Against Simple Universal Health-Care Identifiers

Peter Szolovits
MIT Laboratory for Computer Science, Cambridge, MA.
Isaac Kohane
Children's Hospital, Harvard Medical School, Boston, MA.

Medical records are becoming fully computerized; technical, administrative and economic forces are pushing toward standardization on a single identifier, such as the Social Security Number (SSN) to index all records; consequently, the privacy and security of our medical histories will be severely compromised. We argue that there are sensible and effective technological means available to reduce the risks of such compromise, and that it is time to *design* the characteristics we want of our record-keeping systems rather than to fall into the problems caused by unthinking adoption of an overly simplistic approach.

Current Trends and Their Dangers

Over a year ago, a meeting of the American College of Medical Informatics took up the question of how to identify patient records, and concluded that the simplest, most expedient solution was to adopt the Social Security Number (extended by a check digit) as the universal health-care identifier [1]. The advantages of this proposal are mostly that virtually everyone already has an SSN and that there already exists an organization that issues new ones as needed. Recognized disadvantages include relatively frequent cases where more than one person was issued the same SSN, people who have been issued multiple SSN's, newborns and "marginal" people who have health care needs but no SSN, eventual insufficiency of a nine-digit "addressing" scheme, lack of consistency checks leading to easy misidentification, and privacy considerations. Our severe concerns grow mainly out of the privacy considerations.

In any society that values privacy but also keeps records, there must be concern about possible invasions of privacy. It has always been possible to violate the privacy of selected individuals by devoting enough resources to looking into their affairs. Examination of public records, interviews with associates and colleagues, perusal of newspaper stories, illicit access to private files, and physical surveillance all offer the dedicated snoop ways of invading the privacy of a particular person. Fortunately for privacy in general, however, such research is cumbersome and expensive, and can therefore be applied only in relatively rare situations.

If we organize our records in such a way that the indexing of information is routine, then we make the job of the snoop much simpler and less expensive. Credit bureaus, for example, with their voluminous data sets on our consumer behavior make it easy to violate the privacy of any specific individual [2]. The reporter J. Rothfeder, for example, was able to use a simple ruse to gain access to such national information repositories and to track down specific details of the life of then Vice President Dan Quayle [3]. The existence and penetrability of such databases to date provides, however, mostly a quantitative, not a qualitative deterioration of individual privacy. A public figure such as the Vice President could always be selected as the target of investigation; the conveniently-collected data, used as Rothfeder demonstrated, merely make this less costly.

The more threatening consequence of large, insecure databases is the ability to search them easily and cheaply for groups of previously anonymous people with certain

This paper was published as P. Szolovits and I. Kohane. Against Universal Health-care Identifiers. *Journal of the American Medical Informatics Association* 1:316–319, 1994.

characteristics [4,5]. With this ability, the snoop can not only invade the privacy of someone already targeted, but can in fact develop new lists of interesting targets who meet specific criteria. The techniques are already widely (and legitimately) used by marketing organizations who select prospective catalog recipients by sorting through records of past purchases, but current use tends to be limited to searches over specific, isolated data sets, not over the lifetime accumulation of information about everyone. Yet, under current proposals, we are making it simpler to collate information from very different sources by indexing all transactions pertaining to an individual under his or her SSN. Future snoops may be able to develop lists of people with certain educational and job backgrounds who suffer from specific maladies and like to spend money on certain kinds of entertainment. The opportunities for abuse are enormous. Yet the more and more widespread adoption of a single identifier facilitates and encourages just this situation.

The SSN was created in support of the social security program of the Federal Government in the 1930's. Originally limited in use to recording individual contributions to the social security plan, its approved Federal use has been broadened to identifying taxpayers and their tax transactions, civil service employment, Defense Department personnel, recipients of some forms of public assistance, and other functions. In addition, states use the SSN for their own tax-related records, and many also index drivers' licenses, motor vehicle registration, and criminal history to the same identifier. Non-government uses include records holding an individual's history of employment, insurance, credit, and education. If current trends continue, health records will join this list.

With growing interoperability of database systems, we are getting close to the time when a single SQL query can, at not very great cost, find a selection of individuals based on any or all of the characteristics indexed by the SSN listed above. To anyone who values privacy even slightly, this is a frightening prospect.

The Falsely-Perceived Technological Imperative

When ACMI's recommendation first appeared, one of us (P.S.) began a discussion of these issues on a mailing list of ACMI members. We were surprised that many of our colleagues did not share our concerns. Our impression of the electronic discussion is that responses split into three camps, with the following (caricature) positions:

1. The universal use of the SSN is bad, because it makes the collation of large databases easy, and is likely to lead to intolerable abuses. (I.e., these are the people who agreed with us.)
2. It's too bad that using a universal ID has these undesirable consequences, but the horse is out of the barn already, so we may as well just learn to live with it.
3. "No problem." And anyway, any reasonable alternatives would be inordinately costly[1].

In addition, we were struck by the dearth of suggestions (from a community normally brimming with technical ideas) about how one might use technological means to ameliorate these problems.

The ideas we present below focus on one specific, narrow issue: How to make it more difficult for unauthorized individuals to perform massive searches across large databases, collating information from multiple sources. We do not specifically address the "targeted" invasions of privacy where the individual to be investigated is already identified, but mainly the search for interesting populations of people who share certain health (and other) characteristics but who were not initially known.

---

[1]Even so, verifying the accuracy of all SSN's and ensuring future accuracy would require a major reorganization and investment estimated at $1.0 to $2.5 billion [6].

Technological Alternatives

One of the major technical advances in computer science occurred in the late 1970's and is just now beginning to influence commercial practice: the development of *public-key cryptography*. In any cryptographic method, we try to assure that the sender of a message can encode it in some way such that the recipient can decode it, but such that anyone else who intercepts the coded message cannot make sense of it. Simple codes have been in use since Roman times, and very effective coding schemes have been a routine part of military communications in our century. Such schemes require, however, that both the sender and the recipient of a message agree, ahead of time, on the coding and decoding method they will use. This makes spontaneous encrypted communication impossible, and leaves the difficult question of how to agree on a secret encoding/decoding method before there is an effective encrypted means of communicating. (In most cases, this has been done by physical exchange of codes; this certainly requires much advance planning.)

Public-key cryptography does away with the need for pre-arranged cryptographic codes in a clever way that relies on the complexity of certain mathematical computations. Anyone who wants to use this method needs to acquire two keys that allow arbitrary messages to be encoded and decoded. The keys define functions which are mathematical inverses of each other, but with the characteristic that it is phenomenally costly to calculate one of them if given only the other [7].[1] For Alice, a typical user, we call these A and A'. Alice keeps A a well-guarded secret, but publishes A' so that anyone interested in communicating a secret message to her may find it. If Bob, say, wishes to send her a message m, he can compute $A'(m)$ and transmit it to Alice, who then applies her secret key A to it to recover the original message, because $A(A'(m)) = m$. A potential snoop, Charley, cannot reconstruct m even if he intercepts $A'(m)$ because he does not know A, and because it is very costly to compute A even knowing A'.

An interesting corollary of the fact that public and private keys are inverses is that the same keys form the basis for a reliable means of authenticating the sender of a message. Thus, if Bob's keys are B and B', Bob can assure Alice that a message he sends her really came from her by encoding (part of) it with his private key and asking her to decode it with his public key, which she can easily find. Thus, Bob sends $A'(B(m))$, and Alice computes $B'(A(A'(B(m)))) = m$. If this computation yields a sensible message, Alice can conclude that its sender must have known B, and must therefore be Bob. The system is robust, individuals can change their keys simply by publishing a new public key, and elaborations are possible to create escrow agents, trusted intermediaries, messages that require cooperation among several parties to decode, and a broad range of other interesting secure communications mechanisms.

This approach is becoming commercially popular, and is the basis of *digital signatures*, which now form an essential part of various computer-based authentication systems such as the Apple Macintosh PowerTalk *Signers*, PGP, a widespread public-domain digital signature scheme for Internet email, General Magic's Telescript communication language, and MIT Project Athena's Kerberos user authentication system.

Cryptography-Based Health-Care Identifiers

We believe that some method based on the ideas of public key cryptography can be fruitfully applied to the problem of collecting and keeping comprehensive medical records

---

[1] The currently most popular method uses the computational difficulty of factoring very large (hundreds of digits) numbers to make it impossible to calculate one key from its inverse. The best currently-known algorithms for factoring require a time that grows exponentially in the size of the number being factored. Thus, breaking one of these codes can be made practically impossible simply by choosing keys that are large enough. It has not been proven that there is no more efficient factoring algorithm, but none have been found in many years of research, so the technique is thought to be quite safe.

that are easy to integrate when it is appropriate to do so, but that are difficult to collate in service of widespread searches that can invade individuals' privacy. We provide two example designs here, to illustrate feasible alternative approaches.

The first is a completely decentralized scheme, in which the individual patient has ultimate control over the degree to which the lifetime collection of medical information about him or her is made available to others. Instead of an SSN, every individual is issued, at birth, a private key K, unique to that person. Every institution that maintains health records is issued a public key, H'. Thus, each hospital, clinic, doctor's office, HMO, insurance company, other third-party payer, government regulatory agency, epidemiologic data center, research study, etc., has its unique H'. When an individual first needs to deal with an institution, a universal cryptographic function f is applied to K and H' to compute the ID number under which that individual's records will be kept at that institution. The computation of ID does not, however, reveal the patient's private key K, just as Alice's sending $A(m)$ did not reveal A. Therefore, any specific institution can have access only to those records it has itself collected about a patient, to those the patient has asked other institutions to forward to this one, and to those at other institutions for which the patient has provided this one their old ID.

Because $ID = f(K,H')$, it depends on both the individual and the institution. Therefore, no two people will have the same ID at the same institution, and no two institutions will have the same ID for the same individual. Furthermore, people can validate their identities by demonstrating the ability to compute their ID given the institution's H'. Under this scheme, if one institution needs to gain access to a patient's records at another, they must ask the patient to compute his or her ID at that other institution, by applying f to the secret K and the public H' of the other institution. The same mechanism is used to tell a hospital, say, the appropriate ID number under which to report reimbursement claims to an insurance company or epidemiological data to the CDC. Any institution holding an unauthorized patient record could be exposed by demonstrating that $f(K,H')$ does not correspond to the identifier used by that institution.

Practicalities of a scheme such as this would require patients to carry something like a "smart card" instead of the current social security card, because the computation of f is not practical without a small computer, and the keys involved are lengthy. Longitudinal records can be assembled by chaining records back from those of current health care providers to those of past providers, and a past provider could not find out newer information without the patient's cooperation. Mechanisms would also need to be established to deal with people losing their cards and keys, though the simplest expedient is just to issue someone a new K and then include in their new records references to the old.

A second possible scheme centers on the institution rather than the individual, and assumes that all patients' private keys are held by a centrally-organized *ID server.* For identification purposes, each patient has something that serves as their public key, and can be as widely known as the SSN is today. The ID server must be a trusted institution, perhaps established and administered by the government or some semi-public consortium. Any institution may request (by authenticating its identity) an ID for a particular patient to be used at that institution, and the ID server returns it, protected by encrypted communications. In this case, neither the patient nor any health care provider know the patient's private key or the patient's ID at any other institution. The ID server is responsible for arranging all transfer of data from one institution to another, according to an agreed-upon and authenticated protocol.

Neither of the above two simple schemes is likely to be satisfactory, and the design of a coherent scheme that meets the data communication and privacy needs of the health care system is a difficult task. For example, we have treated the patient's records at a single institution as indivisible, but it might be necessary to provide different protections for different parts of a medical record. Psychiatric records and HIV status may not need to be accessible to all accessors of the record.

We have focused on the use of technology to prevent unauthorized access. In some settings it may be more appropriate to give presumptive access to anyone who claims a need to know, but to use authentication methods to assure an unforgeable “signed” log of everyone who has actually accessed the data. Improper access is then not prevented, but a record of unjustifiable access implies a posteriori liability for invasion of privacy.

The technology of public-key cryptography is a very powerful tool, and its creative application can lead to many interesting systems that provide different degrees of privacy, convenience, and flexibility. Some of these capabilities will certainly be developed piecemeal in any case, because American industry, including the health care industry, will adopt useful and easily-available methods to enhance the security and reliability of its communications. Adopting technically naive solutions such as universal use of the SSN as a health care identifier will simply lock us in to a system that sacrifices personal privacy because of the mistaken impression that nothing better is feasible. Our peers on the “Information Superhighway” will laugh at our missed opportunities, at least until they realize that we have also compromised their privacy.

We call for a serious and informed discussion of the needs of health care record keeping. Let us decide what we want, define the trade-offs that will arise, and use the rich technologies available to design solutions that best achieve our desires.

Acknowledgments

Peter Szolovits's research has been supported, in part, by a grant from IBM Corporation and by grant R01 LM 05296 from the National Library of Medicine.

References

1. Board of Directors of the American Medical Informatics Association. Standards for medical identifiers, codes and messages needed to create an efficient computer-stored medical record,” *Journal of the American Medical Informatics Association 1*: 1–7, 1993.

2. Kratka, J. “For their eyes only: The insurance industry and consumer privacy,” Massachusetts Public Interest Research Group, 1990.

3. Rothfeder, J. *Privacy for sale: How computerization has made everyone's private life an open secret*; Simon and Schuster: New York, 1992.

4. Curran, W. J.; Stearns, B.; Kaplan, H. Privacy, confidentiality and other legal considerations in the establishment of a centralized health-data system. *New England Journal of Medicine* 1969, *281*, 241-8.

5. Gostin, L. O.; Turek-Brezina, J.; Powers, M.; Kozloff, R.; Faden, R.; Steinauer, D. D. Privacy and security of personal information in a new health care system. *JAMA* 1993, *270*, 2487-93.

6. Statement of Gwendolyn S. King, commisioner of Social Security. *Hearings Before the House Subcommittee on Social Security, Committee on Ways and Means.* 102nd Congress, 1st session: 1991.

7. Rivest, R.; Shamir, A.; Adleman, L. A method for obtaining digital signatures and public-key cryptosystems. *Comm. ACM* 1978, *21*, 120-6.