Linking anonymous databases for national and international multicenter epidemiological studies: A cryptographic algorithm

Catherine Quantin, Maniane Fassa, Gouenou Coatrieux, Benoît Riandey, Gilles Trouessin, François André Allaërt

Linking anonymous databases for national and international multicenter epidemiological studies: A cryptographic algorithm

Chaînage de bases de données anonymisées pour les études épidémiologiques multicentriques nationales et internationales : proposition d'un algorithme cryptographique

C. Quantin [a,b,*], M. Fassa [b], G. Coatrieux [c], B. Riandey [d], G. Trouessin [e], F.A. Allaert [f,1]

[a] *Inserm U 866, université de Bourgogne, Dijon, France*
[b] *Service de biostatistique et informatique médicale, CHU de Dijon, BP 77908, 21079 Dijon cedex, France*
[c] *Inserm U 650, LaTIM, institut Télécom, Télécom Bretagne, Bretagne, France*
[d] *Institut national d'études démographiques (Ined), Paris, France*
[e] *OPPIDA Sud, Toulouse, France*
[f] *Chaire d'évaluation médicale Ceren Esc, Dijon, France*

Background. – Compiling individual records coming from different sources is very important for multicenter epidemiological studies; however, European directives and other national legislation concerning nominal data processing must be respected. These legal aspects can be satisfied by implementing mechanisms that allow anonymization of patient data (such as hashing techniques). Moreover, for security reasons, official recommendations suggest using different cryptographic keys in combination with a cryptographic hash function for each study. Unfortunately, this type of anonymization procedure is in contradiction with common requirements in public health and biomedical research because it becomes almost impossible to link records from separate data collections where the same entity is not referenced in the same way. Solving this paradox using a methodology based on the combination of hashing and enciphering techniques is the main aim of this article.

Methods. – The method relies on one of the best-known hashing functions (the Secure Hash Algorithm) to ensure the anonymity of personal information while providing greater resistance to dictionary attacks, combined with encryption techniques. The originality of the method lies in how the hashing and enciphering techniques are combined: as in asymmetric encryption, two keys are used but the private key depends on the patient's identity.

Results. – The combination of hashing and enciphering techniques greatly improves the overall security of the proposed scheme.

Conclusion. – This methodology makes the stored data available for use in the field of public health for the benefit of patients, while respecting legal and security requirements.

Résumé

Position du problème. – Pour conduire des études épidémiologiques multicentriques nationales ou internationales, il est souvent nécessaire de rapprocher des informations d'un même patient, provenant de plusieurs sources. En Europe, le chaînage des fichiers nominatifs, dans le cadre de la recherche médicale, est soumis à la directive européenne du 24 octobre 1995, qui requiert que l'information soit rendue anonyme avant son utilisation à des fins de chaînage. La méthodologie du hachage permet de résoudre le problème de l'anonymisation des données, notamment en

* Corresponding author.
E-mail address: catherine.quantin@chu-dijon.fr (C. Quantin).
[1] Chairman IMIA WG IV "Data security".

santé. Par ailleurs, pour des raisons de sécurité, il est recommandé d'utiliser des clés différentes pour chaque étude. Malheureusement, cette recommandation est en contradiction avec les besoins de chaînage. L'objectif de cet article est de proposer une méthodologie innovante pour répondre à la fois aux exigences en matière de sécurité des informations médicales, tout en permettant le chaînage des données relatives à un même patient et leur exploitation statistique.

Méthodes. – La méthode repose sur l'utilisation, pour le hachage, de la fonction Secure Hash Algorithm (SHA), qui permet d'assurer l'anonymat des données personnelles, qui est combinée avec des techniques de chiffrement. L'originalité de la méthode réside dans la manière dont le hachage et le chiffrement sont combinés : comme dans les méthodes de chiffrement asymétrique, nous proposons l'utilisation de deux clés, mais avec une différence fondamentale, puisqu'une des deux clés va dépendre de l'identité du patient.

Résultats. – La combinaison du hachage et des techniques cryptographiques assure une amélioration importante dans la sécurité des données, tout en permettant le chaînage des données multicentriques.

Conclusion. – Cette méthode rend disponibles les informations rendues anonymes et stockées dans des bases de données multicentriques nationales et internationales, pour une exploitation à des fins épidémiologiques et de recherche clinique. Cela, en respectant les exigences de sécurité imposées par les lois nationales et européennes.

Keywords: Security; Patient identification; Encryption; Hashing; Linkage; Multicenter studies; Anonymized data

Mots clés : Sécurité ; Identification du patient ; Chiffrement ; Hachage ; Chaînage de données ; Études multicentriques ; Chaînage ; Données anonymisées

1. Introduction

To conduct national or international multicenter epidemiological studies, one must often compile individual patient records coming from different sources. In Europe, linking nominative files for medical research purposes is subjected to the European directive of 24 October 1995, which requires that the information be made anonymous before it can be used for linkage purposes. To respect this legislation, anonymization procedures must be used. The solution that we have proposed [1] uses an irreversible cryptographic method that is applied to each file before linking.

The hashing technique can solve this problem of data anonymization, notably in the healthcare sector [1–3]. Hashing provides irreversible transformation of patient identity and thus protects patients' privacy. However, since hashing functions are in the public domain, dictionary attacks are a major security problem. This is particularly problematic when the data are collected from several sources and combined at the national level, as when data must be collected by several institutions and chained together. For the Program for the Medicalization of Information Systems (*programme de médicalisation des systèmes d'information*, PSMI), for example, all French healthcare institutions need to be able to use the same cryptographic key so that these keys can later link all the usable data on a single patient at a national level. Yet, if the same cryptographic key is used by different institutions and these data have not been secured, it is possible for one of them to apply the hashing algorithm with this cryptographic key to a dictionary (a large number of identities) in a dictionary attack [4,5] and thus find the identity of the data stored at the national level. If data are not secured and a member institution manages to access the national data, it could know the clientele of competing institutions, for example. To secure this nationally stored data, the French Data Protection Authority (la Commission nationale de l'informatique et des libertés, CNIL) has recommended using a second-level anonymization function, the same hashing function but with another cryptographic key that is only known by the national center.

In addition, for security reasons, using different cryptographic keys is recommended (particularly, for the second hashing) for each study. Unfortunately, this recommendation thwarts data linkage because it becomes almost impossible to link records from separate data collections where the same entity is not referenced in the same way.

The objective of this article is to propose an innovative methodology to respond to the requirements in terms of medical information security while making it possible to link individual patient data and use them for statistical purposes. This means being able to securely and legally make it technically possible to match secured data that were not initially intended to be matched and thus link data between databases, which is very useful in multicenter studies, in the conditions stipulated by the law and the CNIL.

2. Methodology

This article does not deliberate the question of the choice of patient identifiers, which depends on the applications or studies concerned. In particular, this choice should be adapted to the legislation, norms, and usage in force in France and to the structures concerned (healthcare institutions, regional networks, national organizations).

2.1. Review of the cryptographic function of hashing

The use of hashing functions is recent in the world of modern cryptology [6,7]. They were developed so that digital secured signature techniques could be elaborated. Hashing functions are said to be one-way because the calculation of their inverse is considered impractical within a reasonable time for reasons related to the Shannon theory, given today's technology. The hashing function transforms a plain text of any length into a hashing value of a fixed length, often called a fingerprint (e.g., 160-bit output of the SHA-1 function).

Among the many hashing functions proposed by cryptologists, the function that is considered the most secure is the Secure Hash Algorithm (SHA) [8,9], recognized as the American standard by the National Institute for Standards and Technology (NIST). The SHA-1 hashing function, which provides a 160-bit signature, is integrated into the Digital Signature Algorithm (DSA), which was proposed by the NIST in 1991. SHA-1 has since shown security flaws and was improved with a new series of algorithms, SHA-2. Since 2006, the NIST has recommended replacing SHA-1 with an SHA-2 series hashing function (notably, SHA-256).

The probability that two different identifiers have the same fingerprint after hashing (the collision rate) is on the order of 10^{-48} for the SHA-1 function and remains even lower for the SHA-2 function since the length of the result of SHA-2 hashing is even longer.

Given these properties, this hashing function is usually used to verify data integrity. The fingerprint obtained after hashing is specific to the initial message. In particular, a slight modification of the message leads to a radically different fingerprint (a principle referred to as the avalanche effect). To ensure that data are secured during their transmission, an encryption method can be used. Encryption corresponds to the transformation, using an encryption key, of a message expressed without encryption (called plaintext) into a message expressed in an incomprehensible format (called ciphertext) if one does not have the cryptographic key available. Symmetrical cryptographic keys are founded on a single key to encrypt and decrypt a message. The problem with this technique is that the key, which must remain strictly confidential, must be transmitted to any correspondent in a secure fashion. To resolve the problem of key exchange, asymmetric decryption was developed in the 1970s. This method is based on the principle of two keys: one public key allowing encryption and one private key allowing decryption. As indicated by its name, the public key is made available to anyone who wishes to encrypt a message. This message can only be decrypted with the private key, which must remain confidential. The sender sends both the plaintext message and the signed fingerprint (a signature corresponds to an asymmetric encryption carried out by the sender with the private key). To be sure of the message's source and integrity, the addressee first calculates the message's fingerprint with the same hashing algorithm as that used by the sender, then compares the resulting fingerprint with the fingerprint extracted during signature authentication (signature authentication corresponds to the asymmetric description carried out by the addressee with the sender's public key). The addressee can thus be sure:

- that the message sender is indeed the signatory of the message received, since this person is the only one to know the private key used to sign the fingerprint (using asymmetric encryption with the private key);
- that the corresponding public key is the only key that can carry out the decryption (through signature authentication by asymmetric decryption with the public key).

A public key authority (also called the public key infrastructure) generates or receives a public key and certifies it: it generates a certificate containing the public key and signs everything with its private signature key so as to ensure the authenticity and integrity of this public key.

2.2. Use of hashing functions to anonymize and link patient data

We proposed using hashing techniques to ensure the anonymity of personal information as early as 1995, to solve the problem of nominative medical information linkage in multicenter epidemiological studies. When grouping medical information within an organization outside of the care institution, the CNIL [9] recommends irreversible data transformation.

After attempting to improve existing methods proposed by Thirion et al. [10], in 1995 we suggested to the CNIL that one-way hashing methods be used to ensure anonymity. Soon after in 1996, the CESSI/CNAMTS[2] designed and provided an anonymization function called Nominative Information Removal Function (*fonction d'occultation d'information nominatives*, FOIN) [11] to set up the PMSI for private organizations as recommended by the CNIL (which suggested using the algorithm developed by the Dijon University Hospital's Department of Medical Informatics). Contrary to the encryption methods that should be reversible so that the legitimate addressee can decrypt the message, hashing methods are irreversible. The result of hashing is a strictly anonymous code (which cannot return to the patient's identity), but which is always the same for a given individual so that a patient's information can be compiled. In agreement with the Central Service for Information Security (*service central de la sécurité des systèmes d'information*, SCSSI), we chose the SHA algorithm, which, to our knowledge, is the most secure public-domain hashing algorithm to counteract decrypting attempts [1,6–9,12,13].

The progression of cryptography led us to use the SHA-2 family algorithms, more relevant than the SHA-1 version, which had become obsolete. The procedure was announced to the CNIL and SCSSI in March 1996. This solved the problem of linking information from a single patient within a multicenter study. However, since the hashing algorithm was public, data security depended on the cryptographic key used. Indeed, as explained in the introduction, someone who knew the cryptographic key could apply the hashing to a large number of identities and proceed to a dictionary attack: this person could compare the codes obtained to the codes of a given individual in the hashed file and find this person's identity. To prevent this dictionary attack, the use of different hashing keys for each study was recommended. The same hashing algorithm (e.g., SHA) was therefore used with a cryptographic key that could differ from one study to another. Thus, with the same

[2] Center for the study of information system security (CESSI) of the French National Health Insurance system for salaried workers (CNAMTS).

identifier and the same algorithm, different fingerprints were obtained depending on the key used. Since the results of hashing the same identity from two different keys were completely disconnected, it was not possible to chain the same person's data from different studies. For example, in Belgium, based on hashing the social security number, authorities have planned to use different keys to make up distinct identifiers for reimbursing healthcare acts, medical care, and an identifier for each type of research. It then becomes very difficult to link the data produced by the different sources. This article attempts to propose an innovative methodology to overcome these problems based on the combined use of hashing techniques and encryption, as we have already proposed in collaboration with our Swiss colleagues [14]: a person's identity data (name, date of birth, and sex) are first made anonymous by hashing and then secured by an encryption method. The originality of the approach proposed in this article is based on the encryption key protection method.

2.3. Proposal for an encryption method combined with hashing

Fig. 1 shows that the method starts with the patient identity number (PIN) made anonymous by double hashing DH(PIN), as was obtained in the preceding step. To secure this ID number and prevent a dictionary attack on this number, a double encryption procedure will be set up. This system is therefore based on two cryptographic keys, a single key P_w defined for the study and a variable key I_k, which depends on the patient's identity; thus, providing additional security. Only the knowledge of these two keys will allow one to encrypt the anonymized ID number and, vice versa, to decrypt it. This means that only those who are duly authorized can access the ID number that has been anonymized, the linking landmark. It is possible to hash the message again, using the hashing function H, the anonymized ID number DH(PIN). The I_k value obtained is the new encryption key. Then the encryption function C will be applied, with the key I_k, to the anonymized ID number DH(NIP) to secure it and thus obtain C_{I_K} [DH(PIN)]. It is also possible to apply the same encryption function C to I_k itself to secure it, but this time with the study's encryption key, that is, P_w. This makes potential attacks much more difficult to carry out on the stored number, since the encryption key I_k that the attacker would have to discover depends on the patient's identity. In addition, transmission of this key is also protected by use of the encryption key C on the same key.

Nevertheless, it remains possible to relate a single patient's data, in conditions secured by returning to the anonymized ID number DH(PIN). If the key P_w and the encryption key C are known, it suffices to decrypt I_k (Fig. 2). Then, the anonymized ID number DH(PIN) can be recovered. Clearly, this does not mean obtaining the patient's identity in an uncoded form, but rather the anonymous ID number DH(PINP), which had been protected from any dictionary attack by encryption.

This being said, this procedure is extremely powerful: fully secure and legal, it can match protected data from several centers. This assumes the existence – both legal and CNIL-approved – of an authority that manages the cryptographic keys, which would hold all the P_{wi} encryption keys used by the different studies.

Knowledge of this anonymous ID number DH(PIN) would allow this authority to link the data from different studies, without knowing the identity of the patients involved in these studies. The key authority would know the P_{wi} keys used by the different studies. Following a request made jointly by the two principal investigators and CNIL authorization, it can return the common denominator of the patient identifier in the different studies to the anonymous identity DH(PINP). However, despite their common source, a patient's identifiers are completely different from one study to another. As shown in Fig. 3, the key authority can find I_{k1} (respectively, I_{k2}) by decrypting $C_{Pw1}(I_{k1})$ (respectively, $C_{Pw2}I_{k2}$). Consequently, it

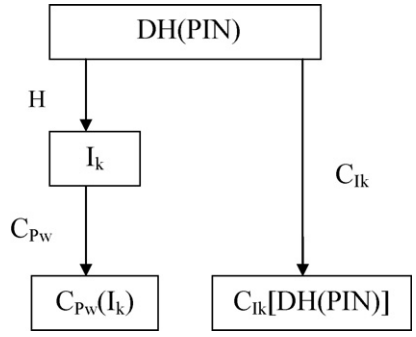

Fig. 1. Securing data by encrypting the anonymized ID number.
H = hashing function; C = encryption function; DH: double hashing function; I_K: variable key depending on the patient's identity; P_W: single key defined for the study.

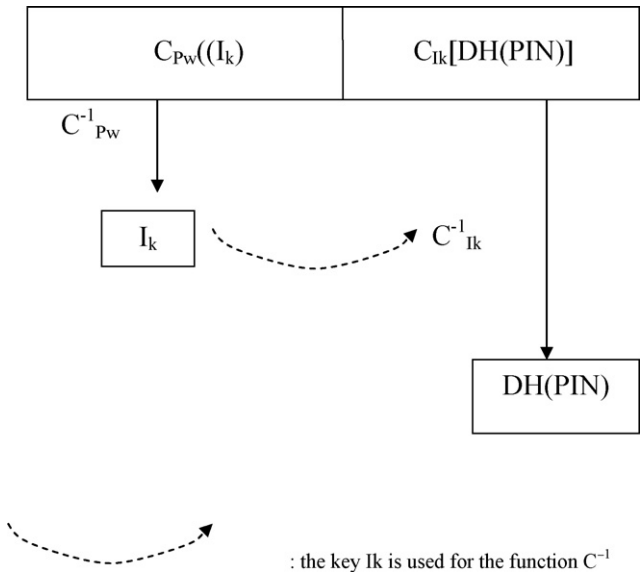

Fig. 2. Decryption of the anonymous ID number.
H = hashing function; C = encryption function; C^{-1} = decryption function (associated with C); DH: double hashing function; I_K: variable key depending on the patient's identity; P_W: single key defined for the study.

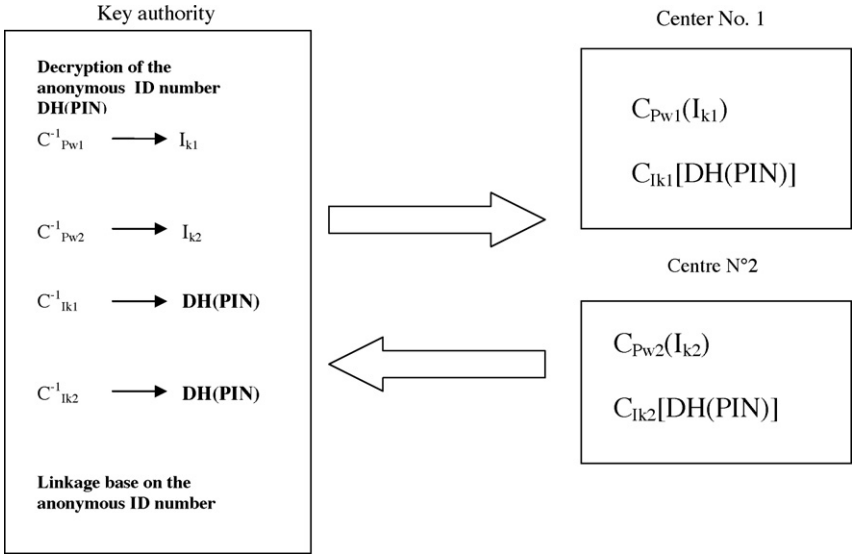

Fig. 3. Inter-center linking based on the anonymous ID number by the key authority.
H = hashing function; C = encryption function; DH: double hashing function; I_K: variable key depending on the patient's identity; P_W: single key defined for the study.

can decrypt C_{Ik1}[DH(PIN)] as well as C_{Ik2}[DH(PIN)] and find DH(PIN) for each of the centers. The key authority is therefore able to link the inter-center data. Remaining anonymous, it then suffices to proceed to recording newly matched information (to reuse it according to the epidemiological objectives to be met) using the technique combining double hashing and encryption that we have described above.

3. Discussion

3.1. Advantages of the data linking method

The advantage of this method is that it can relate a patient's data, when this match is authorized by the CNIL, while ensuring the security of these data. Each center will store patient data with an ID number that belongs only to that patient. It is therefore impossible for someone outside the center to discover the patient corresponding to the data stored by the center. Nevertheless, given that the construction of each center's identifier is based on a similar method, particularly on the same hashing key that anonymizes the patient's identity, it is possible for the key authority, but only this authority, to return to this anonymous ID number by decrypting. This decryption relies on two conditions:

- that the two centers wishing to match their data have authorized this authority to do so;
- that they have obtained the CNIL's agreement.

As for the key authority, it cannot come back to the patient ID data because it does not store the different centers' data. The centers only transmit anonymized and encrypted identity data for matching, with no medical data.

If necessary, this organization unblocks the specialization of data created by irreversible hashing of a SSN[3] hashed by the PMSI or the SNIIRAM and, in addition, hashed as a cohort identifier. The solution that had been proposed previously was the Identifier Coordination Authority (*instance de coordination des identifiants*, ICI), a generalized trusted third party,[4] that would preserve the lookup tables, a logic that does not fit hashing well [15]. The interest of the method lies in the fact that abandoning these lookup tables for simple conversion of these keys (hence, the name of the authority), the method is much better secured (i.e., less at risk) than the method stemming from the use of these tables. The utility of this authority stems from the CNIL's policy of sectoring the identifiers of public statistics and from the need to be able to communicate between sectors to transmit data or validate the sectorial identifier.

Linking data requires substantial resources that can be put to work over the long term. To assume the role to be played by a secured matching center of individual local, regional, and national databases, a national agency for linking data should be created, as was done in Australia at the end of the 1990s, which Marcel Goldberg et al. [16] unambiguously showed to be highly advantageous for France.

The value of linking a patient's data coming from large databases; notably, within multicenter epidemiological or clinical research studies, extends well beyond a single country's

[3] Social Security Number (*numéro d'inscription au répertoire*) of the National Insititute for Statistical and Economic Studies (INSEE), National Identification Register of Private Individuals (*répertoire national d'identification des personnes physiques*, RNIPP).

[4] For ethical reasons with regard to patients, epidemiological studies can demand that a re-identification procedure be maintained. Conversion of a nominative correspondence table by a trusted third party is hence unavoidable.

reference. It seems illusory to wait for a single identifier to be created for healthcare at the European level (the United States does not have one). Europe's current policy is to promote the interoperability of the existing identifiers [17–19], which runs the risk of requiring a great deal of time. Also, the method proposed in this article should make it possible to immediately pull down the main obstacles to conducting European and international multicenter research studies.

3.2. Choice of the encryption method

The choice of the encryption method made in this article deserves to be debated. In particular, the advantages of an encryption method with a public rather than a secret key should be discussed. However, the latter are, in principle, faster. Nevertheless, here the objective is encryption to store data: the speed of encryption is not the priority. Using a public key algorithm that provides a different key for encryption and decryption without doubt improves security, without slowing down encryption excessively. This is where the problems of protecting the transmission of the secret keys are found, as with any symmetric cryptographic method. The asymmetric processes avoid this problem. Only the public key [20], corresponding to a given study, is transmitted to those in charge of the data sources to encrypt – that is, secure – the anonymized ID number DH(PIN) (Fig. 1). The data stored in this manner at the source level but also at the study's data processing center can only be decrypted by those holding the private key, that is, the person in charge of the study's data processing center and the key authority. Thus, only the key authority knows the private keys of all the studies and is therefore able to find the anonymized ID number DH(PIN) that is common to all the studies. This is the principle that should be generalized.

4. Conclusion

This methodology can respond to the need to establish a balance between two of the main pillars of information security: protection of privacy and availability of information. What would be the use of having enormous cemeteries of data that are extremely well protected, but unusable for public health purposes? Even if the solutions may seem somewhat awkward to implement and time-consuming, this is nothing compared to the time required to collect the thousands of data that could be used while respecting citizens' privacy and making use of the expected benefits to public health in both epidemiology and clinical research.

French version

A French version of this article is available at doi:10.1016/j.respe.2008.10.010.

References

[1] Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods Inf Med 1998;37:271–7.

[2] Blakely T, Woodward A, Salmond C. Anonymous linkage of New Zealand mortality and Census data. Aust N Z J Public Health 2000;24:92–5.

[3] Churches T, Christen P. Some methods for blindfolded record linkage. BMC Med Inform Decis Mak 2004;28:4–9.

[4] Quantin C, Gouyon B, Allaert FA, Dusserre L, Cohen O. Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales. Courrier Stat 2005;113–114:15–26. Journal de la SFdS 2005;146(3):15–26.

[5] Quantin C, Allaert FA, Bouzelat H, Rodrigues JM, Trombert-Paviot B, Brunet-Lecomte P, et al. La sécurité des réseaux d'informations médicales : application aux études épidémiologiques. Rev Epidemiol Sante Publique 2000;48:89–99.

[6] Schneier B. Applied Cryptography. Paris, France: International Thomson Publishing; 1994.

[7] Bellare M., Canetti R., Krawczyk H., Message authentication using hash functions. The HMAC construction. RSA laboratories' CyrptoBytes 1996;2:1–5. Available at http://www.cs.ucsd.edu/users/mihir/papers/hmac.html/.

[8] http://en.wikipedia.org/wiki/SHA_hash_functions.

[9] Vuillet-Tavernier S. Réflexion autour de l'anonymat dans le traitement des données de santé. Med Droit 2000;40:1–4.

[10] Thirion X, Sambuc R, San Marco JL. Epidemiology and anonymity: a new method. Rev Epidemiol Sante Publique 1988;36:36–42.

[11] Trouessin G, Allaert FA. FOIN: a nominative information occultation function. Stud Health Technol Inform 1997;43:196–200.

[12] Zhou JH, Zhu GL. Research and realization for certification of EHR based on ECC and SHA-1. Zhongguo Yi Liao Qi Xie Za Zhi 2008;32:117–9.

[13] Buyl R, Nyssen M. An electronic registry for physiotherapists in Belgium. Stud Health Technol Inform 2008;136:383–8.

[14] Borst F, Allaert FA, Quantin C. The Swiss solution for anonymously chaining patient files. In: Proc MEDINFO 2001, IMIA. 2001. p. 1239–41.

[15] Gensbitel MH, Riandey B, Quantin C. Appariements sécurisés : statisticiens ayez de l'audace ! Courrier Stat 2007;121–122:49–58.

[16] Goldberg M, Quantin C, Zins M. Base de données médico-administratives et épidémiologie – intérêt et limites. Courrier Stat, in press.

[17] Quantin C, Allaert FA, Gouyon B, Cohen O. Proposal for the creation of a European healthcare identifier. Stud Health Technol Inform 2005;116:949–54.

[18] Quantin C, Cohen O, Riandey B, Allaert FA. Unique patient concept: a key choice for European epidemiology. Int J Med Inform 2007;76:419–26.

[19] Quantin C, Allaert FA, Fassa M, Riandey B, Fieschi M, Cohen O. How to manage a secure direct access of European patients to their computerised medical record and personal medical record? Technol Inform 2007;127:246–55.

[20] Ethridge Y. PKI (public key infrastructure) how and why it works. Health Manag Technol 2001;22:20–1.