Three Computational Systems for Disclosing Medical Data in the Year 1999

Latanya Sweeney

Laboratory for Computer Science, Massachusetts Institute of Technology, USA

Abstract

Today most 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. We examine three computer programs that do maintain 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 data within the record 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. The techniques presented in these systems help protect confidentiality in the face of a changing globally-networked society with immediate access to volumes of personal data.

Keywords:

Computational disclosure control; Confidentiality; Medical record linkage; Computer security; Databases

Introduction

The proliferation of data through immediate electronic means poses a tremendous challenge to sharing medical records while maintaining patient confidentiality. Even in countries like Canada and the Netherlands, that invoke centralized control over the collection and release of medical data, growing demands from researchers, policy makers and others for more and more detailed information threaten confidentiality and trust [1]. In the United States where medical data is autonomously collected and controlled, the challenge is even more precarious since public expectations may not be consistent with actual practice [2] and there have been many abuses [3].

Table 1 contains a list of fields which are commonly collected and distributed to companies, researchers, economists, policy makers and administrators. Unfortunately, as we will show, the data are incorrectly believed to be anonymous. The goal of this work is to present tools for extracting needed information from medical records while maintaining patient confidentiality.

Table 1. Data fields commonly distributed.

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

Background

We begin by stating our definitions of de-identified and anonymous data. In de-identified data, all explicit identifiers, such as name, address and phone number, are removed 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, 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. Population registers, such as local census data, voter registration lists, city directories, as well as information from motor vehicle agencies, tax assessors, real estate agencies and the World Wide Web are publicly available and often include a postal code and birth date along with the accompanying name and address. These registers can be linked and matched to the fields in Table 1 to identify patients. For example, Table 2 shows which fields in Table 1 were used to uniquely identify the names and addresses of individuals in the 1997 voting list for Cambridge, Massachusetts USA. 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. 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.

Table 2. Demographic uniqueness in Cambridge voting list.

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

Measuring the degree of anonymity in released data poses a third problem when producing anonymous data for practical use. The United States 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 [4]. 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

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

The Scrub System provides a methodology for removing personally identifying information in text documents and in textual fields of the database so that the integrity of the medical information remains intact even though the identity of the patient remains confidential [5]. This process is termed "scrubbing." A close examination of two different computer-based patient record systems 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.

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 3 shows a sample letter to a referring physician and its scrubbed result. Actual letters are often several pages in length. In 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 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 [5]. 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 identity of an individual. 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

The Datafly System [6] concerns the release of field-structured records and provides the most general information useful to the recipient by automatically generalizing, substituting and removing information as appropriate. 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. 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 since each record matches as many people as the user had specified.

The overall anonymity level provided by the user 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 unique identifiers.

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. 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.

Table 3. On the left is a sample letter to a referring physician that contains the name and address of the referring physician, a typo in the salutation line, the patient's nick name, references to another care-taker, the patient's school and mother and her mother's employer and phone number. On the right is the result from the Scrub System. Notice 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.

Wednesday, February 2, 1994 Marjorie Long, M.D. RE: Virginia Townsend St. John's Hospital CH#32-841-09787 Huntington 18 DOB 05/26/86 Boston, MA 02151 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

February, 1994 *Erisa Cosborn*, M.D. RE: *Kathel Wallams* *Brighaul* Hospital CH#*18-512-32871* *Alberdam Way* DOB *05/86* *Peabon*, MA *02100* 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*

Numerous tests were conducted using the Datafly System to access a pediatric medical record system [6]. 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 and other groupings. 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.

Table 4. Anonymity generalizations for Cambridge voters data.

Anonymity	BinSize	BirthDate	Drop%
1			
Ã .9 ´	493	24	4%
Ã .8 ´	438	24	2%
Ã .7 ´	383	12	8%
Ã .6 ´	328	12	5%
Ã .5 ´	274	12	4%
Ã .4 ´	219	12	3%
Ã .3 ´	164	6	5%
Ã .2 ´	109	4	5%
Ã .1 ´	54	2	5%
0			

Table 4 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 in months as shown. Outliers were excluded from the released data and their corresponding percentages of the total are noted. The user sets the anonymity level, as depicted by the slide bar at 0.7 in Table 4. This setting 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.

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 a first version of a program named µ-Argus that seeks to accomplish this goal [1]. The µ-Argus program is considered the official confidentiality software of the European community even though Statistics Netherlands admittedly considers this first version a rough draft.

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.

Table 5a. There is only one Caucasian female.

SSN	Ethnicity	Birth		ZIP	Problem
819181496	Black	9/2/65	m	02141	short breath
195925972	Black	2/1/65	m	02141	chest pain
902750852	Black	1/8/65	f	02138	hypertension
985820581	Black	8/4/65	f	02138	hypertension
209559459	Black	1/7/64	f	02138	obesity
679392975	Black	2/4/64	f	02138	chest pain
819491049	Caucasian	1/5/64	m	02138	chest pain
749201844	Caucasian	3/1/65	f	02139	hypertension
985302952	Caucasian	8/3/64	m	02139	obesity
874593560	Caucasian	5/5/64	m	02139	short breath
703872052	Caucasian	2/6/67	m	02138	chest pain
963963603	Caucasian	3/9/67	m	02138	chest pain

In Table 5a there are many Caucasians and many females, but only one female Caucasian in the database. Tables 5b and 5c 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 5c. 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 and those that contain the "more identifying" fields. 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.

Table 5b. Results from applying the Datafly System to Table 5a. The minimum bin size is 2. The profile identifies only the demographic fields as being likely for linking. The Caucasian female record was dropped as an outlier.

SSN	Ethnicity	Birth		ZIP	Problem
902387250	Black	1965	m	02140	short 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	short breath
410968224	Caucasian	1967	m	02130	chest pain
664545451	Caucasian	1967	m	02130	chest pain

Table 5c. Results from applying the µ-Argus system to Table 5a. The minimum bin size is 2. The profile for fields was: SSN, "most

identifying;" birth, sex and ZIP, "more identifying;" and, ethnicity, "identifying." The uniqueness of the Caucasian female is suppressed; but, there remains a unique record for the Caucasian male born in 1964 in 02138.

SSN	Ethnicity	Birth		ZIP	Problem
	Black	1965	m	02141	breath shortness
	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	breath shortness
	Caucasian	1967	m	02138	chest pain
	Caucasian	1967	m	02138	chest pain

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 appears as Table 5c.

In comparing the results of these two systems, 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 5c 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.

Discussion

The Scrub System demonstrated that textual medical documents, can be de-identified, but de-identification alone is not sufficient to protect confidentiality. 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 and we can even provide anonymous data for public use. However, 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. Still, these systems offer us a good start in facing the challenges of sharing medical information in a globally-networked society.

Acknowledgments

The author is grateful to God for the opportunity to present this work. The author thanks Beverly Woodward, Ph.D., for discussions, Professor Peter Szolovits for support and Sylvia Barrett for editorial suggestions. This work was supported by Medical Informatics Training Grant 1-T15-LM07092 from the National Library of Medicine.

References

[1] Hundepool, A. and Willenborg, L. μ- and τ-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality.* Bled: 1996.

[2] Woodward, B. Patient privacy in a computerized world. *1997 Medical and Health Annual 1997;* Chicago: Encyclopedia Britannica, Inc., 1996:256-259.

[3] Woodward, B. The computer-based patient record and confidentiality. *The New England Journal of Medicine;* Boston: Mass. Medical Society, 1995:1419-1422.

[4] Alexander, L. and Jabine, T. Access to social security microdata files for research and statistical purposes. *Social Security Bulletin.* 1978 (41) No. 8.

[5] Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. Proceedings, *American Medical Informatics Association.* Washington: Hanley & Belfus, 1996.

[6] Sweeney, L. Guaranteeing anonymity when sharing medical data, the datafly system. Proceedings, *American Medical Informatics Association.* Nashville: Hanley & Belfus, Inc, 1997.

Address for correspondence

Email: sweeney@medg.lcs.mit.edu
URL: http://www.medg.lcs.mit.edu/people/sweeney/