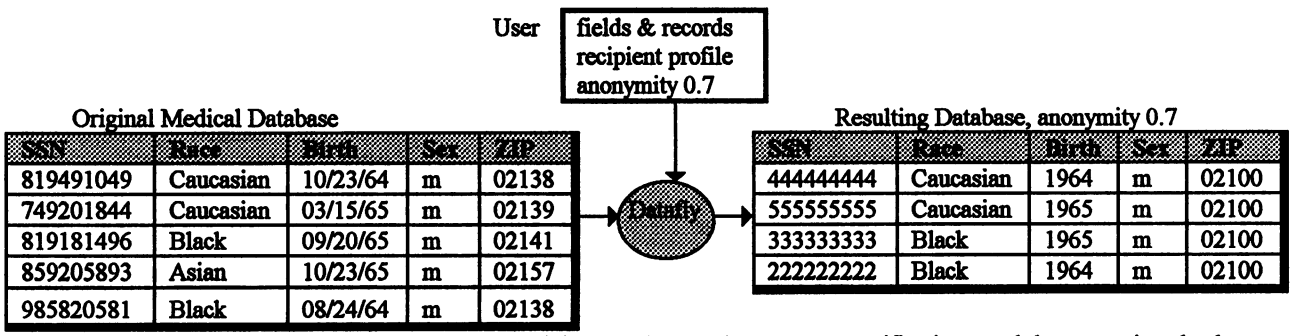
Determining the optimal bin size to ensure anonymity is tricky. It certainly depends on the frequencies of characteristics found within the data as well as within other sources for re-identification. In addition, the motivation and effort required to re-identify released data in cases where virtually 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.

METHODS

We constructed a computer program named Datafly that interfaces a user with an Oracle server, which in turn, accesses a medical database. Datafly was written using Symantec C and Oracle's Pro*C Precompiler. It processed all queries to the database. Diagram 1 provides a user-level overview. The original database appears on the left. A user requests specific fields and records and provides a profile of the person who is to receive the data and 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 unique Asian entry was removed; Social Security numbers were replaced with made-up alternatives; and birth dates were generalized to the year and ZIP codes to the first three digits. In the next paragraphs, we discuss the values the user provides.

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. 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, such as Social Security numbers.

Table 3 shows the relationship between bin sizes and selected anonymity levels using the Cambridge voters database. As the anonymity level 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 to the aggregate months shown. Outliers were excluded from the released data and their percentages of the total 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



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

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

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.

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.

Anonymity	BinSize	BirthDate	Drop%
1			
├.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			

Table 3. The slide bar highlights a 0.7 anonymity level for the birth date field of the Cambridge voters data.

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. As we discussed earlier, 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 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. 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

The database we used was a de-identified subset of a pediatric medical record system[8]. It consisted of 300 patient records; we were primarily concerned with the fields listed in Table 1. 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 were generalized using the International Classification of Disease (ICD-9) hierarchy and other semantic groupings. Geographic replacements for states and ZIP codes generalized to use regions and population size. Continuous variables, such as dollar amounts and clinical measurements, were treated as categorical values; however, their replacements were based on meaningful ranges; of course this was only done in cases where generalizing these fields was necessary.

Table 4 reports the sensitivity of each field within this database to Datafly's anonymity parameters. The values within each cell ranging from 0.1 to 0.8 are the minimum anonymity levels that adhered to the designated bin size. The columns labeled 0.1 to 0.9 correspond to increases in the required bin size due to an increase in the field's profile value or the overall anonymity level.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
bin size	3	6	9	12	15	18	21	24	27
Hosp#	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
SSN	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Gender	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
VisitDate	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Ethnicity	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.8
Diagnosis	0.0	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Birthdate	0.1	0.2	0.3	0.3	0.5	0.5	0.5	0.5	0.5

Table 4. Results as the overall anonymity level increases. Each cell reports the minimum anonymity level that achieved the required bin size.

Consider the first two rows which report on the number assigned to the patient by the hospital (Hosp#) and the patient's Social Security number (SSN). No matter what the required bin size, an anonymity level as low as 0.1 required the replacement algorithm for the SSN and Hosp# fields to provide replacement values. Since an SSN and Hosp# should be unique, this result is consistent with expectations. An anonymity level as low as 0.1 forced the fields associated with gender and visit date to drop their outliers. Both of these fields had few bins with lots of occurrences in each bin, so once the outliers were removed, no further generalization was necessary to achieve the higher bin sizes.

The ethnicity field also had few bins with lots of occurrences in each, but the distribution across bins was not even, so eventually further generalization was required. The original values found within the diagnosis field already adhered to the smallest bin size, but achieving bin sizes beyond that required generalizing the values as shown. Of the fields listed In Table 4, the birth date field was the most sensitive, having lots of bins with few values in each; so it is not surprising that generalization produced stepwise improvement.

DISCUSSION

We have demonstrated that the Datafly System offers a practical approach to maintaining confidentiality by providing the most general version of the data useful to the recipient in both administrative and research releases of data. We found in our own work that if we approach some hospitals as researchers, we must petition the hospital's internal review board (IRB) to state our intentions and methodologies, then they decide whether we get data and in what form; but if we approach these same hospitals as administrative consultants, data are given to us with no IRB review. The decision is made locally and acted on. Datafly helps enforce consistent policies and ensures each release provides the most practically specific data to the recipient. Since Datafly explicitly quantifies "trust" in the recipient, associated risk becomes clear but the remedy against abuse lies outside the Datafly System and resides in contracts, laws and policies.

Acknowledgments

The author thanks Beverly Woodward, Ph.D., for discussions, Professor Peter Szolovits for support, Isaac Kohane, M.D. Ph.D., for the use of his data, and Patrick Thompson and Sylvia Barrett for editorial suggestions. This work has been supported by Henry Leitner and Harvard University DCE and by Medical Informatics Training Grant 1-T15-LM07092 from the National Library of Medicine.

References

1. Woodward, B. Patient privacy in a computerized world. *1997 Medical and Health Annual 1997;* Chicago: Encyclopedia Britannica, Inc., 1996:256-259.
2. Woodward, B. The computer-based patient record and confidentiality. *The New England Journal of Medicine;* Boston: Massachusetts Medical Society, 1995:1419-1422.
3. Linowes, D. and Spencer, R. Privacy: the workplace issue of the '90s. *The John Marshall Law Review;* 23 (1990): 591-620.
4. Grady, D. Hospital files as open book. *The New York Times;* New York, March 12, 1997:C8.
5. National Association of Health Data Organizations. A guide to state-level ambulatory care data collection activities. Falls Church: 1996 (October).
6. Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. Cimino, J., ed. Proceedings, *American Medical Informatics Association.* Washington, DC: Hanley & Belfus, Inc, 1996:333-337.
7. Alexander, L. and Jabine, T. Access to social security microdata files for research and statistical purposes. *Social Security Bulletin.* 1978 (41) No. 8.
8. Kohane, I. Getting the Data In: Three-Year Experience with a Pediatric Electronic Medical Record System. Ozbolt J., ed. Proceedings, *Symposium on Computer Applications in Medical Care.* Washington, DC: Hanley & Belfus, Inc, 1994:457-461.