MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Working Paper No. AIWP-WP344b May 1997

Maintaining Patient Confidentiality When Sharing Medical Data Requires a Symbiotic Relationship Between Technology and Policy

Latanya Sweeney[1]

ABSTRACT

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; the greater the number of candidates per record, the more anonymous the data. We examine three general-purpose computer programs for maintaining patient confidentiality when disclosing electronic medical records: the Scrub System which locates personally-identifying information in letters between doctors and notes written by clinicians; the Datafly System which generalizes values based on a profile of the recipient at the time of access; and, the μ-Argus System which is becoming a European standard for disclosing public use data. Despite the possible effectiveness of these systems, completely anonymous data may not contain sufficient details for all uses, so care must be taken when released data can identify individuals and such care must be enforced by coherent policies and procedures.

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. A few years ago, in 1994, we surveyed some college students at Harvard and in Taiwan. We posed the question: "Does your school have the right to read your electronic mail?" Students at Harvard (18 of 19 or 95%) stated that Harvard had no right to read their electronic mail. They argued that electronic mail was like regular mail, and since Harvard had no right to read their regular mail, Harvard had no right to read their electronic mail. Taiwanese students on

[1] Email address: sweeney@mit.edu. This work was performed at the Laboratory for Computer Science, MIT. It is currently under review for publication and is reprinted here with special permission. Related publications by Sweeney can be found at <http://carrie.lcs.mit.edu/people/sweeney/publications.html>.

the other hand, voiced an opposing opinion (16 of 17 or 94%). They felt their electronic mail reflected the school, and so the school had every right to make sure students were behaving honorably.

These findings are not surprising since they mirror the ethical systems of these two societies. It seems we map old expectations onto new technical entities, believing the new version adheres to the same social contract. In the case of electronic medical records, the public's expectations may not be consistent with actual practice and the public may not be aware that their perceived social contract is tenuous.

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 personal medical information about themselves had ever been disclosed without their permission. In a second question, 87% said laws should be passed that prohibit health care organizations from giving out medical information without first obtaining the patient's permission.

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[2], and the integration of applicable data from multiple sources into a unified display for clinicians[3]; 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.

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 are not likely to accept that their records are kept "confidential" if large numbers of people have access to their contents. As more HMOs and hospitals merge, the number of people with authorized access often increases dramatically since most of these systems allow full access to all records by any authorized person[4]. For example, assume a billing clerk in hospital X can view all information in all medical records within the institution. When hospital X merges with hospitals Y and Z, the same clerk may then be able to view all records at all three hospitals even though there is no reason for the clerk to know information about the patients at the other institutions.

As one would expect, there have been many abuses. For example, in 1995, Woodward[5] 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[6] 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* reported cases of snooping by insiders in large hospital computer networks[7], even though the use of a simple audit trail, a list of each person who looked up a patient's record, could curtail such behavior[4].

Why are identified data so available? Most electronic medical records are really two medical records in one bundle. This duality came about primarily for historical reasons. In terms of the medical record, computers were first introduced as a billing system only, and the record was basically used and controlled by administrators. Compiled for remuneration from insurance companies, these records typically included diagnosis, procedure and medication codes along with the name, address, birth date, and Social Security number of each patient. Medical billing records today usually have more than 100 such fields of information per patient.

The clinical condition of the patient was maintained separately, in written form, by doctors and nurses. Some cite fear from legal retaliation and others the refusal to type on a computer keyboard as reasons for clinical information being maintained outside the computer system. In fact, even today, the "real" clinical record can often be found on index cards located in the doctor's pocket.

This trend is changing rapidly and more clinical information is routinely included in the electronic medical record, which has led to even more confusion in the social contract of patient confidentiality. In our own work, 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.

When the clinical record joins the billing record, is the resulting electronic medical record governed by administrators, who pass it along in part to independent consultants and outside agencies as the administrators deem appropriate? Or, is it governed by the doctor-patient confidentiality contract? Who governs the records maintained by insurance companies? Pharmaceutical companies run longitudinal studies on identified patients and providers. What happens when these records are bought and sold? What about individualized prescription records maintained by local drug stores? State governments are insisting on maintaining their own encounter-level records for cost analysis. Who should get copies and for what purposes? On the one hand, 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 paper is to examine available tools that extract needed information from medical records, so we may then elicit the accompanying policies needed to maintain a commitment to patient confidentiality.

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.

970202	4973251	n
970202	7321785	y
970202	8324820	n
970203	2018492	n
970203	9353481	y
970203	3856592	n

Table 1. Possibly anonymous HIV test data.

There are three major difficulties in providing anonymous data. One of the problems is that anonymity is in the eye of the beholder. Knowledge a viewer of the data may hold or bring to bear on the data is usually not known beforehand by the person producing the data and such knowledge may be useful in identifying patients. Consider an HIV testing center located in a heavily populated community within a large metropolitan area. If Table 1 shows the results for two days, then it may not appear very anonymous if the leftmost column is the date, the middle column is the patient's phone number, and the rightmost column holds the results. An electronic phone directory can match each phone number to a name and address. Although this does not identify the specific member of the household tested, the possible choices have narrowed to a particular address.

Alternatively, if the middle column in Table 1 holds random numbers assigned to samples, then identifying individuals becomes more difficult, but we still cannot guarantee the data are anonymous. If a person with inside knowledge (e.g., a doctor, patient, nurse, attendant or even a friend of the patient) recognizes a patient and recalls the patient was the second person tested that day, then the results are not anonymous to the insider. In a similar vein, medical records distributed with a provider code assigned by an insurance company are often not anonymous with respect to the provider, because hundreds of administrators typically have directories that link the provider's name, address and phone number to the assigned code.

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

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

As another 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 this table 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 year-round. The ZIP code 20612 may have only one Asian family. In these cases, information outside the data identifies the individuals.

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 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, census data are usually not very accurate in college towns and areas that have large transient communities, 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.

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

Table 3. Uniqueness of demographic fields in Cambridge voter list.

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.

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 a reporter, private investigator 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. In an interview, for example, the janitor may recall an Asian patient whose last name was Chan and who worked as a stockbroker because the patient gave the janitor some good investing tips. Any single uniquely occurring value or group of values can be used to identify an individual. Remember that the unique characteristic may not be known beforehand. It could 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.

As another example, consider the medical records of a pediatric hospital in which only one patient is older than 45 years of age. Suppose a de-identified version of the hospital's records is to be released for public-use that includes age and city of residence but not birth date or zip code. Many may believe the resulting data would be anonymous because there are thousands of people of age 45 living in that city. However, the rare occurrence of a 45 year-old pediatric patient at that facility can become a focal point for anyone seeking to discredit the

anonymity of the data. Nurses, clerks and other hospital personnel will often remember unusual cases and in interviews may provide additional details that help identify the patient.

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

Table 4. Sample database in which Asian is a uniquely identifying characteristic.

As a final example, 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. If her clinical data contains sensitive information about medical complications, then any release of clinical data containing her record may identify her and provide additional information about her medical condition even though the released data may not contain any references to her age or residence. When releasing data for public and semi-public use, records containing notable characteristics must be suppressed or masked.

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[8]. 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 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. Unfortunately, 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 and combinations of characteristics found in the general population because it contains almost all the general population; small, specialized databases, however, must estimate these values. In the next section, we will present computer programs that generalize data based on bin sizes and estimates. Following that, we will report results using these programs and discuss their limitations and the need for complementary policies.

METHODS

There are many possible tools for maintaining confidentiality when disclosing medical data such as changing singletons to median values, inserting complementary records, generalizing codes, swapping entries, scrambling records, suppressing information and encrypting fields. Which technique, or combination of techniques, is best to use depends on the nature of the data and its intended use, but these techniques are narrowly focused and little literature exists concerning their use with medical data. The three systems presented here are among the few complete architectures currently available for use. Not only do they provide effective solutions but they also help us understand many of the underlying issues. The Scrub system locates and replaces personally-identifying information in letters and notes. The Datafly System generalizes database information to satisfy bin size requirements based on a profile of the recipient. And, the μ-Argus System generalizes information for disclosing public use data. We will now examine each of these in turn and then discuss their limitations.

The Scrub System.

Last year, Sweeney presented the Scrub System[9] which locates and replaces personally identifying information in text documents and in textual fields of the database. A close examination of two different computer-based patient record systems[10,11] quickly revealed that much of the medical content resided in the letters between physicians and in the shorthand notes of clinicians since this is where providers discussed findings, explained current treatment and furnished an overall view of the medical condition of the patient.

At present, most institutions have few releases of data that include these notes and letters, but new uses for this information is increasing, and therefore, the desire to release this text is also increasing. After all, these letters and notes are a valuable research tool and can corroborate the rest of the record. The fields containing the diagnosis, procedure and medication codes when examined alone can be incorrect or misleading. A prominent physician stated at a recent conference that he purposefully places incorrect codes in the diagnosis and procedure fields when such codes would reveal sensitive information about the patient. Similarly, the diagnosis and procedure codes are often up-coded for billing purposes. If these practices become widespread, they will render the administrative medical record useless for clinical research and may already be problematic for retrospective investigation. Clinical notes and letters may prove to be the only reliable artifacts.

The Scrub System provides a methodology for removing personally identifying information in medical writings so that the integrity of the medical information remains intact even though the identity of the patient remains confidential. This process is termed "scrubbing." Protecting patient confidentiality in raw text is not as simple as searching for the patient's name and replacing all occurrences with a pseudo name. References to the patient are often quite obscure, consider for example, "he developed Hodgkins while acting as the U.S. Ambassador to England and was diagnosed by Dr. Frank at Brigham's." Clinicians write text with little regard to word-choice and in many cases without concern to grammar or spelling. While the resulting "unrestricted text" is valuable to understanding the medical condition and treatment of the patient, it poses tremendous difficulty to scrubbing since the text often includes names of other care-takers, family members, employers and nick names.

Table 5 shows a sample letter and its scrubbed result. Actual letters are often several pages in length. In the case of clinical notes, the recorded messages are often cryptic abbreviations specific to the institution or known only among a group of physicians within the facility. The traditional approach to scrubbing is straightforward search and replace which misses these references.

The Scrub System was modeled after a human approach to the problem. It uses templates and localized knowledge to recognize personally-identifying information. In fact, the Scrub work showed that the recognition of personally-identifying information is strongly linked to the common recording practices of society. For example, Fred and Bill are common first names and Miller and Jones are common last names and knowing these facts makes it easier to recognize them as likely names. Common facts along with their accompanying templates of use are considered commonsense knowledge and the itemization and use of commonsense knowledge is the backbone of Scrub.

Wednesday, February 2, 1994

Marjorie Long, M.D.
St. John's Hospital
Huntington 18
Boston, MA 02151

RE: Virginia Townsend
CH#32-841-09787
DOB 05/26/86

Dear Dr. Lang:

I feel much better after seeing Virginia this time. As you know, Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins in a week at Marina Corp 473-1214 to schedule a follow-up visit for her daughter.

Patrick Hayes, M.D. 34764

Wednesday, February 2, 1994

Marjorie Long, M.D.
St. John's Hospital
Huntington 18
Boston, MA 02151

RE: *Kathel Wallams*
CH#*18-512-32871*
DOB 05/26/86

Dear Dr. Lang:

I feel much better after seeing *Kathel* this time. As you know, Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins in a week at Marina Corp 473-1214 to schedule a follow-up visit for her daughter.

Mank Brones, M.D. *21075*

February, 1994

Erisa Cosborn, M.D.
Brighaul Hospital
Alberdam Way
Peabon, MA *02100*

RE: *Kathel Wallams*
CH#*18-512-32871*
DOB *05/86*

Dear Dr. *Jandel*:

I feel much better after seeing *Kathel* this time. As you know, *Cob* is a 7 and 6/12 year old female in follow-up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. *Wandel* at *Namingham*'s. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the . We will contact Mrs. *Learl* in a week at *Garlaw* Corp *912-8205* to schedule a follow-up visit for her daughter.

Mank Brones, M.D. *21075*

Table 5. Sample letter reporting back to a referring physician. On the top is a made-up original containing the name and address of the referring physician, a typo in the salutation line, the patient's nick name, and references to another care-taker, the patient's athletic team and mother and her mother's employer and phone number. On the lower left is the result from simple search and replace and on the right is the result from the Scrub System. Notice in the Scrub System that the name of the medication remained but the mother's last name was correctly replaced. The reference "U.S. Junior Gymnastics team" was suppressed since Scrub was not sure how to replace it.

The Scrub System accurately found 99-100% of all personally-identifying references in more than 3,000 letters between physicians, while the straightforward approach of global search-and-replace properly located no more than 30-60% of all such references[9]. The higher figure for search and replace includes using additional information stored in the database to help identify the attending physician's name, identifying number and other information. Since the letters were properly formatted, the heading block was easily detected and compositional cues were available using keywords like "Dear." This dramatically improved the results of the search-and-replace method[9] to around 84%; however, most references to family members, additional phone numbers, nick names and references to the physician receiving the letter were still not detected, whereas Scrub was able to correctly identify and replace these instances. However, the Scrub System merely de-identifies information and cannot guarantee anonymity. Even though all explicit identifiers such as name, address and phone number are removed or replaced, it may be possible to infer the identify of an individual. Consider the following text.

> "At the age of two she was sexually assaulted. At the age of three she set fire to her home. At the age of four her parents divorced. At the age of five she was placed in foster care after stabbing her nursery school teacher with scissors."

If her life continues to progress in this manner, by the age of eight she may be in the news, but nothing in this text required scrubbing even though there would probably exist only one such child with this history. An overall sequence of events can provide a preponderance of details that identify an individual. This is often the case in mental health data and discharge notes.

The Datafly System.

Although Scrub reliably de-identifies clinical notes and letters, the greatest volume of medical data found outside the originating institution flows from administrative billing records, which Scrub does not address. In 1996, the National Association of Health Data Organizations (NAHDO) reported that 37 states had legislative mandates to gather hospital-level data[12]. Last year, 17 of these states reported they had started collecting ambulatory care (outpatient) data from hospitals, physician offices, clinics, and so on. Table 6 contains a list of the fields of information which NAHDO recommends these states accumulate. Many of them have subsequently given copies to researchers and sold copies to industry. As stated earlier, there are many other sources of administrative billing records with similar fields of information. What remains more alarming of course is that most of these de-identified records can be re-identified since patient demographics and other fields often combine uniquely to identify individuals.

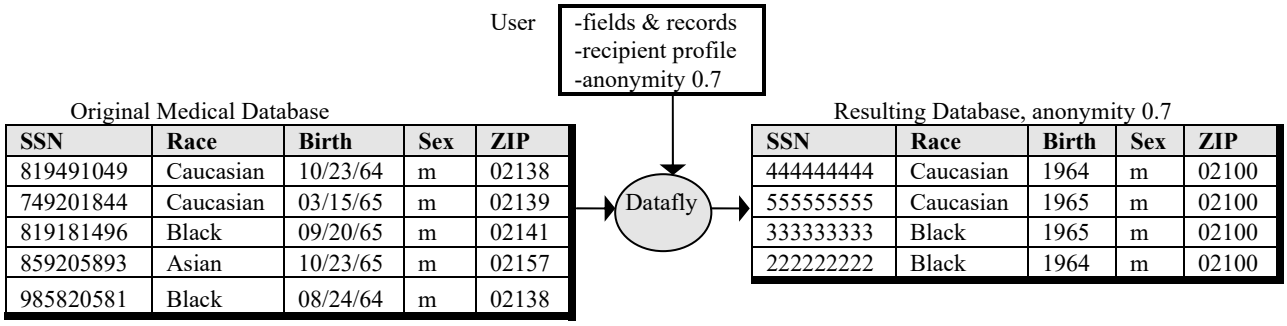
Earlier this year Sweeney presented the Datafly System[13] whose goal is to provide the most general information useful to the recipient. Datafly maintains anonymity in medical data by automatically aggregating, substituting and removing information as appropriate. Decisions are made at the field and 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.

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 6. Data fields recommended by NAHDO for state collection of ambulatory data.

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 two paragraphs we examine the overall anonymity level and the profile of the recipient, both of which the user provides when requesting data.



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.

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 anonymity level to actual bin sizes though Sweeney[14] 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 7 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.

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.

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 7. Anonymity generalizations for Cambridge voters data with corresponding bin sizes. 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.

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.

Numerous tests were conducted using the Datafly System to access a pediatric medical record system[14]. 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 6, for approximately 135,000 patients consisting of state employees and their families[15]. 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.

The μ-Argus System.

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[16]. 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.[17]

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 all 2- or 3-combinations across all fields. 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; this process is called cell suppression[18]. 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 8a there are many Caucasians and many females, but only one female Caucasian in the database. Tables 8b and 8c 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 8c.

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.

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 8a. There is only one Caucasian female, even though there are many females and Caucasians.

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

Table 8b. Results from applying the Datafly System to the data in Table 8a. 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.

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

Table 8c. Results from applying the approach of the μ-Argus system to the data in Table 8a. 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.

We will briefly compare the results of these two systems, but for a more in-depth discussion, see Sweeney[14]. In the Datafly System, the generalization across a subset of sensitive fields ensures that the combination across those fields will adhere to the minimal bin size. This is demonstrated in Table 8b. The μ-Argus program however, only checks 2 or 3 combinations; there may exist unique combinations across 4 or more fields that would not be detected. For example, Table 8c 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 data. Further, since the number of fields, especially demographic fields, in a medical database is large, this may prove to be a serious handicap when using the μ-Argus system with medical data.

DISCUSSION

The Scrub System demonstrated that medical data, including textual documents, can be de-identified, but as we have shown de-identification alone is not sufficient to protect confidentiality. Not only can de-identified information often be re-identified by linking data to other databases, but also releasing too many patient-specific facts can identify individuals. Unless we are proactive, the proliferation of medical data may become so widespread that it will be impossible to release medical data without further breaching confidentiality. For example, the existence of rather extensive registers of business establishments in the hands of government agencies, trade associations and firms like Dunn and Bradstreet has virtually ruled out the possibility of releasing database information about businesses [18].

The Datafly and μ-Argus systems illustrated 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[19].

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[20]. 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, the Datafly and μ-Argus 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.

What is needed is a contractual arrangement between the recipient and the producer to make the trust explicit and share the risk. Table 9 contains some guidelines that make it clear which fields need to be protected against linking. Using this additional knowledge and the techniques presented in the Datafly and μ-Argus systems, the producer can best protect the anonymity of patients in data even when sensitive information is released. It is surprising that in most releases of medical data there are no contractual arrangements to limit further dissemination or use of the data. Even in cases where there is an IRB review, no contract usually results. Further, 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 technology and statistical disclosure techniques and resides instead in contracts, laws and policies.

1. 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.

2. The recipient must be strictly and legally accountable to the producer for the security of the data and must demonstrate adequate security protection.

3. The data must be de-identified. The release must contain no explicit individual identifiers nor should it contain data that would be easily associated with an individual.

4. 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 also identify those fields for which the recipient will link the data. If such linking identifies patients, then patient consent may be warranted.

5. The data provider should have the opportunity to review any publication of information from the data to insure that no potential disclosures are published.

6. At the conclusion of the project, and no later than some specified date, the recipient must destroy all copies of the data.

7. The recipient must not give, sell, loan, show or disseminate the data to any other parties.

Table 9. Contractual requirements for restricted use of data based on federal guidelines and the Datafly System.

In closing we consider an alternative to autonomous database systems since the burden of determining the risk of disclosure may appear cumbersome. Suppose instead we had a centralized federal repository for medical data. Though institutions and businesses could maintain their own data for internal purposes, they could not sell or give data away in any form, except of course for disclosure to the federal repository, remuneration for services and required reporting. The recipients of these data would, in turn, be equally restricted against further dissemination. The trusted authority that maintains the central repository would have nearly perfect omniscience and could confidently release data for public use. Questions posed by researchers, administrators and others could be answered without releasing any data; instead the trusted authority would run desired queries against the data and then provide non-compromising results to the investigators. In releases of de-identified data, the exact risk could be computed and accompanying penalties for abuse incorporated into the dissemination process. While this type of system may have advantages to maintaining confidentiality, it requires a single point of trust or failure. Current societal inclinations suggest that the public would not trust a sole authority in such a role and would feel safer with distributed, locally controlled data. Ironically, if current trends continue, a handful of independent information brokers may assume this role of the trusted authority anyway. If information brokers do emerge as the primary keepers of medical data, then they may eventually rank among the most conservative advocates for maintaining confidentiality and limiting dissemination, since their economic survival would hinge on protecting what would be their greatest asset, our medical records.

Acknowledgments

The author is grateful to God for the opportunity to present this work because the journey that made it possible was remarkable. The author thanks Beverly Woodward, Ph.D., for many discussions and comments. The author also thanks Professor Peter Szolovits at MIT for providing an environment that made it possible for me to explore my own ideas and Patrick Thompson and Sylvia Barrett for editorial suggestions. We also acknowledge the continued support of Henry Leitner and Harvard University. This work has been supported by a 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. Cooper, G., et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine,1997*.
3. Kohane, I., et al. Sharing electronic medical records across heterogeneous and competing institutions. In: Cimino, J., ed. Proceedings*, American Medical Informatics Association.* Washington, DC: Hanley & Belfus, Inc, 1996:333-337.
4. Clayton, P., et al. Protecting electronic health information. National Research Council. Washington, DC: National Academy Press, 1997.
5. Woodward, B. The computer-based patient record and confidentiality. *The New England Journal of Medicine;* Boston: Massachusetts Medical Society, 1995:1419-1422.
6. Linowes, D. and Spencer, R. Privacy: the workplace issue of the '90s. *The John Marshall Law Review;* 23 (1990): 591-620.
7. Grady, D. Hospital files as open book. *The New York Times;* New York, March 12, 1997:C8.
8. Alexander, L. and Jabine, T. Access to social security microdata files for research and statistical purposes. *Social Security Bulletin.* 1978 (41) No. 8.
9. Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. In: Cimino, J., ed. Proceedings*, American Medical Informatics Association.* Washington, DC: Hanley & Belfus, Inc, 1996:333-337.
10. Kohane, I. Getting the data in: three-year experience with a pediatric electronic medical record system. In: Ozbolt J., ed. Proceedings*, Symposium on Computer Applications in Medical Care.* Washington, DC: Hanley & Belfus, Inc, 1994:457-461.
11. Barnett, G. The application of computer-based medical-record systems in ambulatory practice. *New England Journal of Medicine.* 1984;310(25):1643-1650.
12. A guide to state-level ambulatory care data collection activities. *National Association of Health Data Organizations.* Falls Church: 1996 (October).
13. Sweeney, L. Computational disclosure control for medical microdata, the Datafly System. *Bureau of the Census, Record Linkage Bulletin.* Washington, DC: Bureau of the Census, 1997.
14. Sweeney, L. Guaranteeing anonymity when sharing medical data, the Datafly system. Proceedings*, American Medical Informatics Association.* Nashville: Hanley & Belfus, Inc, 1997.
15. Michael Lasalandra, Panel told releases of med records hurt privacy. *Boston Herald*. Boston: March 20, 1997 (35).

16. Hundepool, A. and Willenborg, L. μ- and τ-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality.* Bled: 1996.
17. Willenborg, L. and De Waal, T. *Statistical disclosure control in practice.* New York: Springer-Verlag, 1996.
18. Kirkendall, N. et al. Report on statistical disclosure limitation methodology. *Statistical Policy Working Paper.* Washington: Office of Management and Budget, 1994 (No. 22).
19. Duncan, G. and Lambert, D. The risk of disclosure for microdata. *Proceedings of the Bureau of the Census Third Annual Research Conference.* Washington: Bureau of the Census, 1987.
20. Skinner, C. and Holmes, D. Modeling population uniqueness. *Proceedings of the International Seminar on Statistical Confidentiality.* International Statistical Institute, 1992:175-199.