Encrypting Personal Identifiers

Eleanor Meux

Study Setting. A statewide patient discharge database contained only one unique identifier: the social security number (SSN). A method was developed to transform (encrypt) the SSN so that it could be made publicly available, for purposes of linking discharge records, without revealing the SSN itself. The method of encrypting the SSN into a Record Linkage Number (RLN) is described.

Principal Findings. The same RLN will always result from the same SSN; it is highly improbable that the same RLN would be produced by two different SSNs; the SSN cannot be derived from the RLN, even given access to the encryption program; the encryption method cannot be determined through knowledge of a number of SSN/RLN combinations; and the method can be described, evaluated, and adapted for use by other researchers without compromising confidentiality of the RLNs resulting from the method.

Keywords. Unique identifier, social security number, encryption, linkage of records, confidentiality

Various state governments and state hospital associations collect data about patients discharged from hospitals (Agency for Health Care Policy and Research 1991). The data include demographic (e.g., race, gender, age), clinical (e.g., diagnoses, procedures), and utilization (e.g., charges, length of stay, source of payment) variables. When collected by state agencies, these data are commonly made available to users (e.g., researchers, government agencies, private purchasers) in the form of computer files. The uses for patient discharge data include: study of the costs of treating certain diagnoses; comparison of the costs, efficiency, and quality of care among hospitals; study of the outcomes of medical care; assistance for purchasers of care in making selective contracting decisions; guidance for policy decisions by government; and the study of diagnoses, payment sources, and discharge status of specific population groups.

In most of the patient discharge data programs, the intent is to make data available that will be useful in answering statistical questions, without allowing the identification of individuals. Thus, the programs may contain

only a rudimentary form of unique identifier, which could not be used to identify any individual. For example, the state of Washington uses an identifier composed of the first two letters of the last name, birth date, and first two letters of the first name. California had no unique identifier in its patient discharge data system, and found that this limited the usefulness of the data because people were unable to identify readmissions, follow patients between hospitals, or link mortality files in order to obtain postdischarge deaths (among other potential uses of a unique identifier). In 1988, California added patient social security number (SSN) to the legally required patient discharge data elements, to enable linking within discharge records and thus to provide information relevant to quality and cost of health care. The legislation also contained a restriction that the patient's rights of confidentiality should not be violated in any manner. Implementation of the legislation occurred on June 30, 1991, the due date of the first reporting period containing SSNs.

Confidentiality rights in California are perhaps stronger than in some other states. California's constitution contains an explicitly stated and recognized right to privacy clause, added by voters in 1972. In related legislation, the California Information Practices Act of 1977 established restrictive conditions for the disclosure by a state agency of personal information in a manner that could link the information to an individual. The Act includes criteria such as legal authority, informed consent, and use of the information consistent with the purposes for which the information was collected.

The Office of Statewide Health Planning and Development (OSHPD), which is the department of state government responsible for the California patient discharge data program, now has statutory authority to collect the patient's SSN, diagnoses, procedures performed, zip code of residence, date of birth, and other information. It does not collect the name or address of the patient. The purpose of this data program is for public disclosure–but without violating patient confidentiality in any way. When the SSN was added to its list of data elements, OSHPD needed to find a way to provide the benefits of a unique identifier to the users of its data, without revealing the SSN itself. It needed a method of transforming the SSN through encryption into some other unique identifier that would have no external utility in identifying an individual.

Address correspondence and requests for reprints to Eleanor Meux, Ph.D., Chief, Patient Discharge Data Section, Office of Statewide Health Planning and Development, California State Health and Welfare Agency, 818 K Street, Room 100, Sacramento, CA 95814. This article, submitted to *Health Services Research* on December 15, 1992, was revised and accepted for publication on July 30, 1993.

CRITERIA FOR ENCRYPTION

The method of encrypting the SSN had to meet four restrictive criteria:

1. *The method had to be reliable.* The same SSN must be encrypted into the same unique identifier every time, even over a number of years of data (so long as the encryption program is unchanged). This criterion is necessary so that records can be linked over time.
2. *The method had to produce truly* unique *identifiers.* That is, the probability that the same unique identifier might be produced from two different SSNs should be exceedingly low. This criterion is necessary to minimize records erroneously matched as an artifact of the encryption program.
3. *The method had to be irreversible.* No one, not even the computer programmer with access to the actual program, should be able to go backward and determine the SSN from its encrypted value. This criterion is necessary to protect confidentiality, so that the encrypted value can be publicly released without any possibility that a data user could determine the SSN. The SSN can be replaced with its encrypted value, in order to prevent any possible future access to the SSN (or the original file containing the SSN can be retained, in case of future need).
4. *The method had to be secure, in various ways.* First, the method should be describable (as in this article) so that it can be evaluated, without revealing critical parameters. Second, it must be possible to control access to the program, so that no one can deliberately run a specific SSN through it and so obtain the encrypted value. Third, it must not be possible to derive the method from knowledge of the actual encrypted value for a number of SSNs.

The third and fourth criteria partially originate in the unusual security problems raised by the public release of patient discharge data. These problems must be addressed by the encryption method. One obvious problem is that someone may seek to identify patients (in the publicly released data) having specific characteristics, such as AIDS or other conditions of special interest to health insurers and employers. In order to prevent this, publicly released data must not contain any information that might serve, singly or in combination, to identify a specific person. Another problem is that someone may want to locate discharge records of a specific individual in the publicly released data, using specific known characteristics of the person. Identification with certainty would require access to the patient's SSN and/or name and address. By never releasing the SSN, name, and address, this risk to confidentiality can be minimized.

Due to the "small cell" problem the risk to confidentiality can never be totally eliminated so long as individual records are made public. If the "cell" (the conjunction of age, gender, geographic area, diagnosis, facility, admission/discharge dates) is small enough, someone with external information (e.g., from a newspaper) may be able to say that a particular discharge record probably belongs to a particular individual. But, so long as personal identifiers are not released, the identification can only be probabilistic and not certain.

Both of these problems might occur in any kind of publicly released data containing records of individuals. But hospital discharge data present another type of problem: hospitals with a copy of the publicly released database can match at least some of the records in the public data with their own internal medical record database. Their own records include not only their patient's name and address, but also his or her SSN. If, by examining the combinations of SSN and the released unique identifier for its own patients, a hospital can determine the "rule" by which the SSN was transformed (such as a substitution code or algorithm), it might thereby be able to determine the SSN for all patients in the statewide database. It is therefore necessary that the encryption technique use multiple coding algorithms in such a way that the algorithms or substitution codes cannot be derived from knowledge of a number of SSN/identifier matches. This requirement has been made a specific aspect of the fourth criterion.

ALTERNATIVES CONSIDERED

Three alternatives for transforming the SSN were considered.

1. The SSN could be encrypted using available commercial encryption packages (Computer Security Institute 1991). Most work done in the area of cryptography (Meyer and Matyas 1982) is concerned with how to store or transmit information securely so that it can be read only by an authorized recipient who has been given some key or password. The requirements described here do not fit standard situations being served by usual cryptographic models. After exploring the commercial products and consulting services available in the marketplace (Computer Security Institute 1991), this alternative was rejected because none of those contacted had available an irreversible encryption method that would meet the criteria described.

2. Every possible SSN could be randomly (without replacement) assigned a corresponding nine-digit number, or alphanumeric value, to serve as the unique identifier. There would be no encryption method per se, just a one-identifier-for-one-SSN correspondence file, which would have to be securely maintained. While this method would meet some of the criteria,

it would be costly to store the file and costly to search through almost one billion values as many times as would be needed (California's discharge data program receives one million different SSNs each six months). This alternative was rejected because it would be inefficient and hard to keep secure.

3. The third alternative was a method developed for this purpose by the author and described in this article: encrypting the SSN reliably but irreversibly into a unique "record linkage number" (RLN). The resulting RLN can be released to the public without violating confidentiality; is generalizable to other types of identifiers; and can be described in sufficient detail, as in this article, to be adopted as a methodology by any other program needing to encrypt personal identifiers, without technical assistance or purchase of software.

THE METHOD

Overview

The RLN produced from an SSN consists of nine alphanumeric characters (a combination of letters and numbers). Each digit in the SSN is translated into a letter or number for the RLN using a three-step process. The first step selects an arrangement of the ten digits 0 through 9. The second step selects an arrangement of ten letters/digits (alphanumerics). The third step consists of using the combination of the two arrangements selected in a substitution code for the values in the SSN, with the RLN as the result.

In the following description of these steps, the nine positions in the SSN are identified as *a–i*, from left to right. Before selecting specific positions to be used, the author obtained a sample of 867,535 SSNs and obtained the distribution of digits in each position. It was found that the digits 0–9 do not have equal probability of appearing in positions *a–e*, nor is there an equal probability of odd/even values in each position. Thus, the author elected to use only the positions *f–i* for encryption.

Step 1. The result of this step is the selection of a particular arrangement from a large set of random arrangements, of the ten digits 0–9. The digits 0–9 can be arranged in 3,628,800 unique arrangements ($10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$, mathematically expressed as 10!). The number of SSNs to be uniquely encrypted into RLNs can be used to decide whether it will be sufficient to work with only 10 of the possible 10! arrangements, with 100, or with some other number, up to 10!. Numbering the randomly produced arrangements from 1 to N: if $N = 10$, then the digit at one particular position in the SSN to be transformed can be used to select a particular arrangement; if $N = 100$, the value of the digits in a combination of two positions can be used. By arithmetically transforming digits from the

SSN, various values of N can be used. Thus formulas using positions *f–i* in various combinations can be constructed to select up to 10! arrangements. While the positions selected and their order are arbitrary, once chosen they cannot be changed, or the encryption will result in a different RLN. For example purposes, the digit in position *h* will be used to select one of ten random arrangements of the digits 0–9, as listed in Table 1.

Using the arrangements in Table 1, if the SSN = 315-24-2181, then the value of position *h* is "8" and so the arrangement in line 8 of Table 1 is selected as the result of Step 1: "6937108452."

Step 2. The result of this step is the selection of a particular arrangement of ten characters taken from a large table of randomly produced alphanumeric values, with no duplicates within the ten characters selected. Symbols (#, @, %, etc.) could also be used, as could both uppercase and lowercase characters. The larger the possible number of such arrangements, the more likely it is that the resulting RLN will be unique to the SSN. The alphanumeric values are placed randomly by computer into a computer file consisting of multiple "pages," each consisting of an equal number of rows and columns. One requirement is that an alphanumeric value not be duplicated within any ten consecutive positions within any column or within any row, reading in either direction and wrapping around at row and column ends. If the number of pages does not exceed 100, the value of two of the digits in positions *f–i* in the SSN can be used to indicate the page. If the number is less than 100, then some algebraic transformation can be used to yield the desired range of page numbers. Similarly, two more pairs of digits in positions *f–i* can be algebraically manipulated to yield the desired range of row and column numbers. A fourth pair of digits is read not for numeric value but to indicate the direction of movement from the starting

Table 1: Ten Random Arrangements of the Ten Digits 0 through 9

Arrangement #	*Arrangement*									
0	9	0	3	6	7	5	1	2	8	4
1	6	1	7	2	9	3	8	5	4	0
2	5	8	7	9	1	0	4	3	6	2
3	7	9	5	0	6	1	8	4	3	2
4	4	6	5	8	2	7	0	3	9	1
5	9	8	0	1	3	6	4	2	7	5
6	0	3	6	5	4	2	1	7	9	8
7	4	9	2	3	5	8	0	1	7	6
8	6	9	3	7	1	0	8	4	5	2
9	4	1	5	7	2	8	9	3	6	0

point. By reading these digits as "odd-odd," "odd-even," "even-odd," or "even-even," each combination can be associated with a specific direction (left, right, up, down). Once the "starting point"–the (page/row/column intersection)–is located, ten alphanumeric values are read, starting with the value in the selected cell (as indicated by the page/row/column) and reading in the direction indicated.

For example purposes, we will again use the SSN = 315-24-2181, and will assume that our file consists of 100 pages, each containing 100 rows and 100 columns. The page, row, and column can be read directly from selected pairs of digits. For example, we will use positions *hg* to direct us to page 81, positions *gf* to direct us to row 12, and positions *ig* to direct us to column 11. A fourth pair, *hf*, is even-even, which we will assume had been defined as "down." So we will look on page 81, and locate the intersection of row 12 and column 11. We will record the alphanumeric value in that cell, and in the nine cells below it (reading down within column 11). If an edge of the table on a "page" is reached before the tenth value is obtained, the program "wraps around" to the opposite edge and continues to read in the same direction, until ten values are obtained. For example purposes, let the alphanumeric values obtained from Step 2 for the SSN being transformed (315-24-2181) be "6SC18BMJXG."

Step 3. The third step consists of using the results of the first two steps to produce the actual RLN. In our example, where the SSN was 315-24-2181, the result of Step 1 was the arrangement "6937108452," and the result of Step 2 was the arrangement "6SC18BMJXG." The encryption for this SSN will consist of substituting the corresponding ten alphanumeric values for each occurrence in the SSN of the corresponding digits 0–9. Thus, we substitute a "6" for each "6" in the SSN, an "S" for each "9" in the SSN, a "C" for each "3" in the SSN, and so on. Using this substitution code on the example SSN results in the RLN of "C8XGJG8M8." Once the SSN is encrypted, it can be erased and only the RLN kept on the file.

EVALUATION

The encryption method meets the four criteria described at the outset.

1. *The method may be considered reliable.* Because the method relies on a computer program using stored tables for reference, and does not rely on any random process except for the original preparation of the tables, the encryption will always produce the same RLN for the same SSN.
2. *Virtually unique identifiers are produced.* The criterion of uniqueness was tested by running a file containing 867,535 different SSNs

through the encryption program, which resulted in 867,535 different RLNs. Thus, while it is not impossible that the same RLN can be produced from two different SSNs, it is highly improbable ($p < .00000115$).

3. *The encryption is irreversible.* Even a programmer with access to the tables used in the encryption process, and with knowledge of which digits in the SSN were used in the program, cannot feasibly determine the SSN from which an RLN was derived.
4. *The method can be secured, in several ways.* As this article demonstrates, the encryption method can be described in terms specific enough for its assessment, without compromising its security. The computer program (which identifies the specific digit combinations used in each step) and the two tables used in the method can be separated from each other and maintained in different locations with separate access controls.

DISCUSSION

A researcher may need to identify individual records in a research database, but would have no need to identify individual persons except to contact the person(s) directly or to link future data for those persons with the originally collected data. Some form of individual identifier is required to enable the researcher to link data. Possible identifiers include: complete or partial name, address, SSN, date of birth, parents' names, and, where relevant, date and location of service. Clearly, the more identifiers available, the greater the likelihood of a correct match. Identifiers have been used to link files within one program's data (Welch and Larson 1991) or within programs operated in a state under confidentiality rules that permit sharing of data across programs (Buescher et al. 1991). This article addresses another type of situation, where a single unique identifier needs to be made publicly available to data users under stringent confidentiality rules prohibiting the release of the personal identifier itself. The solution presented is a method to encrypt the personal identifier prior to public release, in a way that meets specific and stringent criteria. Because the parameters of the described method can be varied, the encrypted values will be different for each data program using the method. The method is designed to permit uniquely identified, unduplicated data to be disseminated, linking records over time *within a publicly released* database, without compromising stringent concerns for individual privacy.

Linking records across databases, without violating confidentiality, could possibly be facilitated by use of the method, if the files contain a common identifier, for example, the SSN. If a database is not considered confidential, it could be merged with discharge records, using the SSN to

link the records. The merged file could be made available to researchers with the SSNs encrypted into record linkage numbers (RLNs). When the other file is also confidential, another version of the encryption program could be created on a computer system accessible by the different agencies responsible for the files. Each agency could pass its file with SSNs through the encryption program, which would replace the SSN with the RLN (so that the RLN for a common SSN would be the same). The linking of the two newly encrypted files, using the common RLN, could be done by either agency. The challenge to be met is to construct restrictions during the encryption that will thwart the ability of either agency to copy the method intact, will prevent access by either agency to the other's SSNs or other confidential data, and will gain the trust of each agency that the confidentiality of its data has not been compromised. Obviously, programs without confidentiality constraints are always free to link their files, using any identifiers their data have in common.

Other states and data collection programs are encouraged to explore variations of this methodology as a means of providing individual patient identifiers on publicly released data while maintaining confidentiality by replacing confidential individual identifiers with coded values that can safely be shared. Records can be released after encryption of SSNs and deletion or transformation of other data items that could permit the identification of an individual, thus permitting linkage of records over time. The encryption method has been presented in sufficient detail so that it can be adopted by other programs and implemented by their computer programmers without the need to purchase software or seek technical assistance. The method was described as applied to SSNs, but could be generalized for use on alphanumeric data. Depending on the number of unique values to be encrypted, the number of possible substitution codes could be increased in order to minimize the possibility of different values being encrypted into the same RLN.

ACKNOWLEDGMENT

The author wishes to acknowledge the assistance of Stephanie Majuk-Fox, who translated the method described in this article into a working computer program, and who offered suggestions that resulted in significantly reducing the rate of duplicate RLNs produced by the encryption program.

REFERENCES

Agency for Health Care Policy and Research. *Report to Congress: The Feasibility of Linking Research-Related Data Bases to Federal and Non-Federal Medical Administrative Data Bases.* Rockville, MD: U.S. Department of Health and Human Services, 1991.

Buescher, P., M. Roth, D. Williams, and C. Goforth. "An Evaluation of the Impact of Maternity Care Coordination on Medicaid Birth Outcomes in North Carolina." *American Journal of Public Health* 81, no. 12 (December 1991): 1625–29.

Computer Security Institute. *Buyers Guide.* San Francisco: Miller Freeman, Inc., 1991.

Meyer, C., and S. Matyas. *Cryptography: A New Dimension in Computer Data Security–A Guide for the Design and Implementation of Secure Systems.* New York: John Wiley & Sons, 1982.

Welch, H., and E. Larson. "Patients Requiring at Least Five Admissions in 1 Year." *Medical Care* 29, no. 6 (June 1991): 578–82.