Extraction and Anonymity Protocol of Medical File

Hocine Bouzelat, Catherine Quantin, and Liliane Dusserre
Department of Medical Informatics (Pr. L. Dusserre), Teaching Hospital of Dijon France
Tel. : 80.29.36.29, Fax : 80.29.39.73, e-mail : cquantin@satie.u-bourgogne.fr

To carry out the epidemiological study of patients suffering from a given cancer, the Department of Medical Informatics (DIM) has to link information coming from different hospitals and medical laboratories in the Burgundy region.

Demands from the French department for computerized information security (Commission Nationale de l'Informatique et des Libertés : CNIL), in regard to abiding by the law of January 6, 1978, completed by the law of July 1st, 1994 on nominal data processing in the framework of medical research have to be taken into account. Notably, the CNIL advised to render anonymous patient identities before the extraction of each establishment file.

This paper describes a recently implemented protocol, registered with the French department for computerized information security (Service Central de la Sécurité des Systèmes d'information : SCSSI) whose purpose is to render anonymous medical files in view of their extraction. Once rendered anonymous, these files will be exportable so as to be merged with other files and used in a framework of epidemiological studies. Therefore, this protocol uses the Standard Hash Algorithm (SHA) which allows the replacement of identities by their imprints while ensuring a minimal collision rate in order to allow a correct linkage of the different information concerning the same patient. A first evaluation of the extraction and anonymity software with regard to the purpose of an epidemiological survey is described here. In this paper, we also show how it would be possible to implement this system by means of the Internet communication network.

INTRODUCTION

To carry out the epidemiological study of patients suffering from a given cancer, the Department of Medical Informatics (DIM) has to link information coming from different hospitals and medical laboratories. Thus, for example, for the follow-up of lung cancer patients, it is necessary to have anatomo-pathological information in order to check the diagnosis of this cancer but also to get information concerning hospital care of these patients in this region.

Each hospital in the Burgundy region has a file with discharge abstract. Anatomo-pathological laboratories have centralized their files in a regional center which allows the obtaining of information about all patients suffering from a given cancer. Moreover, if we manage to link anatomo-pathological files and hospital files, we will be able to keep up with the hospital care and cost of all patients. However, on the one hand, the linkage of nominal files within the framework of medical research is submitted to French law n° 78-17 of January 6th 1978, completed by law n° 94-548 of the July 1st 1994 relative to nominal data processing. Thus, the French department for computerized information security (Commission Nationale de l'Informatique et des Libertés : CNIL), advised us to render anonymous each file, in an irreversible way, before its export, so that returning to the nominal information becomes impossible. On the other hand, implementation of encryption methods in France is submitted to the law which specifies that the use of any encryption system has to be declared to the French department for computerized information security (Service Central de la Sécurité des Systèmes d'information : SCSSI). This department advised us to use a hash method[1] for which an authorization is not necessary but which only needs to be registered.

This paper describes a recently implemented protocol that we have developed on the basis of a one-way hash function with the purpose of rendering anonymous medical files before their extraction, in view of their linkage. Once rendered anonymous, these files will be exportable so as to be merged with other files and used in a framework of epidemiological studies. Therefore, this protocol uses a one-way hash function[2] which allows the replacement of identities by their imprints while ensuring a minimal collision rate, i.e. the ratio between the number of differents strings producing the same result, and the total number of strings, in order to allow a correct linkage of the different information concerning the same patient. A declaration of this protocol has been sent to the SCSSI on January 1996. A first evaluation of the extraction and anonymity software with regard to the purpose of an epidemiological survey is described here. In this paper, we also show how it would be possible to implement this system by means of the Internet communication network.

Although the exchange of files from the different laboratories and those from the DIM has

0195—4210/96/$5.00 © 1996 AMIA, Inc.

been done on floppy disk, we are displaying here a more general protocol which utilizes network communication tools. In this case, we assume that each laboratory participating in the study has access to public-key cryptosystem. The protocol developed here can be used for any file with a view to linkage. It is not restricted to medical files linkage and may therefore serve as a model for comparable applications as census for example.

METHOD

For each study, the DIM has to generate a pad, i.e. a large random file, (k_1[3]) and send it to each hospital or laboratory participating to the study by using an encrypt function, i.e. to transform an intelligible cleartext in an unintelligible ciphertext, so that only the legitimate recipient is able to decipher this ciphertext, thus, recovering the cleartext. RSA, for its inventors Rivest, Shamir and Adleman)[4],[5] (see Figure 1) is one of very first public-key cryptosystem. This type of cryptosystems uses two keys: a secret-key and a public-key. Only the recipient is in possession of the secret-key that is the decrypting key.

The same pad k_1, initially generated by the DIM, is transmitted, after encryption (k'_1), to the different laboratories and hospitals participating in the study. Each laboratory decrypts what it has received by using its RSA secret-key to find the pad k_1. The pad k_1 must be kept secret by the different laboratories participating in the study and destroyed after utilization. Each laboratory can then perform the extraction and anonymity software on their file and address this file to the DIM.

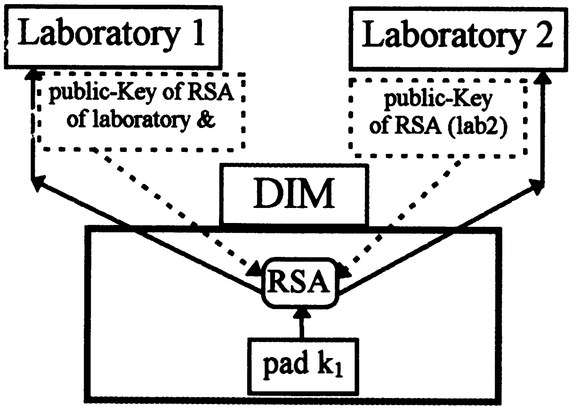

Figure 1

Dispatch of the pad k_1 to the different laboratories

Once received in the DIM, the file is hashed[6] again with the same function (using SHA) but with a second pad k_2, and the initially received file is destroyed. The confidentiality of information stored in the DIM is thus ensured since even members of laboratories that participate in the study and that hold the pad k_1 cannot return to the identity information. The physician of the DIM takes on responsibility for the security of the two pads k_1 and k_2.

The following diagram (see Figure 2) shows the various transformations of anonymity in a medical file.

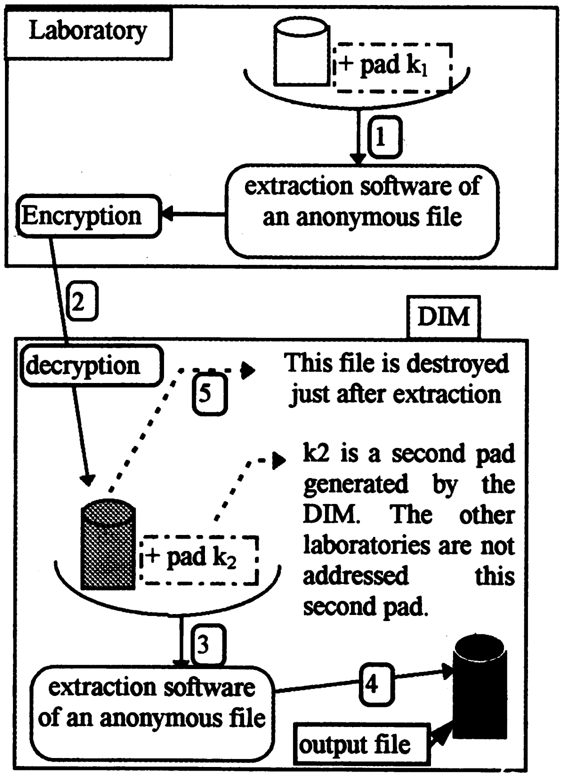
Laboratory
+ pad k_1
1
extraction software of an anonymous file
Encryption
2
DIM
decryption
5
This file is destroyed just after extraction
k2 is a second pad generated by the DIM. The other laboratories are not addressed this second pad.
+ pad k_2
3
extraction software of an anonymous file
4
output file

Figure 2

Import, to the DIM, of an anonymous file from a laboratory

Nominal information is therefore hashed twice: the first time in the laboratory with pad k_1 and the second time in the DIM with pad k_2.

Description of the extraction and anonymity software and its utilization

The extraction of an anonymous file necessitates three stages: the first one consists in a conversion program from the storage format of the laboratory to a standard format. The second stage consists in performing the extraction and anonymity program which includes of 8 steps:

step 1: read fields corresponding to name, date of birth, address zip code and sex,
step 2: orthographic processing [7], used to minimize typing errors and to merge homonyms. It consists of several rules[8], for example in order to reduce repeated characters into a single character.
step 3: concatenate the different fields in a variable "string" of permanent size,
step 4: calculate the exclusive-or (XOR) with a portion of the pad according to Figure 3:

The determining of this portion is as follows:

Pad k_1 is considered as a set of pages[9], each page constituted of several rows and columns. The idea consists in choosing a page, then a row and a column of this page. From this character (the intersection of the selected row and column) and by following a direction chosen according to one character of the string 's' to hash, a portion with the same length as the size of this 's' is extracted from this page. If the number of pages does not exceed 255, the ASCII code of a character of the 's' to hash can be used to indicate the page to be selected.

The parity of the two ciphers of the ASCII code of one character of 's' gives the direction. As for example: even-even, even-odd, odd-even, odd-odd to designate the four directions left, right, high and low respectively. For characters with ASCII code inferior to 10, the cipher 0 is added, 1 becomes 01.

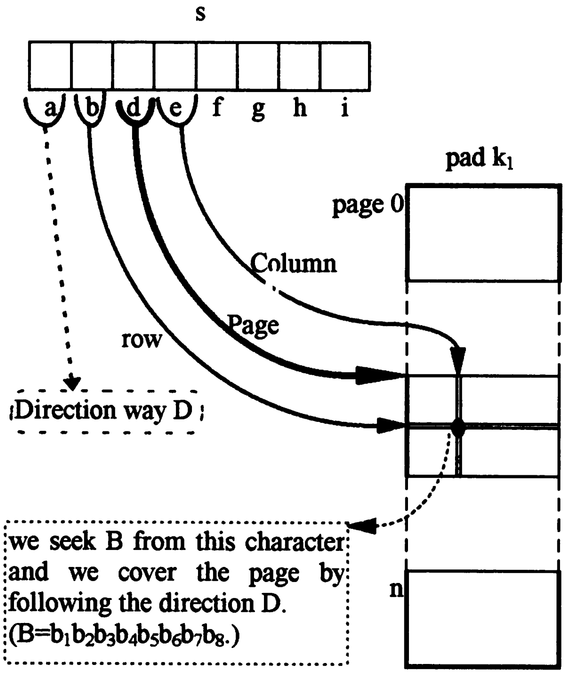
s
a
b
d
e
f
g
h
i
pad k1
page 0
Column
Page
row
Direction way D
we seek B from this character and we cover the page by following the direction D. (B=b1b2b3b4b5b6b7b8.)
n

Figure 3

After having selected the salt B (salting is the technique that nearly eliminates dictionary attack), we can apply the bitwise exclusive-or (⊕) between 's' and 'B' i.e. $0\oplus0=1\oplus1=1$, $0\oplus1=1\oplus0=1$
step 5: to apply the hash function to the chain produced by the preceding step
step 6: to put the date of birth in the variable string and to retake steps 4 and 5
step 7: to put the zip code in the variable string and to retake steps 4 and 5
step 8: the different code obtained and the remaining anonymous information of input file are inserted in the corresponding record of the output file. This output file is then sent[1] to the DIM.

All these steps can be schematized as follows:

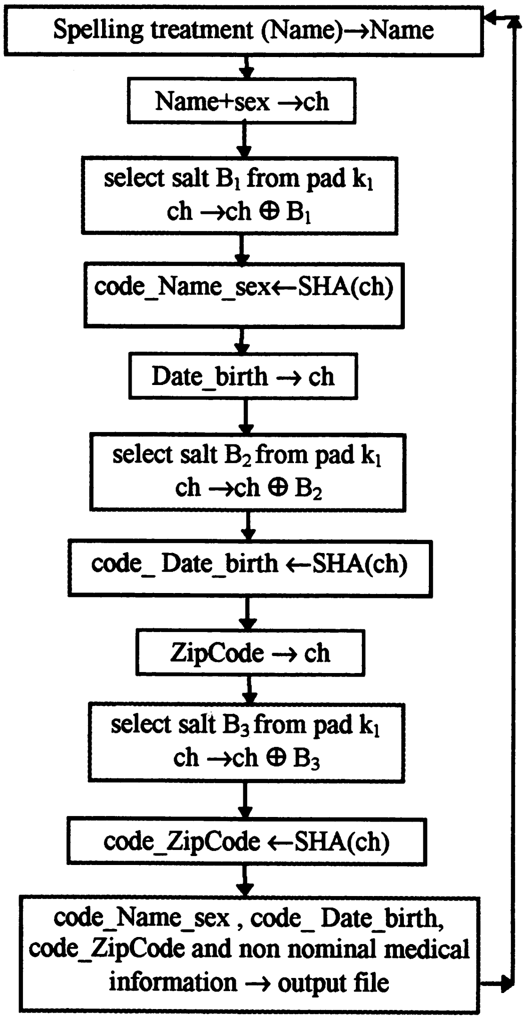
Spelling treatment (Name)→Name
Name+sex →ch
select salt B1 from pad k1
ch →ch ⊕ B1
code_Name_sex←SHA(ch)
Date_birth → ch
select salt B2 from pad k1
ch →ch ⊕ B2
code_ Date_birth ←SHA(ch)
ZipCode → ch
select salt B3 from pad k1
ch →ch ⊕ B3
code_ZipCode ←SHA(ch)
code_Name_sex , code_ Date_birth, code_ZipCode and non nominal medical information → output file

Figure 4

Flow chart of the software of the first extraction with pad k1

[1] The dispatch of the imprint will be made by floppy disk. The piracy of this floppy disk and the pad will allow an dictionary attack.

3. Encryption of hashed file
The hashed file is encrypted by using RSA algorithm with the public key of the DIM and exported to it.

This software is used by laboratories and by the DIM for its own files.

Each time that the DIM receives an anonymous file, it decrypts it and hashes it a second time with a second pad k_2 by therefore coding variables code_Name_sex, code_Date_birth and code_Zipcode.

RESULTS

Application

We have applied our method to two files, the first concerning the administrative data base of Dijon public hospital containing 264 458 records and the second concerning the anatomo-pathology center (for patients with digestive cancer) containing 305 records without doubloons.

Security

Files and transmitted laboratory key are protected according to several levels:

1. Pad k_1 is encrypted before being sent to several laboratories. For each laboratory we use its RSA public-key and the encryption security is that RSA.
2. If the laboratory file is intercepted before its arrival at the DIM, an 'only-ciphertext attack' would be very difficult, not only due to the complexity of cryptanalysis of RSA but also due to the fact that the deciphering provides a file hashed with a one-way function using a pad.
3. Once the file is at the DIM, it is deciphered using the RSA private key, then hashed a second time. So that, a dictionary attack is not possible.

DISCUSSION

Patient level data is critical to epidemiological research, particulary within the framework of survival studies requiring the follow-up of each patient through a long period. The hash method that we proposed for the linkage of nominal files must respect two important constraints: the first constraint is to ensure anonymity, thus providing collision. The seconde one is to allow a correct linkage of the information concerning the same patient, and requires, as a consequence, a minimization of the collision rate in order to reduce typing errors. Regarding the first constraint of anonymity, the SHA function was chosen because it produces collisions and it is impossible to find a message with a specific imprint by computation. Moreover, no attack is more efficient than the exhaustive attack. The SHA is not based on any hypothesis, such as the difficulty of factorization used in RSA and produces an imprint of 160 bits which renders the exhaustive attack impossible. Moreover, the addition of the fifth variables (as compared to MD4 and MD5) renders the attack of DEN Boer-Bousselaers against MD5 impossible against SHA.

Concerning the second constraint of correct linkage, the importance of the spelling treatment is easy to understand. It allows a reduction in the number of codes which can be obtained from phonetically identical names. This is profitable when the same name is spelled in different ways due to typing errors, and precludes assigning several codes to the same name, which would render the linkage of one patient's records impossible. Indeed, of the 17% collision rate due to spelling treatment, 6.4% is beneficial, i.e. allows correction of typing errors.

- The use of SHA is not a simple adaptation but we have brought several modifications, namely:

- The use of the spelling treatment that allows correction 2% due to typing errors.

- Optimization of SHA code, obtained from SCSSI, has been brought by using powerful operators of language C.

Although the collision rate of SHA is not completely null, the combined rate tends to zero, since we use it on several variables.

In order to estimate the collision rate due to spelling treatment and considered beneficial for correcting typing errors, we have established the hypothesis that records with the same family name after spelling treatment, the same date of birth and the same sex are connected with the same person. Of course, this definition is not perfect because cousins or twins who were born on the same date may be confused, if first names are not taken into account.

However, the first name was not included in these criteria because it may be filled in with the maiden name and because one of the two names only is sometimes recorded in a compound given name.

For this reason, fields have been crossed so as to correct this type of errors as follows. In Table 1, the name which appears in the first column (in file 1) is compared with the name in the second column (in file 2). For example, 3 records have been corrected because the family name of the first file corresponded to the maiden name of the second file.

File 1	File 2	corrected records
Family name	Maiden name	3
Maiden name	Family name	4
First name	Family name	2
Family name	First name	0

Table 1

Furthermore, the comparison is not made character by character but rather by inclusion. Thus, the comparison of a composite family name with another name allows the identification of these names if the second one is included in the first one, for example, « Robert Kevin » and « Robert » are supposedly identical if they have the same sex and date of birth.

To each variable is attributed a weight[10]: two records having, for example, the same name are not considered in the same way as two records having equal sexes.

CONCLUSION

An adaptation of a one-way hash-function SHA, is proposed to allow the chaining of medical information of patients within the framework of epidemiological follow-up. The SHA function is proposed by the NSA to be used in Digital Signature Algorithm (DSA) which was part of the Digital Signature Standard (DSS).

The linkage of patient information can then be accomplished while respecting the confidentiality of medical data which becomes anonymous by the irreversible encryption of the patient's identity.

The use of a pad discourages decryption by dictionary attack. The security of this method is reinforced by the use of two pads which can be separated from each other and maintained in different locations with separate access controls. In our example, the first pad is shared by laboratories, producers of the information, and the second one by the recipient (DIM) only. So that the producer could not decrypt the ciphertext.

Acknowledgment

This research has been supported by means of a grant from the Burgundy Regional Council.

References

1 . Vaudeney S. La sécurité des primitives cryptographiques. Thesis of University of Paris 7. Sustained in April 1995.

2 . Beckett B. Introduction aux méthodes de cryptologie, Masson, 1990.

3 . Schneier B. Applied Cryptography, Protocols, Algorithms, and Source Code in C. John Wiley & Sons, Inc., 1994.

4 . Rivest RL, Shamir A, Adleman L. A Method for obtaining Digital Signatures and Public-Key Cryptosystems, *CACM*, 1978; 2: 120

5 . Zimmermann P. A Proposed Standard Format for RSA Cryptosystems, *Boulder Software Engineering, Computer*, 1986; 9: 21

6 . Brassard G. Modern Cryptology. *Lecture Notes in Computer Science,* 1993: 325.

7 . Thirion X, , R. Sambuc, San Marco JL. Epidemiology and anonymity: a new method. Rev. Epidém et Santé Publ, 1988; 36: 36-42

8 Dusserre L., Quantin C., Bouzelat H., A one way public-key cryptosystem for the linkage of nominal files in epidemiological studies. *MEDINFO'95, R.A. Greenes, H.E. Peterson, D.J. Protti (editors), Elsevier Science Publishers (North-Holland),* 1995; 661-665

9 . Meux E. Encrypting personal identifiers. *Health Services Research*, 1994; 29(2): 247-56.

10 . Fiona J, Stanley, Maxine L, Croft. A population database for maternal and child health research in Western Australia using record linkage. *Paediatric and Perinatal Epidemiology* 1994; 8: 433-447