

**CHAÎNAGE DE BASES DE DONNÉES ANONYMISÉES
POUR LES ÉTUDES ÉPIDÉMIOLOGIQUES
MULTICENTRIQUES NATIONALES ET INTERNATIONALES :
PROPOSITION D'UN ALGORITHME CRYPTOGRAPHIQUE**

Quantin C1,2, Fassa M2, Coatrieux G3, Riandey B4, Trouessin G5, Allaert F.A6.

1. INSERM U 866 Université de Bourgogne, Dijon

2. Service de Biostatistique et Informatique Médicale, CHU de Dijon

3. Inserm U650, LaTIM ; INSTITUT TELECOM, TELECOM BRETAGNE

4. Institut National d'Etudes Démographiques (INED), Paris;

5. OPPIDA Sud, Toulouse

*6. Department of Epidemiology and Biostatistics, Mc Gill University, Montreal Canada
& Chairman IMIA WG IV "Data security".*

Adresse de correspondance

Professeur Catherine QUANTIN
Service de Biostatistique et Informatique Médicale
Centre Hospitalier Universitaire
BP 77908
21079 DIJON CEDEX
Tel: 33 3 80 29 36 29
Fax: 33 3 80 29 39 73
Email: catherine.quantin@chu-dijon.fr

Abstract:

Objectives: Compiling individual records which come from different sources remains very important for multi-centre epidemiological studies, but at the same time European directives or other national legislations concerning nominal data processing have to 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, such an anonymization procedure is in contradiction with the common requirement in public health and biomedical research as 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 by using 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 relies on the way the combination of hashing and enciphering techniques is performed: like 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 provides a great improvement in 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 security requirements.

Keywords: Security, Patient identification, Encryption, Hashing, Linkage, Multicentre studies, Anonymized data.

Résumé :

Objectif : 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 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 SHA (Secure Hash Algorithm), 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. Ceci, en respectant les exigences de sécurité imposées par les lois nationales et européennes. Cette méthode pourrait utilement être employée par la statistique publique.

Mots clés : Sécurité, Identification du patient, Chiffrement, Hachage, Chaînage de données, Etudes multicentriques, Chaînage, Données anonymisées.

Introduction

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. Pour respecter cette législation, des procédures d'anonymisation doivent être utilisées. La solution que nous avons proposée [1] repose sur l'utilisation d'une méthode cryptographique irréversible, qui est appliquée à chaque fichier avant chaînage.

La méthodologie du hachage permet de résoudre le problème de l'anonymisation des données, notamment en santé [1-3]. En effet, le hachage permet une transformation irréversible de l'identité et donc de protéger la vie privée des patients. Toutefois, les fonctions de hachage étant du domaine public, l'attaque par dictionnaire constitue un problème de sécurité majeur. En effet, ceci est particulièrement gênant lorsque les données sont collectées par plusieurs sources et rassemblées au niveau national. C'est le cas notamment lorsque les données doivent être collectées par plusieurs établissements et chaînées entre elles. Pour le Programme de Médicalisation des Systèmes d'Information (dit « le PSMI »), par exemple, il est important que tous les établissements de santé français puissent utiliser la même clé de hachage, de façon à pouvoir ensuite, au niveau national, relier entre elles toutes les données utilisables concernant un même patient. Toutefois, si la même clé de hachage est utilisée par des établissements différents, il est alors possible à l'un d'entre eux, grâce à une attaque par dictionnaire [4,5], d'appliquer l'algorithme de hachage avec cette même clé sur un dictionnaire (un grand nombre d'identités) et de retrouver ainsi l'identité des données stockées au niveau national, si celles-ci ne sont pas sécurisées. En effet, en l'absence de cette sécurisation, un établissement participant à la collecte, pourrait, s'il arrivait à accéder aux données nationales, connaître la clientèle des établissements concurrents. Afin de sécuriser les données stockées au niveau national, la CNIL a donc recommandé l'utilisation d'une fonction d'anonymisation de second niveau, soit la même fonction de hachage mais avec une autre clé, qui n'est connue que par le centre national.

Par ailleurs, pour des raisons de sécurité, il est recommandé d'utiliser des clés différentes (tout particulièrement pour le second hachage) pour chaque étude. Malheureusement, cette recommandation est en contradiction avec les besoins de chaînage, dans la mesure où l'utilisation d'une clé différente rend le chaînage ultérieur impossible entre des fichiers provenant de différentes sources, puisque l'identifiant d'un même patient va alors être complètement différent d'une source à l'autre.

L'objectif de cet article publié dans la RESP [6] et de la communication dans le cadre des Journées de Méthodologie Statistique est de proposer une méthodologie innovante pour répondre à la fois aux exigences en matière de sécurité des informations médicales ou sociales, tout en permettant le chaînage des données relatives à un même patient et leur exploitation statistique.

Il s'agit de pouvoir, en toute sécurité et parfaite légalité, rendre techniquement possible des appariements sécurisés, alors que ceux-ci n'avaient initialement pas été prévus, et de permettre ainsi le chaînage des données entre bases de données, très utile pour les études multi-centriques, dans les conditions de sécurité requises par la loi et par la CNIL. Cette technique permettrait à la statistique publique de dépasser les difficultés inhérentes à l'usage du NIR dans d'autres domaines que celui du PMSI et de l'assurance maladie. Comme l'évoque la communication précédente, cette technique rend envisageable des avancées considérables dans le domaine des statistiques de l'emploi grâce à un appariement sécurisé entre enquêtes et fichiers administratifs. Le décalage entre ces deux types de sources en serait éclairé au niveau individuel et les estimations locales bénéficieraient des techniques statistiques d'estimation sur petits domaines, par exemple pour les zones d'emploi. Ce serait donc l'opportunité de marier les avantages respectifs des fichiers administratifs exhaustifs et des données d'enquêtes par sondage individuellement plus précises

1 – Méthodologie

Dans cet article, nous ne traitons pas la question du choix de l'identifiant des patients, qui va bien sûr dépendre des contraintes des applications ou des études concernées. En particulier, ce choix devra s'adapter à la législation, aux normes ou aux usages en vigueur dans le pays et relatif aux structures concernées (établissements de santé, réseaux régionaux, organismes nationaux).

1-1- Rappel sur la fonction cryptographique de hachage

L'utilisation des fonctions de hachage est récente dans le monde de la cryptologie moderne [7,8]. Elles ont été plus particulièrement développées de façon à permettre l'élaboration de techniques de signature numérique sécurisée. Les fonctions de hachage sont dites à sens unique car le calcul de leur inverse est considéré comme irréalisable dans des délais « raisonnables », pour des raisons liées à la théorie de l'information de Shannon, et ce même avec la technologie actuelle. La fonction de hachage transforme un texte en clair d'une longueur quelconque en une valeur de hachage de longueur fixe, souvent appelée empreinte (par exemple 160 bits en sortie de la fonction SHA-1).

Parmi les nombreuses fonctions de hachage proposées par les cryptologues, la fonction considérée comme la plus sûre est le Secure Hash Algorithm (SHA) [9,10] reconnu comme standard américain par le National Institute for Standard and Technology (NIST). La fonction de hachage SHA-1, qui fournit une signature de 160 bits, est intégrée dans l'algorithme de signature DSA (Digital Signature Algorithm), qui a été proposé par le NIST en 1991. SHA-1 a montré des faiblesses depuis et a été améliorée à travers la nouvelle série d'algorithmes SHA-2. Depuis 2006, le NIST a recommandé de remplacer SHA-1 par une fonction de hachage de la série SHA-2 (notamment SHA-256 bits).

La probabilité que deux identifiants différents aient la même empreinte après hachage (taux de collision) est de l'ordre de 10^{-48} pour la fonction SHA-1 et est encore plus faible pour la fonction SHA-2 puisque la longueur du résultat du hachage par SHA-2 est encore plus grande.

Grâce à ces propriétés, cette fonction de hachage est habituellement utilisée pour vérifier l'intégrité des données. En effet, l'empreinte obtenue après hachage est spécifique du message initial. En particulier, une légère modification du message conduit à une empreinte radicalement différente (principe dit de « l'effet avalanche »). Pour assurer la sécurisation des données pendant leur transmission, une méthode de chiffrement peut être utilisée. Le chiffrement correspond à la transformation, à l'aide d'une clé de chiffrement, d'un message exprimé en clair (dit texte clair) en un message exprimé de façon incompréhensible (dit texte chiffré) si on ne dispose pas de la clé de déchiffrement. Les algorithmes de chiffrement symétrique se fondent sur une même clé pour chiffrer et déchiffrer un message. Le problème de cette technique est que la clé, qui doit rester totalement confidentielle, doit être transmise au correspondant de façon sûre. Pour résoudre le problème de l'échange de clés, le chiffrement asymétrique a été mis au point dans les années 1970. Cette méthode se base sur le principe de deux clés : une publique, permettant le chiffrement ; une privée, permettant le déchiffrement. Comme son nom l'indique, la clé publique est mise à la disposition de quiconque désire chiffrer un message. Ce dernier ne pourra être déchiffré qu'avec la clé privée, qui doit quant à elle rester confidentielle. L'expéditeur envoie à la fois le message en clair et l'empreinte signée (une signature correspond à un chiffrement asymétrique effectué par l'expéditeur avec sa clé privée). Pour s'assurer de l'origine et de l'intégrité du message, le destinataire va tout d'abord recalculer l'empreinte du message avec le même algorithme de hachage que celui utilisé par l'expéditeur, puis il comparera l'empreinte ainsi obtenue avec l'empreinte qu'il aura préalablement extraite lors de la vérification de la signature (une vérification de signature correspond au déchiffrement asymétrique effectué par le destinataire avec la clé publique de l'expéditeur). Le destinataire peut ainsi s'assurer que l'expéditeur est bien le signataire du message reçu, puisque ce dernier est le seul à connaître sa clé privée, utilisée pour signer (par chiffrement asymétrique avec la clé privée) l'empreinte, et que la clé publique correspondante est la seule à permettre le déchiffrement (par vérification de signature par déchiffrement asymétrique avec la clé publique). Une autorité de gestion des clés (appelée aussi infrastructure à clé publique) génère ou reçoit une clé publique et la certifie : elle génère un certificat contenant la clé publique et signe le tout avec sa clé privée de signature, afin d'assurer l'authenticité et l'intégrité de cette clé publique.

1-2- Utilisation des fonctions de hachage pour l'anonymisation et le chaînage des données du patient

Nous avons proposé l'utilisation des techniques de hachage pour assurer l'anonymat des informations à caractère personnel, dès 1995, pour résoudre le problème du chaînage d'informations médicales nominatives pour la mise en œuvre d'études épidémiologiques multicentriques. En effet, lors du regroupement d'informations médicales au sein d'une structure extérieure aux soins la CNIL [10] préconise une transformation irréversible des données.

Après avoir tenté d'améliorer les méthodes existantes telle que la méthode proposée par Thirion X. et coll [11], nous avons proposé à la CNIL en 1995 d'utiliser les méthodes de hachage, à sens unique, pour assurer cet anonymat. Peu après, le CESSI/CNAMTS1 « a conçu et fourni dès 1996 une fonction d'anonymisation appelé FOIN (Fonction d'Occultation d'Information Nominatives) [12] pour la mise en place du PMSI établissements privés, sur recommandation de la CNIL (qui a suggéré l'utilisation de l'algorithme développé par le DIM du CHU de Dijon). En effet, contrairement aux méthodes de chiffrement qui doivent pouvoir être réversibles de façon à ce que le destinataire légitime puisse déchiffrer le message, les méthodes de hachage sont irréversibles. Le résultat du hachage est un code strictement anonyme (ne permettant pas de revenir à l'identité du patient) mais toujours le même pour un individu donné de façon à pouvoir rapprocher les données d'un même patient. En accord avec le Service Central de la Sécurité des Systèmes d'Information (SCSSI), nous avons choisi l'algorithme SHA qui, à notre connaissance, est l'algorithme de hachage du domaine public le plus sûr vis-à-vis des tentatives de déchiffrement [1,7-10, 13,14].

L'évolution de la cryptographie nous a conduits à utiliser les algorithmes de la famille SHA-2 plus pertinents que la version SHA-1 obsolète. La procédure a été déclarée auprès de la CNIL et du SCSSI en mars 1996. Cette solution permet de résoudre le problème du chaînage des informations d'un même patient à l'intérieur d'une étude multi-centrique. Toutefois, l'algorithme de hachage étant public, la sécurité des données dépend de la clé utilisée. En effet, comme nous l'avons expliqué dans l'introduction, une personne connaissant la clé pourrait appliquer le hachage à un grand nombre d'identités et procéder ainsi à une attaque par dictionnaire : elle pourrait confronter les codes obtenus aux codes d'un individu donné du fichier haché et retrouver ainsi son identité. Pour éviter cette attaque par dictionnaire, il est recommandé d'utiliser des clés de hachage différentes pour chaque étude. Le même algorithme de hachage (par exemple SHA) est alors utilisé avec une clé qui peut être différente d'une étude à l'autre. Ainsi, avec le même identifiant et le même algorithme on obtiendra des empreintes différentes selon la clé que l'on utilise. Les résultats du hachage de la même identité, à partir de deux clés différentes, étant complètement disjoints, il n'est alors pas possible de chaîner les informations d'un même patient provenant de différentes études. Ainsi, par exemple, en Belgique, il est prévu, à partir du hachage du numéro de sécurité sociale, de constituer à l'aide de clés différentes, des identifiants distincts pour le remboursement des soins, la prise en charge médicale et un identifiant pour chaque type de recherche. Il devient alors très compliqué de relier les données produites par les différentes sources. L'objectif de cet article est de proposer une méthodologie innovante permettant de pallier ces difficultés. Celle-ci repose sur l'utilisation combinée de techniques de hachage et de chiffrement, ainsi que nous l'avons déjà proposé en collaboration avec nos collègues Suisses [15] : les données d'identité (nom, date de naissance et sexe) du patient sont tout d'abord rendues anonymes par hachage puis sécurisées par une méthode de chiffrement. L'originalité de l'approche proposée dans cet article repose sur la méthodologie de protection de la clé de chiffrement.

1-3- Proposition d'une méthode de chiffrement combinée au hachage

Dans la figure 1, on voit que l'on part du numéro d'identité du patient (NIP) rendu anonyme par double hachage DH(NIP), tel qu'il a été obtenu à l'étape précédente. Afin de sécuriser ce numéro d'identité et éviter une attaque par dictionnaire sur ce numéro, on va mettre en place une procédure comportant un double chiffrement. Ce système repose donc sur deux clés, une clé unique Pw définie pour l'étude et une clé variable lk qui va dépendre de l'identité du patient, ce qui apporte une sécurité

¹ Centre d'études des sécurités du système d'information (CESSI) de la caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS).

supplémentaire. Seule la connaissance de ces deux clés permet de chiffrer le numéro d'identité rendu anonyme et, réciproquement, de le déchiffrer. Ceci signifie que seules les personnes dûment autorisées peuvent accéder au numéro d'identité rendu anonyme, qui constitue le repère de chaînage. On peut, en effet, hacher une nouvelle fois, par la fonction de hachage H, le numéro d'identité rendu anonyme DH(NIP). La valeur Ik obtenue constitue la nouvelle clé de chiffrement. On va ensuite appliquer la fonction de chiffrement C, avec la clé Ik, au numéro d'identité rendu anonyme DH(NIP) pour le sécuriser et obtenir ainsi Clk [DH(NIP)]. Il est également possible d'appliquer la même fonction de chiffrement C sur Ik lui-même, afin de le sécuriser, mais cette fois-ci avec la clé de chiffrement de l'étude, c'est-à-dire Pw. Ceci rend beaucoup plus difficiles les attaques potentielles à mener sur le numéro stocké, puisque la clé de chiffrement Ik à découvrir va dépendre de l'identité du patient. Par ailleurs, la transmission de cette clé est également protégée par l'utilisation de la fonction de chiffrement C sur cette même clé.

Néanmoins, il reste possible de relier les données d'un même patient, dans des conditions sécurisées par retour au numéro d'identité anonymisé DH(NIP). Si l'on connaît, en effet, la clé Pw et la fonction de chiffrement C, il suffit (cf figure 2) de déchiffrer la clé Ik ; puis, grâce à cette clé, on récupère le numéro d'identité anonymisé DH(NIP). Bien évidemment, il ne s'agit pas d'obtenir l'identité en clair du patient mais d'accéder au numéro d'identité anonyme DH(NIP), qui avait été protégé, grâce au chiffrement, de toute attaque par dictionnaire.

Ceci étant dit, ce procédé est extrêmement puissant : il peut permettre, en toute sécurité et légalité, des appariements entre données protégées issues de plusieurs centres. C'est supposer l'existence - légale et approuvée par la CNIL- d'une Autorité de gestion des clés, qui détiendrait l'ensemble des clés de chiffrement Pwi utilisées par les différentes études.

La connaissance de ce numéro d'identité anonyme DH(NIP) permettrait à cette autorité de faire le chaînage liant les données des différentes études, sans pour autant connaître l'identité des patients impliqués dans ces études. L'autorité de gestion des clés connaît, en effet, les clés Pwi utilisées par les différentes études. Suite à une demande effectuée conjointement par deux responsables d'études et à l'autorisation de la CNIL, elle peut retourner à l'identité anonyme DH(NIP), le dénominateur commun à l'origine de l'identifiant du patient dans les différentes études concernées. Toutefois, les identifiants d'un même patient, malgré leur origine commune, sont complètement différents d'une étude à l'autre. Comme le montre la figure 3, l'autorité de gestion des clés est capable de retrouver Ik1 (respectivement Ik2) par déchiffrement de CPw1(Ik1) (respectivement CPw2Ik2). Par conséquent elle peut déchiffrer Clk1[DH(NIP)] ainsi que Clk2[DH(NIP)], et retrouver DH(NIP), pour chacun des centres. L'autorité de gestion des clés est donc alors en mesure de chaîner les données inter-centres. Il suffit ensuite, toujours en restant anonyme, de procéder à l'enregistrement des informations ainsi nouvellement appariées (pour les réutiliser selon les objectifs épidémiologiques à atteindre) grâce à la technique combinant double-hachage et chiffrement que nous avons décrite précédemment.

2 – Discussion

2-1- Intérêt de la méthode pour le chaînage des données

L'intérêt de cette méthode est de permettre de relier les données d'un même patient, lorsque cet appariement est autorisé par la CNIL, tout en assurant la sécurité de ces données. Chaque centre va, en effet, stocker les données du patient avec un numéro d'identification qui lui est propre. Il est donc impossible à une personne extérieure à un centre, de retrouver à quel patient correspondent les données stockées par ce centre. Toutefois, dans la mesure où la construction de l'identifiant de chaque centre repose sur une méthode similaire, et en particulier sur la même clé de hachage permettant d'anonymiser l'identité du patient, il est alors possible à l'autorité de gestion des clés, mais uniquement à cette autorité, de revenir à ce numéro d'identité anonyme, par déchiffrement. Ce déchiffrement repose sur deux conditions : que les deux centres souhaitant apparier leurs données aient donné l'autorisation à cette autorité pour le faire et qu'ils aient obtenu l'accord de la CNIL. Quant à l'autorité de gestion des clés, elle ne peut revenir aux données d'identité des patients car elle ne stocke pas les données des différents centres. Les centres ne lui transmettent pour appariement que des données d'identité rendues anonymes et chiffrées, sans données médicales.

Cette organisation permet, en cas de besoin, de débloquent la spécialisation des données créée par le hachage irréversible d'un NIR2 d'une part haché pour le PMSI ou le SNIIRAM et par ailleurs haché comme identifiant de cohorte. La solution antérieurement proposée était l'Instance de Coordination des Identifiants (ICI), tiers de confiance³ généralisé qui conserverait les tables de correspondance, dans une logique peu conforme au hachage [16]. L'avancée de cet article réside dans l'abandon de ces tables de correspondance au profit de la seule conservation de ces clés (d'où le nom de cette autorité), conservation qui est beaucoup plus sécurisée (i.e., moins à risque) que celle de ces tables. L'utilité de cette instance résulte de la politique de la CNIL de sectorisation des identifiants de la statistique publique, et de la nécessité de pouvoir communiquer entre secteurs pour transmettre des données ou pour valider l'identifiant sectoriel.

La réalisation du chaînage des données nécessite des moyens importants, mobilisables sur une longue durée. Pour assurer le rôle de « centre d'appariement sécurisé » de données individuelles en provenance de bases de données locales, régionales ou nationales, il serait souhaitable de créer une agence nationale de chaînage de données, comme celles créées à la fin des années 1990 en Australie et dont l'intérêt pour la France a été largement démontré par Marcel GOLDBERG [17]. L'intérêt du chaînage des données d'un même patient, provenant de bases de données de grande ampleur, notamment dans le cadre d'études épidémiologiques ou de recherche clinique multicentriques, dépasse largement le cadre d'un pays. Il paraît illusoire d'attendre la mise en place d'un identifiant unique pour la santé à l'échelle de l'Europe (les Etats-Unis n'en disposent pas). La politique actuelle de l'Europe est plutôt de promouvoir l'interopérabilité des identifiants existants [18-20], ce qui risque de demander beaucoup de temps. Aussi la méthode proposée dans cet article doit permettre de lever dès maintenant les principaux obstacles à la réalisation des études multicentriques européennes et internationales.

2-2- Choix de la méthode de chiffrement

Le choix de la méthode de chiffrement, effectué dans cet article, mériterait d'être discuté. On pourrait, en particulier, évoquer l'intérêt d'une méthode de chiffrement à clé publique plutôt qu'à clé secrète. Mais ces dernières sont par principe beaucoup plus rapides. Toutefois, il s'agit ici de chiffrer en vue du stockage ; la rapidité du chiffrement n'est donc sans doute pas la priorité. L'utilisation d'un algorithme à clé publique, permettant d'avoir une clé différente pour le chiffrement et le déchiffrement, permettrait certainement d'améliorer la sécurité, sans pour autant trop ralentir le chiffrement. C'est bien à ce niveau que se situe toute la difficulté de la protection de la transmission des clés secrètes comme pour toute méthode cryptographique symétrique. Les processus de type asymétrique évitent ce problème. Seule la clé publique [21] correspondant à une étude donnée est transmise aux responsables des sources de données pour chiffrer -c'est-à-dire sécuriser- le numéro d'identité rendu anonyme DH(NIP) (cf. figure 1). Les données ainsi stockées au niveau des sources mais aussi au niveau du centre de traitement des données de l'étude ne seront décryptables que par les détenteurs de la clé privée, c'est à dire le responsable du centre de traitement de l'étude et l'autorité de gestion des clés. Ainsi, seule l'autorité de gestion des clés connaît les clés privées de toutes les études et est donc en mesure de retrouver le numéro d'identité rendu anonyme DH(NIP), commun à l'ensemble des études. C'est ce principe qu'il conviendrait de généraliser.

3 - Conclusion

Cette méthodologie s'inscrit dans la nécessité d'établir l'équilibre souhaitable entre deux des principaux piliers de la sécurité des informations, à savoir la protection de la confidentialité et la disponibilité des informations. A quoi servirait en effet de disposer d'énormes cimetières de données extrêmement bien protégées mais inexploitable à toutes fins de santé publique ? Même si les solutions peuvent apparaître un peu lourdes à mettre en œuvre et quelque peu consommatrices de temps, elles ne sont rien par rapport au temps qui a été nécessaire pour recueillir les milliers de données que l'on pourrait ainsi exploiter dans le respect dû aux personnes et aux bénéfiques espérés

2 Numéro d'Inscription au Répertoire de l'INSEE, le répertoire national d'identification des personnes physiques RNIPP, plus communément (et d'ailleurs abusivement) appelé numéro de sécurité sociale.

3 Pour des raisons éthiques à l'égard des patients, les études épidémiologiques peuvent exiger le maintien d'une procédure de réidentification des personnes. La conservation d'une table de correspondance nominative par un tiers de confiance n'est alors pas évitable.

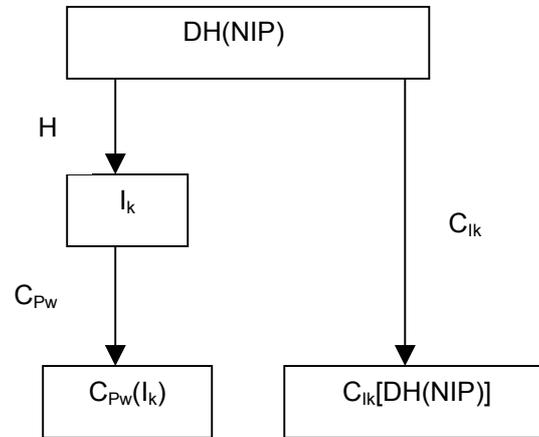
sur le plan de la santé publique, aussi bien dans le domaine de l'épidémiologie que de la recherche clinique.

Compte-tenu du nombre important de fichiers sociaux identifiés par le NIR, l'algorithme cryptographique ainsi proposé ouvre également de réelles potentialités dans le domaine de la statistique publique. Dans cette perspective, une réflexion globale avec le CNIS et la CNIL serait utile pour fixer à moyen terme le cadre de son utilisation. C'est l'une des réponses qu'on peut apporter à la sollicitation de la directrice du service juridique de la CNIL [16] lorsqu'en évoquant les appariements sécurisés, elle nous encourage par ces mots : « Statisticiens, ayez de l'audace ! ».

Référence :

- [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 des statistiques*, 2005 ;113-114 :15-26 ; *Journal de la SFdS* 2005 ;146(3) :15-26.
- [5]. Quantin C., Allaert F.A., Bouzelat H., Rodrigues J.M., Trombert-Paviot B., Brunet-Lecomte P., Gremy F., Dusserre L. La sécurité des réseaux d'informations médicales : application aux études épidémiologiques. *Revue d'Epidémiologie et de Santé Publique*. 2000;48:89-99.
- [6]. Quantin C, Fassa M, Coatrieux G, Riandey B, Trouessin G, Allaert F.A., Chaînage de bases de données anonymisées pour les études épidémiologiques multicentriques nationales et internationales : proposition d'un algorithme cryptographique. A paraître en 2009 dans la *Revue d'Epidémiologie et de santé Publique*.
- [7]. Schneier B., *Applied Cryptography*, International Thomson Publishing, 1, Rue St-Georges, 75009 Paris, France, 1994.
- [8]. Bellare M., Canetti R., Krawczyk H., Message authentication using hash functions. The HMAC construction. *RSA laboratories' CryptoBytes* 1996, 2: 1-5. Available at <http://www.cs.ucsd.edu/users/mihir/papers/hmac.html/>
- [9]. http://en.wikipedia.org/wiki/SHA_hash_functions
- [10]. Vuillet-Tavernier S. – Réflexion autour de l'anonymat dans le traitement des données de santé. *Med et Droit* 2000 ; 40 : 1-4
- [11]. Thirion X., Sambuc R., San Marco J.L. Epidemiology and anonymity: a new method. *Revue d'Epidémiologie et Santé Publique* 1988, 36, 36-42.
- [12]. Trouessin G, Allaert FA. FOIN: a nominative information occultation function. *Stud Health Technol Inform.* 1997;43 Pt A:196-200.
- [13]. Zhou JH, Zhu GL. Research and realization for certification of EHR based on ECC & SHA-1. *Zhongguo Yi Liao Qi Xie Za Zhi.* 2008 Mar;32(2):117-9.
- [14]. Buyl R, Nyssen M. An electronic registry for physiotherapists in Belgium. *Stud Health Technol Inform* 2008;136:383-8.
- [15]. Borst F., Allaert F.A., and Quantin C. The Swiss solution for anonymously chaining patient files. *Proc MEDINFO 2001, IMIA 2001* : 1239-41.
- [16]. Gensbitel M.H, Riandey B, Quantin C. Appariements sécurisés : Statisticiens ayez de l'audace ! *Le Courrier des statistiques*, 2007, 121-122 : 49-58.
- [17]. Goldberg M, Quantin C, Zins M. Base de données médico-administratives et épidémiologie – Intérêt et limites. *Courrier des Statistiques*, 2008, 124 , 59-70.
- [18]. 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.
- [19]. Quantin C, Cohen O, Riandey B, Allaert FA. Unique patient concept: a key choice for European epidemiology. *International Journal of Medical Informatics*, 2007;76:419-426.
- [20]. 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? *Technology and Informatics* 2007;127:246-255.
- [21]. Ethridge Y. PKI (public key infrastructure) how and why it works. *Health Manag Technol* 2001; 22:20-21.

Figure 1 : Sécurisation par chiffrement du numéro d'identité rendu anonyme.



H = fonction de Hachage

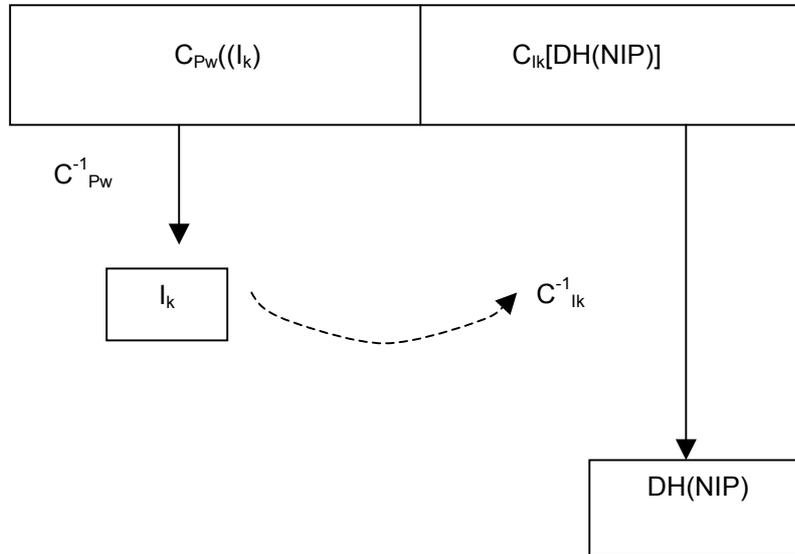
C = fonction de Chiffrement

DH : Double fonction de Hachage

I_k : clé variable dépendant de l'identité du patient

P_w : clé unique définie pour l'étude

Figure 2 : Déchiffrement du numéro d'identité anonyme.



H = fonction de Hachage

C = fonction de Chiffrement

C^{-1} = fonction de Déchiffrement (associée à C)

DH : Double fonction de Hachage

I_k : clé variable dépendant de l'identité du patient

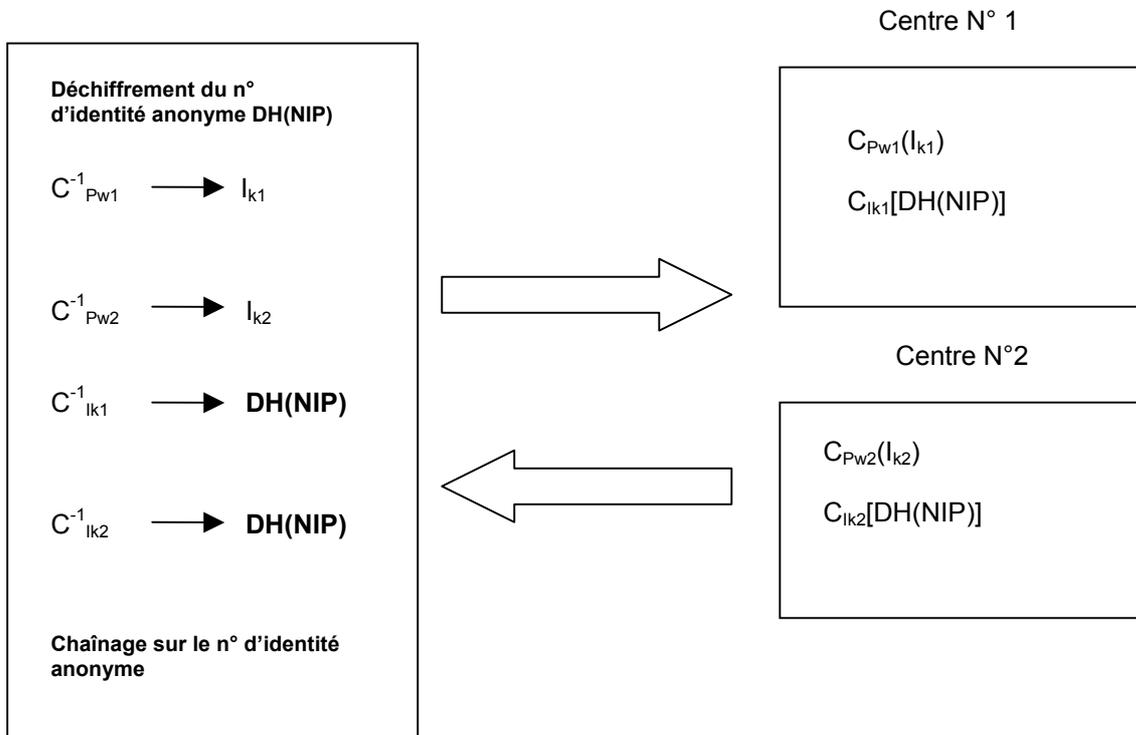
P_w : clé unique définie pour l'étude



: la clé I_k est utilisée pour la fonction C^{-1}

Figure 3 : Chaînage inter-centres à partir du numéro d'identité anonyme, par l'Autorité de Gestion des clés.

Autorité de gestion des clés



H = fonction de Hachage

C = fonction de Chiffrement

DH : Double fonction de Hachage

I_k : clé variable dépendant de l'identité du patient

P_w : clé unique définie pour l'étude