Weaving Technology and Policy Together to Maintain Confidentiality

Latanya Sweeney

Organizations often release and receive medical data with all explicit identifiers, such as name, address, telephone number, and Social Security number (SSN), removed on the assumption that patient confidentiality is maintained because the resulting data look anonymous. However, in most of these cases, the remaining data can be used to reidentify individuals by linking or matching the data to other data bases or by looking at unique characteristics found in the fields and records of the data base itself. When these less apparent aspects are taken into account, each released record can map to many possible people, providing a level of anonymity that the record-holder determines. The greater the number of candidates per record, the more anonymous the data.

I examine three general-purpose computer programs for maintaining patient confidentiality when disclosing electronic medical records: the Scrub System, which locates and suppresses or replaces personally identifying information in letters between doctors and in notes written by clinicians; the Datafly System, which generalizes values based on a profile of the data recipient at the time of disclosure; and the μ-Argus System, a somewhat similar system which is becoming a European standard for disclosing public use data. These systems have limitations. Even when they are completely effective, wholly anonymous data may not contain sufficient details for all uses; hence, care must be taken when released data can identify individuals and such care must be enforced by coherent policies and procedures.

Background

Identifiable personal health information is any information concerning a person's health or treatment that enables someone to identify that person. The expression *personal health information* refers to health information that may or may not identify individuals. As I will show, in many releases of personal health information, individuals can be recognized. *Anonymous personal health information*, by contrast, contains details about a person's medical condition or treatment but the identity of the person cannot be determined.

In general usage, confidentiality of personal information protects the interests of the organization while privacy protects the autonomy of the individual; but, in medical usage, both terms mean privacy. The historical origin and ethical basis of medical confidentiality begins with the Hippocratic Oath, which was written between the sixth century B.C. and the first century A.D. It states:

> Whatsoever I shall see or hear in the course of my dealings with men, if it be what should not be published abroad, I will never divulge, holding such things to be holy secrets.

Various professional associations world-wide reiterate this oath, and by pledging this oath, clinicians—licensed professionals such as doctors, nurses, pharmacists, radiologists, and dentists who access in the line of duty identifiable personal health information—assume the responsibility of securing this information. The resulting trust is the cornerstone of the doctor-patient relationship, allowing patients to communicate with their physicians and to share information regarding their health status. However, the doctor-patient *privilege* offers no real protection to patients regarding the confidentiality of their health information. Legal protection is very narrow, only applying in some cases when a physician is testifying in court or in related proceedings.

Journal of Law, Medicine & Ethics, 25 (1997): 98–110.

The role of information technology is critical to confidentiality. On the one hand, information technology offers comprehensive, portable electronic records that can be easily accessed on behalf of a given patient no matter where or when a patient may need medical care.[1] That very portability, on the other hand, makes it much easier to transmit quickly and cheaply records containing identifiable personal health information widely and in bulk, for a variety of uses within and among health care institutions and other organizations and agencies. The Office of Technology Assessment (OTA) found that current laws generally do not provide consistent or comprehensive protection of personal health information.[2] Focusing on the impact of computer technology, OTA concluded that computerization reduces some concerns about privacy of personal health information while increasing others.

Previous policy efforts to protect the privacy of personal health information were limited to decisions about who gets access to which fields of information. I examine here three new computer programs that attempt to disclose information in such a way that individuals contained in the released data cannot be identified. These programs provide a spectrum of policy options. Decisions are no longer limited to who gets what information, but to how much generality or possible anonymity will exist in the released information.

The public's concern about the confidentiality of personal health information is reflected in a 1993 poll conducted by Harris and Associates for Equifax. The results of the survey found that 96 percent of respondents believe federal legislation should designate all personal health information as sensitive and impose severe penalties for unauthorized disclosure. Eighty percent of respondents were worried about medical record privacy, and 25 percent had personal experience of abuse related to personal health information.[3]

A 1994 Harris-Equifax consumer privacy survey focused on how the American public feels about having their medical records used for medical research and how safeguards would affect their opinions about such systems and uses. Among a list of thirteen groups and organizations, doctors and nurses ranked first in terms of the percentage of Americans who were "very" confident (43 percent) that this group properly handled personal and confidential information. After hearing a description about how medical records are used by researchers to study the causes of disease, 41 percent of those surveyed said they would find it at least somewhat acceptable if their records were used for such research. If a federal law made it illegal for any medical researcher to disclose the identity or any identifiable details of a person whose health records had been used, 28 percent of those who initially opposed having their records used would change their position. This would increase acceptance of this practice to over half those surveyed (58 percent).[4] By extension, this survey implies strong public support for releases of personal health information in which persons contained in the information cannot be identified at all.

Analysis of the detailed information contained within electronic medical records promises many social advantages, including improvements in medical care, reduced institutional costs, the development of predictive and diagnostic support systems,[5] and the integration of applicable data from multiple sources into a unified display for clinicians;[6] 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 that these secondary parties should know some of the information buried in the record, but such disclosure should not risk identifying patients.

Beverly Woodward makes a compelling argument that to the public, patient confidentiality implies that only people directly involved in one's health care will have access to one's medical records and that these health professionals will be bound by strict ethical and legal standards that prohibit further disclosure;[7] the public is not likely to accept the notion that records are "confidential" if large numbers of people have access to their contents. In 1996, the National Association of Health Data Organizations (NAHDO) reported that thirty-seven states had legislative mandates to gather electronically copies of personal health information from hospitals[8] for cost-analysis purposes. Community pharmacy chains, such as Revco, maintain electronic records for over 60 percent of the 2.4 billion outpatient prescriptions dispensed annually. Insurance claims typically include diagnosis, procedure, and medication codes along with the name, address, birth date, and SSN of each patient. Pharmaceutical companies run longitudinal studies on identified patients and providers. As more health maintenance organizations and hospitals merge, the number of people with authorized access to identifiable personal health information will increase dramatically because, as the National Research Council (NRC) recently warned, many of these systems allow full access to all records by any authorized person.[9] For example, assume a billing clerk at hospital X can view all information in all medical records within the institution. When hospital X merges with hospitals Y and Z, that same clerk may then be able to view all records at all three hospitals even though the clerk may not need to know information about the patients at the other institutions.

The NRC report also warns against inconsistent practices concerning releases of personal health information. If I approach a hospital as a researcher, I must petition the hospital's institutional review board (IRB) and state my intentions and methodologies; then the IRB decides whether I get data and in what form. But, if I approach the same hospital as an administrative consultant, data are given to

me without IRB review. The decision is made locally and acted on.

Recent presentations by the secretary of the Department of Health and Human Services emphasize the threats to privacy stemming from misuse of personal health information.[10] There have been abuses; here are just a few. A banker cross-referenced a list of patients with cancer against a list of people who had outstanding loans at his bank. Where he found matches, he called in the outstanding loans.[11] A survey of 87 Fortune 500 companies with a total of 3.2 million employees found that 35 percent of respondents used medical records to make decisions about employees.[12] Cases have been reported of snooping in large hospital computer networks by hospital employees,[13] 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.[14] *Consumer Reports* found that 40 percent of insurers disclose personal health information to lenders, employers, or marketers without customer permission.[15] Abuses like the preceding underscore the need to develop safeguards.

Data and anonymity

I begin by stating definitions of *deidentified data* and *anonymous data*. In deidentified data, all explicit identifiers, such as SSN, name, address, and telephone number, are removed, generalized, or replaced with a made-up alternative. Deidentifying data does not guarantee that the result is anonymous. The term *anonymous* implies that the data cannot be manipulated or linked to identify an individual. Even when information shared with secondary parties is deidentified, it is often far from anonymous.

There are three major difficulties in providing anonymous data. The first problem is that anonymity is in the eye of the beholder. The knowledge a viewer of the data may hold or bring to bear on the data is usually not known beforehand by the person releasing the data, and such knowledge may be useful in identifying individuals. 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 contains the date, the middle column contains the patient's telephone number, and the rightmost column holds the results. An electronic telephone 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 been narrowed to a particular address.

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.

Alternatively, if the middle column in Table 1 holds random numbers assigned to samples, then identifying individuals becomes more difficult; nonetheless, one still cannot guarantee the data are anonymous. If a person with inside knowledge (for example, a doctor, patient, nurse, attendant, or even a friend of the patient) recalls who was the second person tested that day, then the results are not anonymous to the insider. Similarly, 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 telephone number to the assigned code.

For another example, consider Table 2. If the contents of this table are a subset of an extremely large and diverse data base, then the three records may appear anonymous. Suppose the ZIP code 33171 primarily consists of a retirement community. A logical inference is that few young people live there. Likewise, 02657 is the postal code for Provincetown, Massachusetts, where about five black women live year-round. The ZIP code 20612 may contain only one Asian family. In these cases, information outside the data identifies the individuals.

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

Table 2. Deidentified Data that Are Not Anonymous.

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 patients' names, SSNs, and addresses are not present. 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 reidentify deidentified data because 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, which contains the month, day, and year of birth, alone can uniquely identify the name and address of 12 percent of the voters. One can identify 29 percent of the list by just birth date and gender, 69 percent with only a birth date and a 5-digit ZIP code, and 97 percent (53,033 vot-

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

Table 3. Uniqueness of Demographic Fields in Cambridge, Massachusetts, Voter List.

ers) when the full postal code and birth date are used. These values are listed in Table 3. Clearly, the risks of reidentifying data depend both on the content of the released data and on related information available to the recipient.

The second problem in 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, or others to discredit the anonymity of the released data, even when these instances are not unique in the general population. And, unusual cases are often also unusual in other sources of data, making them easier to identify. Consider the data base in Table 4. It is not surprising that the SSN is uniquely identifying, or, given the size of the data base, that the birth date is unique. To a lesser degree, the ZIP codes in Table 4 identify individuals because they are almost unique for each record. What may not have been known without closer examination of the particulars of this data base is that the designation of Asian as a race is uniquely identifying. In an interview, for example, a janitor may recall an Asian patient whose last name was Chan and who worked as a stockbroker, because that 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 minor detail or combination of details available to the memory of a patient or a doctor, or knowledge about the data base from some other source.

As another example, consider the medical records of a pediatric hospital in which only one patient is older than forty-five years. Suppose a deidentified 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 would believe the resulting data is anonymous because thousands of people age forty-five live in that city. However, the rare occurrence of a forty-five-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 Date*	*Sex*	*ZIP Code*
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 Data Base in which Asian is a Uniquely Identifying Characteristic.
* Social Security number.

As a final example, suppose a hospital's maternity records contain 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 contained in 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.

The third problem concerns measuring the degree of anonymity in released data 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, though some may contain regional or size of place designators.[16] SSA recognizes that data containing individuals with unique combinations of characteristics can be linked or matched with other data sources. So, 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 five individuals. This notion of a minimum 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, because, as the bin size increases, the number of people to whom a record may refer often increases, thereby masking the identity of the actual person.

In medical data bases, the minimum bin size should be much larger than the SSA guidelines suggest. Consider these three reasons: (1) most medical data bases are geographically located, hence, one can presume, for example, the ZIP codes of a hospital's patients; (2) the fields in a medical data base provide a tremendous amount of detail, hence any field can be a candidate for linking to other data bases in an attempt to reidentify 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 data base.

Determining the optimal bin size to ensure anonymity is tricky. It depends on the frequencies of characteristics found within the data as well as within other sources for reidentification. In addition, the motivation and effort required to reidentify 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 ten possible people and the ten people can be identified, then all ten candidates could be contacted or visited in an effort to locate the actual person. Likewise, if the mapping is 1 in 100, visits may be impractical, but all 100 could be

telephoned; and in a mapping of 1 in 1000, a direct mail campaign could be employed. The amount of effort a recipient is willing to expend depends on his/her motivation. Some medical files are quite valuable, and valuable data merits 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 data bases like that of SSA, the data base 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 data bases, however, must estimate these values. In the next section, I present computer programs that generalize data based on bin sizes and estimates. I then report results using these programs and discuss their limitations and the need for complementary policies.

Methods

Many possible tools can be used to maintain confidentiality when disclosing medical data. These include 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 depends on the nature of the data and its intended use; but each of these techniques is narrowly focused and there is little literature that addresses their use with medical data. I discuss three systems that 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 data base 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. I examine each in turn and then discuss their limitations.

The Scrub System

In 1996, I presented the Scrub System,[17] which locates and replaces personally identifying information in text documents and in textual fields of the data base. A close examination of two different computer-based patient record systems, one at Boston's Children's Hospital[18] and another at Massachusetts General Hospital,[19] quickly revealed that much of the medical content resided in the letters between physicians and in the shorthand notes of clinicians. In these letters and notes, providers discuss findings, explain current treatment, and furnish an overall view of patients' conditions.

Most institutions have few releases of data that include these notes and letters, but new uses for this information are increasing, and, not surprisingly, so is the desire to release this text. After all, these letters and notes are a valuable research tool and can corroborate the record. The fields containing the diagnosis, procedure, and medication codes when examined alone can be incorrect or misleading. A prominent physician recently stated that he purposely places incorrect codes in the diagnosis and procedure fields when such codes would reveal sensitive information about the patient.[20] Similarly, the diagnosis and procedure codes may be up-coded for billing purposes. The General Accounting Office estimates that as much as 10 percent of annual federal health care expenditures, including Medicare, are lost to fraudulent provider claims.[21] 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 a patient's name and replacing all occurrences with a pseudonym. References to a patient are often quite obscure. Consider, for example, the statement "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 for grammar or spelling. Although the resulting "unrestricted text" is valuable for understanding the medical condition and treatment of the patient, it poses tremendous difficulty to scrubbing because 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. With 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 on a human approach to the problem. It uses templates and localized knowledge to recognize personally identifying information. In fact, Scrub demonstrated that recognition of personally identi-

Wednesday, February 2, 1994	Wednesday, February 2, 1994	*February, 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	Marjorie Long, M.D. St. John's Hospital Huntington 18 Boston, MA 02151 RE: *Kathel Wallams* CH#*18-512-32871* DOB 05/26/86	*Erisa Cosborn*, M.D. *Brighaul* Hospital *Alberdam Way* *Peabon*, MA *02100* RE: *Kathel Wallams* CH#*18-512-32871* DOB *05/86*
Dear Dr. Lang:	Dear Dr. Lang:	Dear Dr. *Jandel*:
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.	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.	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 *Naming-ham*'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.
Patrick Hayes, M.D. 34764	*Mank Brones*, M.D. *21075*	*Mank Brones*, M.D. *21075*
Sample A	*Sample B*	*Sample C*

Table 5. Sample letter reporting back to a referring physician. Sample A is a made-up original text 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 the patient's mother and her mother's employer and telephone number. Sample B is the result from simple search and replace, and Sample C is the result from the Scrub System. Notice in Scrub 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 because Scrub was not sure how to replace it.

fying 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 surnames; and knowing these facts makes it easier to recognize them as likely names. Common facts along with their accompanying templates of use are considered common-sense knowledge and the itemization and use of common-sense knowledge is the backbone of Scrub.

Scrub accurately found 99 to 100 percent of all personally identifying references in more than 3,000 letters between physicians, while the straightforward search-and-replace approach properly located no more than 30 to 60 percent of all such references.[22] The higher figure of 60 percent for search and replace includes using additional information stored in the data base to help identify the attending physician's name, identifying number, and other information. Results of the search-and-replace method located as many as 84 percent[23] by taking advantage of the format of the letter and compositional cues like "Dear." However, most references to family members, additional telephone numbers, nick-names, and references to the physician receiving the letter were still not detected, whereas Scrub correctly identified and replaced these instances. However, Scrub merely deidentifies information; it cannot guarantee anonymity. Even though all explicit identifiers such as name, address, and telephone number are removed or replaced, it may be possible to infer the identify of an individual. Consider the following.

> 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 this child's life progresses in this manner, by age eight she may be headline news; but nothing in the narrative required scrubbing even though only one such child with this exact history would probably exist. 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 deidentifies clinical letters, the greatest volume of medical data found outside the originating institution flows from administrative billing records, which Scrub does not address. In 1996, NAHDO reported that thirty-seven states had legislative mandates to gather hospital-level data, and that seventeen states had started collecting ambulatory care (outpatient) data from hospitals, physician offices, clinics, and so forth.[24] Table 6 contains a list of the fields of information that NAHDO recommends these states accumulate. Many of them have subsequently given data to researchers and sold data to industry. As stated earlier, there are many other sources of ad-

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 the National Association of Health Data Organizations for State Collection of Ambulatory Data.

ministrative billing records with similar fields of information. What remains alarming is that most of these deidentified records can be reidentified because patient demographics and other fields often combine uniquely to identify individuals.

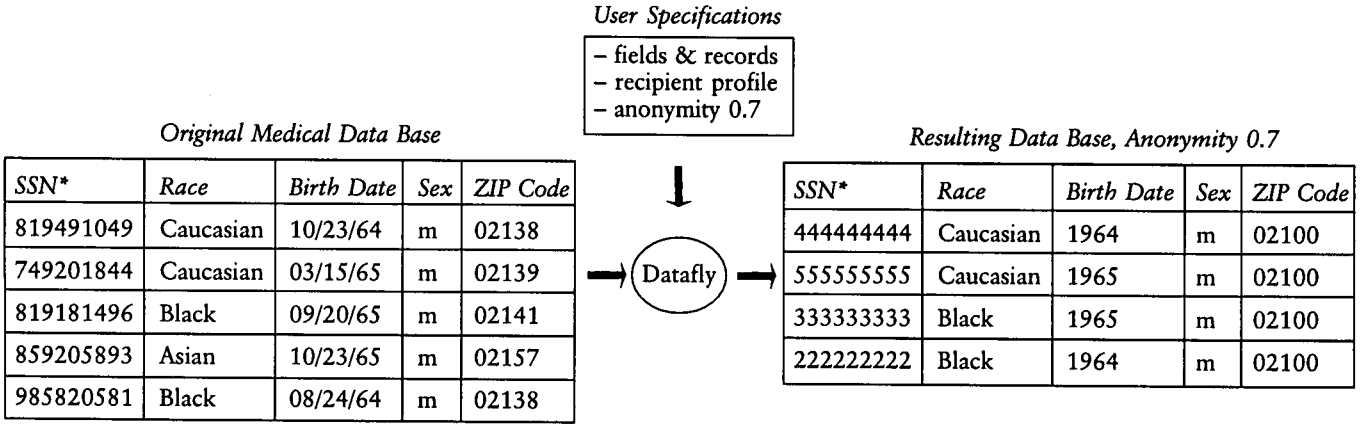
Earlier in 1997, I presented the Datafly System[25] whose goal is to provide the most general information useful to the recipient. Datafly maintains anonymity in medical data by automatically aggregating, substituting, and removing information as appropriate. Decisions are made at the field and the record levels at the time of data base access, so the approach can be incorporated into role-based security within an institution as well as into exporting schemes for data leaving an institution. The end result is a subset of the original data base that permits minimal linking and matching of data because each record matches as many people as the user had specified.

Figure 1 provides a user-level overview of Datafly. The original data base 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 data base 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; SSNs were automatically replaced with made-up alternatives; birth dates were generalized to the year; and ZIP codes were generalized to the first three digits.

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.[26]) Information within each field is generalized as needed to attain the minimum bin size; outliers, which are extreme, atypical values in the data, may be removed. When examining 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 SSNs.

Table 7 shows the relationship between bin sizes and selected anonymity levels using the Cambridge, Massachusetts, voters data base. As the anonymity level increases, the minimum bin size increases; and, in order to achieve the minimum bin size requirement, values within the birth date field, for example, are recoded as shown. Outliers are excluded from the released data and their corresponding percentages of the total number of records are noted. An anonymity level of 0.7, for example, requires at least 383 occurrences of every value in each field. To accomplish this in the birth date field, dates are recoded to reflect only the birth year. Even after generalizing over a twelve-month window, the values of 8 percent of the voters do not meet the requirement, so these voters are 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 data base whether the recipient could have or would use information external to the data base 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



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

Figure 1. The input to the Datafly System is the original data base and some user specifications. The output is a data base whose fields and records correspond to the anonymity level specified by the user, in this example, 0.7.
* Social Security number.

Anonymity	*Bin Size*	*Birth Date*	*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, Massachusetts, voters data with corresponding bin sizes. The birth date generalizations (in months) required to satisfy the minimum bin size is shown and the percentages of the total data base 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 mapping of anonymity levels to bin sizes is determined by the institution.

in the recipient or no concern over the sensitivity of the information within the field, and 1 represents full distrust in 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 that 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. These profiles all differ. Their selection and specificity of fields differ; their sources of outside information on which they could link differ; and their uses for the data differ. From publicly available birth certificates, drivers licenses and local census data bases, the birth dates, ZIP codes, and genders of individuals are commonly available, along with their corresponding names and addresses; so these fields could easily be used for reidentification. Depending on the recipient, other fields may be even more useful, but I limit my examples to profiling these fields. If the recipient is the patient's care-taker 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 care-taker 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 warrant a number as close to 1 as possible but not so much so that the resulting generalizations fail to provide useful data to the recipient. Because researchers or administrators bound by contractual and legal constraints prohibiting their linking of the data can be trusted, 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. Because 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. However, a profile for data being released for public use 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. In so doing, the profile identifies the associated risk to patient confidentiality for each release of data.

Numerous tests were conducted using Datafly to access a pediatric medical record system.[27] Datafly processed all queries to the data base over a spectrum of recipient profiles and anonymity levels to show that all fields in medical records can be meaningfully generalized as needed because any field is 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* or *ICD-10*) 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. If so, their replacements must be based on meaningful ranges in which to classify the values, and this reclassification is only done in cases where generalizing these fields is necessary.

For example, in Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected deidentified, medical encounter-level data with nearly 100 fields of information per encounter, including the fields in Table 6, for approximately 135,000 state employees and their families.[28] In a public hearing, GIC reported giving a copy of the data to a researcher, who in turn stated that 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 reidentified because 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 not yet released, the first version of a program called μ-Argus, which seeks to accomplish this goal.[29] The μ-Argus System is considered by many as the official confidentiality software of the European community, even though Statistics Netherlands considers this a preliminary version.[30]

μ-Argus, like Datafly, 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 Datafly, μ-Argus simply suppresses or blanks out the outlier values at the cell level. This process is called *cell suppression*.[31] The resulting data typically contain all the rows and columns of the original data, but values may be missing from some cell locations.

Table 8a lists many Caucasians and many females, but only one female Caucasian is in the data base. Tables 8b and 8c show the resulting data bases after Datafly and μ-Argus were applied to this data. I now step through how μ-Argus 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 code). Finally, 3-combinations are examined that include the "most" and "more" identifying fields. Obviously, there are many ways to rate these identifying fields, and unfortunately different ratings yield different results. The ratings presented in this example produced the most secure result using μ-Argus, although one could argue that too many specifics remain in the data for it to be released for public use.

Each unique combination of values found within sensitive fields constitutes a bin. When the number of occurrences of such a combination are less than the minimum required bin size, the combination is considered sensitive and hence an outlier. For all combinations that include the SSN, all such combinations are unique. One value of each outlier combination must be suppressed. For optimal results, μ-Argus suppresses values that occur in multiple outliers where precedence is given to the value occurring most

*SSN**	*Ethnicity*	*Birth Date*	*Sex*	*ZIP Code*	*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.
* Social Security number.

*SSN**	*Ethnicity*	*Birth Date*	*Sex*	*ZIP Code*	*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 of 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, hence the Caucasian female record was dropped as an outlier.
* Social Security number.

SSN*	Ethnicity	Birth Date	Sex	ZIP Code	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 of applying the approach of the μ-Argus System to the data in Table 8a. The minimum bin size is 2. The Social Security number was marked as being most identifying; the birth, sex, and ZIP code 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, a unique record still remains for the Caucasian male born in 1964 who lives in the 02138 ZIP code.

often. The final result is shown in Table 8c. Responsibility for when to generalize and when to suppress rests with the user. For this reason, μ-Argus operates in an interactive mode so the user can see the effect of generalizing and may undo a step.

I now briefly compare the results of these two systems.[32] In the Datafly System, generalization across a subset of sensitive fields ensures that the combination across those fields will adhere to the minimum bin size. This is demonstrated in Table 8b. The μ-Argus program, however, only checks 2- or 3-combinations; hence, sensitive combinations across 4 or more fields would not be detected. For example, Table 8c still contains a unique record for a Caucasian male born in 1964 who lives in the 02138 ZIP code, because 4 characteristics 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 is done in Datafly, appears to provide more secure releases of data. Further, because the number of fields, especially demographic fields, in a medical data base is large, this may prove to be a serious handicap when using μ-Argus with medical data. In recent work, I have developed a program that examines combinations of values within sensitive fields and produces an optimal solution with respect to minimum cell suppression.[33] Though more specificity remains in the resulting data, making it more useful to the recipient, the underlying issues remain the same.

Discussion

The Scrub System demonstrates that medical data, including textual documents, can be deidentified, but, as shown, deidentification alone is not sufficient to ensure confidentiality. Not only can deidentified information often be reidentified by linking data to other data bases, but specific individuals can also be identified by releasing too many patient-specific facts. Unless we are proactive, the proliferation of medical data may become so widespread that it will be impossible to release medical data without further breach of confidentiality. For example, the existence of rather extensive registers of business establishments in the hands of government agencies, trade associations, and private businesses like Dun and Bradstreet has virtually ruled out the possibility of releasing data base information about businesses.[34]

The Datafly and μ-Argus systems illustrate that medical information can be generalized, replaced, or suppressed so that fields and combinations of fields adhere to a minimum bin size, and, by so doing, confidentiality can be maintained. By using such systems, we can even provide anonymous data for public use. These systems have two drawbacks, as discussed below, but these shortcomings can be counteracted by policy.

One concern with μ-Argus and Datafly is the determination of the proper bin size and its corresponding measure of disclosure risk. No standard 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 data bases that we assume a recipient is using. Several researchers have proposed mathematical measures of the risk, which compute the conditional probability of the linker's success.[35]

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, public, or semipublic 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 data base itself, the minimum bin sizes and sampling fractions may be far from optimal and 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 data base.[36] These methods are based on subsampling techniques and equivalence class structure. Absent these techniques, uniqueness in the population based on demographic fields can be determined using population registers that include patients from the data base, 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 data base, a producer could use population registers to identify sensitive demographic values within a data base, and thereby obtain a measure of risk for the release of the data.

The second drawback concerns the dichotomy between researcher needs and disclosure risk. If data are explicitly identifiable, the public expects patient permission to be required. If data are released for public use, then the producer must 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 deidentified, but not necessarily anonymous, data are to be released, the likelihood that an effort will be made to reidentify an individual may increase based on the needs of the recipient. The onus, therefore, is on the recipient of the data who should be bound by a trust relationship with society and the producer of the data to handle, store, use, and release resulting information properly. The recipient should be held accountable for the confidentiality of the data.

Datafly and μ-Argus quantify this trust by profiling the fields requested by a recipient. But profiling requires guesswork in identifying fields on which a 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 a recipient, and, as a result, individuals may be more easily identified. This risk cannot be perfectly resolved by the producer of the data because the producer cannot always know what resources a 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 to share the risk. Table 9 contains some guidelines, which, if applied, would clarify which fields need to be protected against linking. Using this additional knowledge and the techniques presented in Datafly and μ-Argus, 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, no contractual arrangements limit further dissemination or use of the data. Even in cases that include IRB review, no contract usually results. Further, because 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 legal and monetary sanctions or penalties for improper use or conduct should apply, because remedy for abuse lies outside technology and statistical disclosure techniques and resides in contracts, laws, and policies.

Conclusion

A few researchers may not find the magnitude and scope of the problems concerning the identifiability and disclosure of medical records surprising, but such revelations have alarmed legislators, scientists, and federal agencies.[37] I must caution, therefore, against overreaction that may lead to inappropriate and inoperable policies. I argue that knowledge of the problems with current practices and the availability of incremental solutions, not ignorance of their existence or nature, provides the best foundation for good policy. What is needed is a rational set of disclosure principles, based on comprehensive analysis of the fundamental issues, which are unlikely to evolve from piecemeal reactions to random incidents. The technology described here is quite helpful, but society must still make informed decisions.

There is a danger in oversimplifying this work. It does not advocate giving all the data on all the people without regard to whether individuals can be identified. It also does not advocate releasing data that is so general it cannot be useful; substantial suppression does not appear to be the norm. From the viewpoint of a person who receives the data, these systems seek to provide the most general data possible that is practically useful to that person. From the viewpoint of privacy, if that level of generality does not

- 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 data base are needed for this purpose.
- The recipient must be strictly and legally accountable to the producer for the security of the data and must demonstrate adequate security protection.
- The data must be deidentified. The release must not contain explicit individual identifiers or data that would be easily associated with an individual.
- 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, and whether the recipient intends to link to such data. The recipient must also identify those fields to which the recipient will link the data. If such linking identifies patients, then patient consent may be warranted.
- The data provider should have the opportunity to review any publication of information from the data to ensure that no potential identifying disclosures are published.
- At the conclusion of the project, and no later than some specified date, the recipient must destroy all copies of the data.
- 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.

provide sufficient protection, then the techniques presented here identify the nature and extent of trust required for a given release of data. Polices and regulations regarding the agreements necessary to make that trust explicit and to enforce its terms lie outside the technology.

Consider, for example, the case of data released to researchers. When anonymous data is useful, the data should be released in that form. In some cases, completely anonymous data is not practically useful; in those instances, we can quantify the trust given to researchers who receive more identifiable data. Changes should be made such that public use files adhere to a reasonably high level of anonymity. In cases where more identifiable data is needed, society should consciously decide how to release such data and a recipient should be held responsible not to violate the contractual agreements that spell out the conditions of trust.

Finally, I warn against doing nothing. The burden of determining the risk of disclosure may appear cumbersome, which is not a realistic assumption given that these systems operate in real-time and that their development costs have been nominal. Nevertheless, consider an alternative to autonomous data base systems in which we have a centralized federal repository for medical data, like those found in Canada and other countries. Though institutions and businesses could maintain their own data for internal purposes, they could not sell or give data away in any form, except for disclosure to the federal repository, remuneration for services, and required reporting. The recipients of these data would, in turn, be equally restricted from 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, or others could be answered without releasing any data; instead, the trusted authority would run desired queries against the data and provide noncompromising results to the investigators. In releases of deidentified data, the exact risk could be computed and accompanying penalties for abuse incorporated into the dissemination process.

This type of system may have advantages for maintaining confidentiality, but it requires a single point of trust or failure. Current societal inclinations suggest that the American public would not trust a single 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 the role of the trusted authority anyway. If information brokers emerge as the primary keepers of medical data, as Dun and Bradstreet does for business data, then they may eventually rank among the most conservative advocates for maintaining confidentiality and limiting dissemination, because their economic survival would hinge on protecting what would be their greatest asset, our medical records.

Acknowledgments

I thank Beverly Woodward, Ph.D., for many discussions and comments. I also thank Peter Szolovits, Ph.D., at the Massachusetts Institute of Technology, for providing an environment that made it possible for me to explore my own ideas, Patrick Thompson and Sylvia Barrett for editorial suggestions, and Professor Pierangela Samarati of the University of Milan for discussions. I acknowledge the continued support of Henry Leitner, Ph.D., of Harvard University. This work has also been supported by a Medical Informatics Training Grant (1 T15 LM07092) from the National Library of Medicine.

References

1. I. Kohane et al., "Sharing Electronic Medical Records Across Heterogeneous and Competing Institutions," in J. Cimino, ed., *Proceedings, American Medical Informatics Association* (Washington, D.C.: Hanley & Belfus, 1996): 608–12.

2. Office of Technology Assessment, *Protecting Privacy in Computerized Medical Information* (Washington, D.C.: U.S. Government Printing Office, 1993).

3. See L.O. Gostin et al., "Privacy and Security of Personal Information in a New Health Care System," *JAMA*, 270 (1993): at 2487 (citing Louis Harris and Associates, *The Equifax Report on Consumers in the Information Age* (Atlanta: Equifax, 1993)).

4. Louis Harris and Associates, *The Equifax-Harris Consumer Privacy Survey* (Atlanta: Equifax, 1994).

5. G. Cooper et al., "An Evaluation of Machine-Learning Methods for Predicting Pneumonia Mortality," *Artificial Intelligence in Medicine*, 9, no. 2 (1997): 107–38.

6. See Kohane et al., *supra* note 1.

7. B. Woodward, "Patient Privacy in a Computerized World," *1997 Medical and Health Annual* (Chicago: Encyclopedia Britannica, 1996): 256–59.

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

9. P. Clayton et al., National Research Council, *For the Record: Protecting Electronic Health Information* (Washington, D.C.: National Academy Press, 1997).

10. See, for example, Donna E. Shalala, Address at the National Press Club, Washington, D.C. (July 31, 1997).

11. B. Woodward, "The Computer-Based Patient Record and Confidentiality," *N. Engl. J. Med.*, 333 (1995): 1419–22.

12. D. Linowes and R. Spencer, "Privacy: The Workplace Issue of the '90s," *John Marshall Law Review*, 23 (1990): 591–620.

13. D. Grady, "Hospital Files as Open Book," *New York Times*, Mar. 12, 1997, at C8.

14. See Clayton et al., *supra* note 9.

15. "Who's Reading Your Medical Records," *Consumer Reports*, Oct. (1994): 628–32.

16. L. Alexander and T. Jabine, *Social Security Bulletin: Access to Social Security Microdata Files for Research and Statistical Purposes*, 41, no. 8 (1978).

17. L. Sweeney, "Replacing Personally-Identifying Information in Medical Records, the Scrub System," in Cimino, *supra* note 1, at 333–37.

18. I. Kohane, "Getting the Data In: Three-Year Experience with a Pediatric Electronic Medical Record System," in J. Ozbolt,

ed., *Proceedings, Symposium on Computer Applications in Medical Care* (Washington, D.C.: Hanley & Belfus, 1994): 457–61.

19. G. Barnett, "The Application of Computer-Based Medical-Record Systems in Ambulatory Practice," *N. Engl. J. Med.*, 310 (1984): 1643–50.

20. Anon., Privacy & Confidentiality: Is It a Privilege of the Past?, Remarks at the Massachusetts Medical Society's Annual Meeting, Boston, Mass. (May, 18, 1997).

21. Government Accounting Office, *Fraud and Abuse in Medicare and Medicaid: Stronger Enforcement and Better Management Could Save Billions* (Washington, D.C.: Government Accounting Office, HRD-96-320, June 27, 1996).

22. See Sweeney, *supra* note 17.

23. See *id.*

24. See National Association of Health Data Organizations, *supra* note 8.

25. See L. Sweeney, "Computational Disclosure Control for Medical Microdata, the Datafly System," *Proceedings of the Bureau of the Census Record Linkage Workshop* (Washington, D.C.: Bureau of the Census, 1997): forthcoming.

26. For guidelines, see L. Sweeney "Guaranteeing Anonymity When Sharing Medical Data, the Datafly System," *Proceedings, American Medical Informatics Association* (Nashville: Hanley & Belfus, 1997): forthcoming.

27. See *id.*

28. M. Lasalandra, "Panel Told Releases of Med Records Hurt Privacy," *Boston Herald*, Mar. 20, 1997, at 35.

29. A. Hundepool and L. Willenborg, "mu- and tau-Argus: Software for Statistical Disclosure Control," *Third International Seminar on Statistical Confidentiality* (1996) (available at <http://www.cbs.nl/sdc/argus1.html>.

30. For a presentation of the concepts on which μ-Argus is based, see L. Willenborg and T. De Waal, *Statistical Disclosure Control in Practice* (New York: Springer-Verlag, 1996).

31. N. Kirkendall et al., *Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper* (Washington, D.C.: Office of Management and Budget, no. 22, 1994).

32. For a more in-depth discussion, see Sweeney *supra* note 26.

33. L. Sweeney, "Towards the Optimal Suppression of Details When Disclosing Medical Data, the Use of Sub-Combination Analysis," *Proceedings of the 9th World Conference on Medical Informatics* (1998): forthcoming.

34. See Kirkendall et al., *supra* note 31.

35. G. Duncan and D. Lambert, "The Risk of Disclosure for Microdata," *Proceedings of the Bureau of the Census Third Annual Research Conference* (Washington, D.C.: Bureau of the Census, 1987): 263–74.

36. C. Skinner and D. Holmes, "Modeling Population Uniqueness," *Proceedings of the International Seminar on Statistical Confidentiality* (Dublin: International Statistical Institute, 1992): 175–99.

37. For example, Latanya Sweeney's testimony before the Massachusetts Health Care Committee had a chilling effect on the proceedings that postulated that the release of deidentified medical records provided anonymity. See *Session of the Joint Committee on Health Care, Massachusetts State Legislature*, (Mar. 19, 1997) (testimony of Latanya Sweeney, computer scientist, Massachusetts Institute of Technology). Though the Bureau of the Census has always been concerned with the anonymity of public use files, they began new experiments to measure uniqueness in the population as it relates to public use files. Computer scientists who specialize in data base security are re-examining access models in light of these works.