JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

CATHERINE QUANTIN
BÉATRICE GOUYON
FRANÇOIS-ANDRÉ ALLAERT
OLIVIER COHEN

Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : des informations médicales

Journal de la société française de statistique, tome 146, n° 3 (2005), p. 19-37

<http://www.numdam.org/item?id=JSFS_2005__146_3_19_0>

MÉTHODOLOGIE POUR LE CHAÎNAGE DE DONNÉES SENSIBLES TOUT EN RESPECTANT L'ANONYMAT : DES INFORMATIONS MÉDICALES

Catherine QUANTIN *, Béatrice GOUYON*,
François-André ALLAERT*,**, Olivier COHEN ***

RÉSUMÉ

Devant la multiplication des recueils informatisés d'informations médicales structurées, les statisticiens et les épidémiologistes souhaiteraient de plus en plus bénéficier, pour leurs études, du regroupement de fichiers nominatifs provenant de plusieurs sources. Toutefois, le croisement de fichiers, pour le regroupement d'informations relatives à un même patient, requiert le respect des législations française et européenne relatives au traitement des données à caractère personnel. Après avoir fait le point sur ces législations, nous décrivons la procédure d'anonymat et de chaînage des informations rendues anonymes développée par le DIM [1] du CHU [2] de Dijon. Chaque fichier est rendu anonyme à la source, à l'aide du logiciel Anonymat, qui utilise la méthode du hachage irréversible de l'identité. Le chaînage des informations d'un même patient repose techniquement sur un modèle statistique par mélange de distributions. Nous présentons ensuite des exemples d'application du hachage de l'identité permettant la constitution de bases de données régionales ou nationales. Ces exemples montrent que l'utilisation du logiciel Anonymat permet de relever le défi de la prise en compte des exigences de la CNIL [3] sans nuire à la disponibilité des données médicales.

Mots clés : chaînage de données, cryptage irréversible, hachage, sécurité.

ABSTRACT

Record linkage allows the compiling of same-person records from various source files, and can thus improve the feasibility of epidemiological research such as population-based studies. Compliance with European legislation on data privacy and data security is, however, essential. Our article describes one way to achieve this : a

* Service de Biostatistique et Informatique Médicale
(Pr. QUANTIN : catherine.quantin@)chu-dijon.fr), CHU de Dijon, BP 1542, 21034 Dijon cedex.
** Chairman TC/251/WGIII Centre Européen de Normalisation, CEN BIOTECH, BP 53077, 21030 Dijon cedex.
*** Laboratoire TIMC - IMAG UMR 5525, CNRS, Université Joseph Fourier, Grenoble.

1. Département d'information médicale.
2. Centre hospitalo-universitaire.
3. Commission nationale de l'informatique et des libertés.

computerized record hash coding and linkage procedure to chain medical information for the purpose of epidemiological monitoring. Before their extraction, files are rendered anonymous using a one-way hash coding based on the standard hash algorithm (SHA) function. Once the patient information is anonymized using ANONYMAT software, it can be linked by means of a mixture model that takes several identification variables into account. Applications of this anonymous record linkage procedure were carried out at the national and regional levels. The applications illustrate how the use of the ANONYMAT program makes it possible to respect data-confidentiality legislation without impeding data availability.

Keywords : hash-coding, non-reversible encryption, record linkage, security.

En dehors des utilisations imposées par l'assurance maladie ou les services de l'État (feuille de soins électronique, Programme de Médicalisation des Systèmes d'Information, PMSI) il est possible d'envisager des utilisations et des circulations d'information propres aux médecins par exemple dans le cadre de réseaux de soins. Toutefois, le regroupement des informations médicales relatives à un même patient par le croisement de divers fichiers existants, ne peut aller à l'encontre des législations européenne et française relatives à la protection des libertés individuelles vis-à-vis du traitement automatisé des données personnelles. Nous montrerons que le respect de la législation conduit au paradoxe suivant : il est possible de réunir les différentes parties du dossier d'un même patient sans pour autant accéder à son identité. Nous verrons comment les techniques cryptographiques, telle que la procédure d'anonymat et de chaînage développée par le Département d'information médicale (DIM) du Centre hospitalo-universitaire (CHU) de Dijon, apportent une solution à ce paradoxe.

1. La législation sur la sécurité des informations nominatives

La directive européenne du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractères personnel et à la libre circulation de ces données substitue au concept d'informations nominatives de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés celui de « *données à caractère personnel* », c'est-à-dire « *toute information concernant une personne physique identifiée ou identifiable* ». « *Est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale.* » Cette définition très exhaustive permet à terme de considérer que toute base de données est indirectement nominative [Quantin *et al.*, 1999]. Cette directive européenne vient d'être transcrite en droit français donnant lieu à la loi du 6 août 2004, relative à la protection des personnes physiques à l'égard des traitements de données à

caractère personnel et modifiant la loi n° 78-17 du 6 janvier 1978. De l'ensemble de ces considérations, il résulte que la notion de données nominatives ou personnelles concerne un très grand nombre d'informations même si le nom n'apparaît pas et qu'aucune table de correspondance entre l'identité en clair et des codes alphanumériques substitutifs n'existe. Sur le plan statistique, le risque d'identification d'un individu à partir d'informations apparemment anonymes est loin d'être nul, en raison de la possibilité de croisement avec une grande diversité de fichiers existants ou à venir. Qui aurait pensé, il y a encore peu de temps, que l'identifiant de la Sécurité sociale pourrait être traité informatiquement par les services fiscaux [Loi n° 98-1266, 1998] ? Toutefois, ce souci que les personnes concernées ne puissent être identifiées lors de l'évaluation ou lors de l'analyse des pratiques et des activités de soins et de prévention, est repris intégralement dans l'article 41 de la loi n°99-641 du 27 juillet 1999 portant création d'une couverture maladie universelle modifiant l'article 40-10 de la loi du 6 janvier 1978.

2. Anonymat : l'essor des méthodes cryptographiques

Plutôt que d'utiliser des méthodes statistiques d'anonymat reposant sur la perturbation de données [Sweeney, 1998; Willenborg *et al.*, 1995; Quantin *et al.*, 2000a] pour empêcher l'identification des personnes, et par conséquent induisant une perte de qualité des informations, il paraît préférable d'utiliser des techniques cryptographiques pour assurer la sécurité des informations. Ces techniques, qui permettent de protéger des informations grâce à un code secret, sont généralement issues de problèmes mathématiques très difficiles à résoudre si l'on ne dispose pas de ce code. Ces méthodes ne sont pas beaucoup plus récentes que les méthodes statistiques d'anonymat mais leur utilisation était jusqu'à présent restreinte par la loi pour des raisons de défense nationale. Les autorisations d'utilisation de ces méthodes n'étaient donc pas faciles à obtenir auprès du Service Central de la Sécurité des Systèmes d'Informations (SCSSI), service qui dépend directement du Premier Ministre. Le domaine de la cryptographie a bénéficié assez récemment d'une libéralisation, d'abord en 1998 [décret n° 98-101 1998; décret n° 98-206 1998; décret n° 98-207 1998] sous forme d'une simplification de la procédure de déclaration auprès du SCSSI. Cet assouplissement en faveur de l'utilisateur, *a priori* non spécialiste du domaine, a été poursuivi en 1999 [décret n° 99-199, 1999] en faisant porter le poids de la réglementation sur les professionnels de la cryptologie. En particulier, l'utilisation des clés de haute sécurité d'une longueur de 128 bits a été rendue possible (jusqu'alors limitée à 40 bits), cette évolution étant devenue obligatoire pour satisfaire aux exigences de la reconnaissance de la signature électronique et faciliter les transactions commerciales dans le cadre de l'Internet. Cette libéralisation a permis de lever l'obstacle de l'utilisation des techniques de cryptage pour assurer la confidentialité des informations médicales directement ou indirectement nominatives et appelées à circuler sur des réseaux informatiques. En effet, si la Commission Nationale de l'Informatique et des Libertés (CNIL) accepte des clés de seulement 40 bits pour le cryptage des informations *indirectement* nominatives, elle exige des

clés d'une longueur d'au moins 56 bits pour celles qui sont *directement* nominatives. Si l'on s'intéresse à la sécurité des informations médicales circulant sur un réseau, les méthodes de cryptage peuvent être utilisées à trois niveaux (figure 1). Le premier concerne le respect de la confidentialité des informations pendant leur transmission. La confidentialité [Fisher et Madge, 1996], selon la définition donnée par le Centre Européen de Normalisation [4], est assurée lorsque seuls les utilisateurs dûment habilités ont accès à l'information.

2.1. Chiffrement ou cryptage, base de la confidentialité

Chiffrer un message, c'est lui appliquer une fonction de transformation qui le rendra illisible au tout-venant. Cette fonction est appliquée par un algorithme de chiffrement [Douglas, 1996 ; Beckett, 1990 ; Brassard, 1993]. Afin de personnaliser le cryptage, on utilise une clé (figure 1). Si l'on prend l'exemple d'un échange d'information entre un établissement de santé et un cabinet médical libéral, le médecin hospitalier sera sûr que seul le médecin généraliste auquel ce message est destiné pourra en prendre connaissance, puisqu'en tant que destinataire légitime il sera le seul à connaître la clé de déchiffrement.

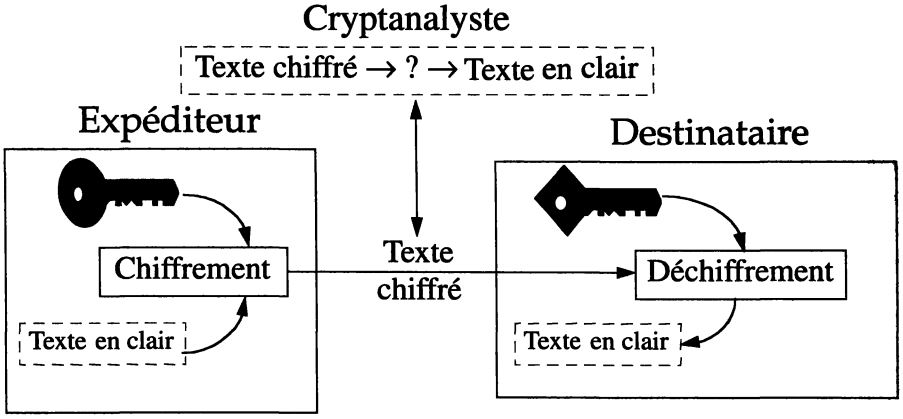

FIG 1. — Chiffrement, déchiffrement et cryptanalyse.

On supposera que l'algorithme est public et que la confidentialité n'est assurée que par la clé de l'utilisateur, qui doit donc être difficile à retrouver, même pour un cryptanalyste expérimenté. Un bon algorithme de cryptage sera un algorithme NP-complet, c'est-à-dire que le calcul inverse (correspondant au déchiffrement du message) n'est possible que par énumération exhaustive des valeurs de clé.

Un algorithme de cryptage est dit symétrique ou à clé secrète lorsqu'une seule clé sert à la fois au cryptage et au décryptage. C'est le cas par exemple de

4. Groupe de travail « Qualité et Sécurité », Working Group III « Quality and Security » de la Communauté Européenne.

l'algorithme Data Encryption Standard (DES) adopté comme standard officiel du gouvernement américain en 1977. L'utilisation de ce type d'algorithme pose le problème du partage de la clé de chiffrement entre l'expéditeur et le destinataire. Au contraire, les algorithmes asymétriques (ou encore à clé publique), qui ont été développés dès 1976, reposent sur l'utilisation de deux clés : la première est dite publique et tout le monde peut l'utiliser pour envoyer un message chiffré à un destinataire donné ; la seconde est dite privée, elle est connue uniquement de ce destinataire et elle seule peut permettre de décrypter le message. Cette procédure supprime le problème de la transmission d'une clé. En effet, seul le destinataire légitime, détenteur de la clé privée, est en mesure de déchiffrer le message. L'algorithme à clé publique le plus connu est l'algorithme RSA [Rivest *et al.*, 1978 ; Zimmermann, 1986] dont la sécurité repose sur l'hypothèse que la factorisation d'un grand nombre en produit de nombres premiers est longue et difficile.

2.2. Signature numérique et contrôle d'intégrité

Le deuxième niveau concerne l'utilisation des méthodes de signature numérique pour permettre au médecin receveur d'authentifier le médecin émetteur du message. Dans l'exemple que nous venons de prendre, ceci signifie que le médecin généraliste pourra s'assurer que le message a bien été adressé par le médecin hospitalier annoncé. La signature numérique a été reconnue comme ayant valeur légale par la loi française n° 2000-230 du 13 mars 2000 portant adaptation du droit de la preuve aux technologies de l'information et relatif à la signature électronique. Ce mécanisme regroupe deux procédures : la signature d'une unité de données et la vérification de la dite signature. La signature d'un message repose sur une clé caractéristique de l'entité émettrice. On exige que la signature ne puisse être produite que par le seul signataire et que la vérification ne puisse pas permettre de reproduire la signature. Généralement, on utilise des algorithmes à clé publique tel que le RSA. L'utilisation de la signature numérique va permettre également de garantir l'intégrité du message, c'est-à-dire être sûr que le message n'a pas été modifié pendant sa transmission. Sur la figure 2, on observe que l'expéditeur crée une empreinte de taille fixe du message qui, lui, est de taille variable, par une technique de hachage [Marsault, 1995].

L'utilisation des fonctions de hachage est récente dans le monde de la cryptologie moderne. Elles ont été surtout développées de façon à permettre l'élaboration de techniques de signature numérique sécurisée. Les fonctions de hachage sont dites à sens unique si le calcul de leur inverse est considéré comme irréalisable avec la technologie actuelle dans des délais « raisonnables ». La fonction de hachage transforme un texte en clair d'une longueur donnée en une valeur de hachage de longueur fixe [5], souvent appelée empreinte.

5. Ceci peut sembler étonnant : comment un texte initial, quelle que soit sa longueur, peut-il, une fois transformé, être de longueur fixe ? C'est dû au fait que le fruit du hachage est un texte « compressé », et que cette compression est telle que son résultat est d'un volume indépendant du texte initial.

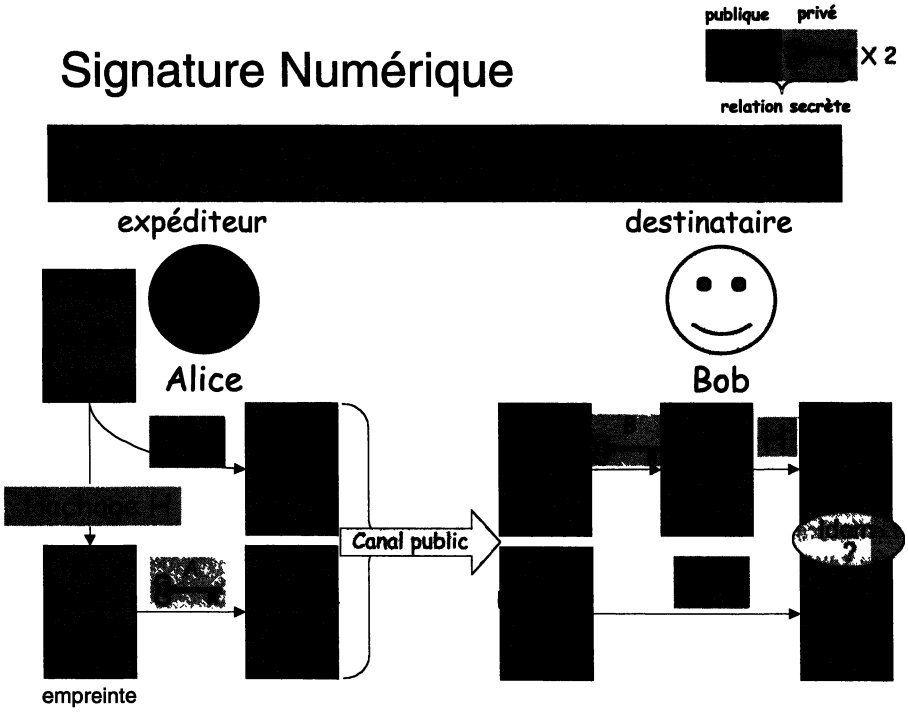

FIG 2. — Contrôle d'intégrité dans la signature numérique.

Parmi les nombreuses fonctions de hachage proposées par les cryptologues, la fonction considérée comme la plus sûre est le *Secure Hash Algorithm* (SHA) reconnu comme standard américain par le *National Institute for Standard and Technology* (NIST). Cette fonction de hachage est intégrée dans l'algorithme de signature DSA (*Digital Signature Algorithm*), qui a été proposé par le NIST en 1991. Dans un premier temps, le message à hacher est complété par une chaîne afin de rendre sa taille multiple de 512 bits. Chaque bloc de 512 bits est ensuite découpé en 16 sous-blocs de 32 bits, eux-mêmes transformés en 80 mots de 32 bits sur lesquels 80 opérations sont appliquées. Le résultat de l'algorithme SHA est une empreinte, c'est-à-dire un message de taille fixe, de 160 bits. L'empreinte est alors spécifique du message. En particulier, une légère modification du message conduit à une empreinte radicalement différente. L'expéditeur envoie à la fois le message en clair et l'empreinte chiffrés. 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 utilisé par l'expéditeur, puis comparera l'empreinte obtenue à l'empreinte qu'il aura préalablement déchiffrée. Le destinataire peut ainsi s'assurer que l'expéditeur est bien le signataire du message reçu, puisque ce dernier est le seul à connaître la clé privée, utilisée pour le chiffrement de l'empreinte, et que la clé publique correspondante est la seule à permettre le déchiffrement.

2.3. L'utilisation des techniques de hachage pour assurer l'anonymat des informations à caractère personnel

Le troisième niveau d'utilisation des techniques cryptographiques concerne le regroupement d'informations médicales au sein d'une structure extérieure aux soins. En effet, le problème du chaînage d'informations médicales nominatives pour la mise en œuvre d'études épidémiologiques multicentriques se pose de plus en plus fréquemment, par exemple dans le cadre d'études coopératives entre la médecine de ville (cabinets) et la médecine hospitalière. Selon les recommandations de la CNIL [Vuillet-Tavernier, 2000], il est alors préférable d'utiliser des techniques cryptographiques garantissant une transformation irréversible des données. Après avoir tenté d'améliorer les méthodes existantes telle que la méthode proposée par Thirion *et al.* [1988], nous avons proposé à la CNIL en 1995 d'utiliser les méthodes de hachage à sens unique pour assurer cet anonymat. En effet, contrairement aux méthodes de cryptage qui doivent pouvoir être réversibles de façon à ce que le destinataire légitime puisse déchiffrer le message, les méthodes de hachage à sens unique 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 SCSSI, nous avons choisi l'algorithme SHA qui, à notre connaissance, est l'algorithme de hachage du domaine public le plus sûr face aux tentatives de déchiffrement. La procédure a été déclarée auprès de la CNIL et du SCSSI en mars 1996. À l'époque, si la législation concernant les fonctions de cryptage était très stricte, l'utilisation des fonctions de hachage relevait du régime de la simple déclaration. En effet, dans la mesure où ces fonctions sont irréversibles, elles ne peuvent être utilisées par des organisations secrètes, souhaitant échanger des informations en échappant au contrôle du gouvernement. Toutefois, bien qu'irréversible, l'opération de hachage ne garantit pas la sécurité parfaite des informations. L'algorithme étant public, le hachage pourrait être appliqué à un grand nombre d'identités. On pourrait confronter les codes obtenus aux codes d'un individu donné du fichier haché et retrouver ainsi son identité. On parle alors d'attaque par dictionnaire. Pour prévenir ce type d'attaque, nous avons proposé d'utiliser non pas une clé unique mais une table de clés, de sorte que la modification introduite varie d'une identité à l'autre. Dans notre étude, le choix de la clé varie selon l'identité à hacher (en fonction des caractères contenus dans cette identité et de leur position). En outre, nous avons proposé un double hachage. Si, par exemple, l'on souhaite rapprocher des fichiers venant de plusieurs sources, chaque expéditeur d'un fichier utilisera une première table de clés appelée K1. Cette « clé » K1, utilisée au moment du hachage des identités pour chaque centre de recueil des informations, permet de protéger les informations vis-à-vis des personnes qui ne connaissent pas cette clé et qui sont donc extérieures à l'étude. Toutefois, l'ensemble des centres participant à l'étude devant utiliser la même clé, il convient alors d'assurer la sécurité des informations centralisées, même vis-à-vis des centres de recueil détenteurs de la clé K1. Les informations reçues par le centre de traitement assurant le croisement des fichiers sont alors à nouveau hachées, par le même algorithme de hachage, mais avec une seconde

table de clés K2. À l'issue des deux hachages de données d'identité, réalisés successivement au niveau des centres recueils et des centres de traitement, l'anonymat des fichiers est ainsi définitivement préservé.

3. Une procédure pour assurer conjointement l'anonymat et le chaînage des informations médicales

La procédure d'Anonymat et de chaînage, développée au DIM du CHU de Dijon, se déroule en deux temps. Une première étape concerne la transformation irréversible (décrite ci-dessus) des variables d'identification [Quantin *et al.*, 1998] (nom, prénom, date de naissance, sexe, ...) pour obtenir un code strictement anonyme, qui constitue le repère de chaînage. La seconde étape est celle du croisement des fichiers [Quantin *et al.*, 2004 ; Quantin *et al.*, 2005] pour chaîner les informations d'une même personne. L'objectif du chaînage est de confronter des fichiers doublement hachés provenant de sources différentes, pour associer les observations qui se rapportent à un même individu. Deux types d'erreurs [Brenner *et al.*, 1997] peuvent survenir dans le processus de chaînage. Le premier correspond au chaînage de deux observations concernant deux individus différents et constitue une erreur « d'homonymie » : par exemple si l'on associe à tort des informations concernant deux personnes dénommées respectivement Dupond et Dupont, du fait d'une erreur dans la saisie de leurs identités. Le deuxième type d'erreur correspond à l'absence de chaînage de deux observations d'un même individu et constitue l'erreur « de synonymie » : par exemple en cas d'utilisation successive du nom de jeune fille et du nom marital pour la même femme. Ces erreurs pourraient être dues soit à des erreurs dans le recueil des données d'identité, soit à la méthode de hachage elle-même. En particulier, des erreurs d'homonymie pourraient résulter de l'existence de collisions lors du hachage, c'est-à-dire de l'obtention du même code à partir du hachage de deux identités différentes. Dans le cas de l'algorithme SHA retenu pour la procédure de hachage, il s'avère que le taux de collision est très faible (de l'ordre de 10^{-48}) et que le risque d'erreur d'homonymie correspondant, égal à ce même nombre, est donc négligeable [Bouzelat, 1998, p. 97]. Pour réduire l'impact des erreurs de saisie de l'identité sur le chaînage, un traitement orthographique a été intégré dans la procédure d'anonymat. La méthode de chaînage « AUTOMATCH » proposée par Jaro [1995], très utilisée aux États-Unis [Sugarman *et al.*, 1996], a été adaptée. Elle tient compte simultanément de plusieurs variables d'identification : le nom, le prénom, le nom de jeune fille, la date de naissance, le sexe et le code postal du lieu de résidence. Bien sûr, chacune de ces variables n'identifie pas un individu de manière univoque, et l'on est ramené au problème connu de la valeur informationnelle d'un signe. Chaque variable est alors pondérée en fonction de la quantité d'information qu'elle apporte. Par exemple, on attribue une valeur plus importante à l'information fournie par la date de naissance qu'à celle fournie par le sexe (la probabilité que deux personnes aient la même date de naissance étant en effet bien plus petite que la

probabilité qu'elles aient le même sexe). Pour déterminer si deux observations doivent être chaînées, on applique un modèle d'analyse statistique qui tient compte des coefficients de pondération de chaque variable utilisée [Quantin *et al.*, 2000b]. Considérons l'ensemble des $n_A \times n_B$ paires d'enregistrement résultant du croisement systématique des fichiers $A(n_A)$ et $B(n_B)$ à chaîner. Nous pouvons définir une partition en deux ensembles M (pour matched) et U (pour unmatched) du produit cartésien $A \times B$. L'ensemble M contient toutes les paires d'enregistrements dites concordantes, c'est-à-dire dont les deux enregistrements correspondent au même individu. L'ensemble U contient toutes les paires qui restent, dites non concordantes. Ainsi, le processus de chaînage des enregistrements consiste à classer les différentes paires d'enregistrements comme appartenant à M ou à U. Si la paire d'enregistrements j est concordante pour la variable d'identification i, c'est-à-dire, par exemple, que les noms des deux enregistrements de la paire sont identiques, alors le poids pour cette variable est donné par la formule (1) :

$$w_{i,j} = \log(m_i/u_i) \tag{1}$$

où les paramètres m_i et u_i représentent respectivement la probabilité que deux enregistrements correspondant au même individu concordent sur cette variable (probabilité appelée « sensibilité ») et la probabilité que deux enregistrements correspondant à deux individus différents, soient concordants sur cette variable (probabilité dont le complément à 1 est appelé « spécificité ») de la variable i considérée. Le poids accordé à cette variable sera alors d'autant plus important que m_i est proche de 1 et u_i voisin de 0. Si, par contre, la paire j est non concordante pour la variable i, c'est-à-dire, par exemple, que les noms des deux enregistrements de la paire sont différents, alors la distribution de la « concordance », variable qualitative dichotomique (0 en cas de concordance des deux enregistrements, 1 en cas de discordance), suit une loi binomiale, de paramètre m dans l'ensemble M et de paramètre u dans l'ensemble U. L'application d'un modèle par mélange de ces deux distributions sur les données recueillies permet alors l'estimation des paramètres m et u, nécessaire au calcul des coefficients de pondération de chaque variable utilisée. La décision à prendre pour classer une paire d'enregistrements dépend de l'ensemble des variables d'identification. Ainsi, on attribue à chaque paire d'enregistrements un poids global appelé *poids composé* égal à la somme des poids correspondant aux différentes variables. Pour chaque variable, ce poids est positif en cas de concordance des deux enregistrements et est négatif en cas de discordance selon la formule (2) :

$$w_{i,j} = \log((1 - m_i)/(1 - u_i)) \tag{2}$$

Après le calcul de la fonction de répartition de ces poids, pour l'ensemble M comme pour l'ensemble U, une paire d'enregistrements est classée [Jaro, 1995] :

- à chaîner si son poids composé dépasse le seuil n° 2 (valeur du poids pour lequel la fonction de répartition conditionnellement à M est égale à 2,5 % [6]);
- à ne pas chaîner si le poids composé est inférieur au seuil n° 1 (valeur du poids pour lequel la fonction de répartition conditionnellement à U est égale à 97,5 %, complément à 1 de 2,5 %);
- en situation d'indécision, si le poids composé est situé entre les valeurs des deux seuils.

Cette situation suppose une validation manuelle des données de chaînage discordantes (*cf.* application au réseau périnatal). En pratique, cette validation peut être réalisée même sur les données rendues anonymes car chaque centre source de l'information est en droit de conserver la correspondance entre le numéro d'anonymat et l'identité du patient. Le coordonnateur de l'étude peut donc demander la vérification ou la correction des données correspondant à un numéro d'anonymat particulier. Le centre source de l'information renvoie alors l'ensemble des enregistrements corrigés, après un nouvel anonymat.

4. Applications dans le domaine médical

4.1. Constitution de bases de données régionales ou inter-régionales

Devant les perspectives offertes par les techniques de hachage, de nombreux acteurs de la recherche médicale se sont employés à la création de bases de données sur des thématiques spécifiques. Seules sont citées ici celles développées en partenariat avec le DIM du CHU de Dijon. Elles concernent le département de la Loire (étude de l'inter-file active des personnes cancéreuses), la région Bourgogne (Réseau Périnatal), les régions Bourgogne et Franche-Comté (réseau hépatite C, suivi des tentatives de suicide, réseau ESPOIR pour l'insuffisance rénale chronique) ou plusieurs régions simultanément (plate-forme HC Forum). Les principales applications sont détaillées ci-dessous.

4.1.1 Étude de l'inter-file active des personnes cancéreuses de trois structures hospitalières dans le cadre de la planification régionale en Rhône-Alpes

Suite à l'approbation du premier schéma régional d'organisation sanitaire (SROS) en 1994 [Abrial, 1998], les principaux établissements hospitaliers du secteur sanitaire n° 6 de la région Rhône-Alpes, le Centre hospitalier régional et universitaire de Saint-tienne (CHRUSE) et l'Union départementale de la mutualité de la Loire (UDML) ont constitué un syndicat inter hospitalier nommé Institut de cancérologie de la Loire (ICL) pour assurer la coordination des soins d'oncologie dans ce secteur sanitaire. Par courrier en date du

6. La valeur 2,5 % retenue ici correspond à l'intervalle de confiance usuel, mais une autre valeur est possible si l'on veut une précision plus petite ou plus grande.

06/11/97 [Agence régionale de l'hospitalisation (ARH), 1997] le directeur de l'ARH demandait à ces établissements d'« accompagner la mise en place de l'Institut de cancérologie de la Loire » par la recherche de « l'élaboration de la file active de cancérologie » de chaque établissement et par l'étude de l'« inter-file active » entre ces établissements. Il était alors convenu de s'appuyer sur le PMSI (Programme de Médicalisation des Systèmes d'Information – voir ci-dessous) qui, s'il ne constitue pas un dossier complet de cancérologie, permet de déterminer le nombre de personnes cancéreuses traitées par chaque établissement. Mais en raison des contraintes d'anonymisation des données du PMSI imposées par la CNIL, se posait alors le problème de dénombrer les malades qui bénéficiaient d'une prise en charge partagée par ces établissements. Depuis 1998, le Service de santé publique et de l'information médicale (Pr. Rodrigues) du CHRUSE applique le logiciel Anonymat aux noms, prénoms, dates de naissance de chacun des enregistrements des trois bases de données issues du PMSI (à partir des données 1996) pour les trois établissements concernés, de façon à les rendre anonymes tout en pouvant les chaîner pour repérer les patients communs aux différents établissements [Quantin *et al.*, 2000b]. De plus, la confrontation entre le nombre de numéros d'Anonymat obtenus dans chaque base et le nombre de patients calculé par l'administration a permis d'estimer le taux de doublons dans chacune des bases administratives.

4.1.2 Développement d'un recueil régional d'indicateurs en périnatalité en région Bourgogne

Un réseau périnatal s'est progressivement développé en Bourgogne depuis 1992 [Gouyon *et al.*, 1999]. Ce réseau inclut les 18 établissements publics et privés prenant en charge les femmes enceintes et les nouveau-nés dans la région. Un recueil régional continu de 42 indicateurs a été mis en place en 1998 sur la base du volontariat pour toutes les naissances prises en charge dans les établissements de la région Bourgogne (environ 18 000 naissances annuelles). Les informations sont extraites du PMSI, sous forme de résumés d'unité médicale (RUM). Les indicateurs n'existant pas dans le PMSI, tels que l'âge gestationnel ou les facteurs de risques psychosociaux, font l'objet d'un recueil supplémentaire sur une fiche adjointe au RUM, constituant un « RUM élargi ». Pour le traitement des données médicales, le chaînage des « RUM élargis » est impératif et ceci à deux niveaux différents. D'une part, les « RUM élargis » d'une même personne, mère ou nouveau-né, doivent pouvoir être reliés lorsqu'il y a hospitalisations successives dans plusieurs unités, y compris lorsqu'il s'agit d'établissements différents. D'autre part, les « RUM élargis » de la mère doivent être reliés à ceux de ses enfants, même si ceux-ci sont hospitalisés dans un établissement différent, afin d'évaluer l'impact postnatal des facteurs de risques et des pathologies de la grossesse. Toutefois, conformément à la législation, les fichiers ne sont transmis au DIM du CHU de Dijon pour exploitation qu'après avoir été rendus anonymes. La procédure de hachage est appliquée avec le système de double clé (*cf.* p.7) permettant d'assurer un anonymat strict c'est-à-dire de rompre définitivement tout lien entre la personne et les données la concernant. Le chaînage de données anonymes a alors été rendu possible par l'utilisation du logiciel Anonymat

(autorisée par la CNIL en 1998) à partir de six variables : le nom de jeune fille de la mère, son prénom et sa date de naissance, le prénom de l'enfant et sa date de naissance, le code postal de résidence de la mère [Cornet *et al*, 2001]. Ces six informations sont saisies de manière identique dans les « RUM élargis » de la mère et de son bébé. Dans le cas de grossesse multiple, tous les prénoms des nouveau-nés doivent être saisis dans le « RUM élargi » de la maman. Ces six variables nominatives sont utilisées par le programme de chaînage, après avoir été rendues anonymes. À ce jour, les 18 établissements effectuent le recueil d'indicateurs en routine, représentant l'ensemble des naissances de Bourgogne. Avant transmission, les fichiers sont validés au sein de chaque établissement par comparaison avec les cahiers de services (maternités et services de pédiatrie). De plus, l'exhaustivité et la qualité du recueil des données de chaînage sont systématiquement contrôlées dans chaque établissement et de façon centralisée par l'équipe coordinatrice (Dr. GOUYON) qui traite les données au DIM du CHU (tests de chaînage mère-enfant, identification des fiches non chaînées, correction des erreurs). Un chaînage mère-enfant est obtenu pour 86,3 % des nouveau-nés, avant validation, et pour 99,9 % des nouveau-nés après l'ensemble des procédures informatisées et manuelles de correction des erreurs.

4.1.3 Le suivi des maladies génétiques : la plate-forme HC Forum

Le Professeur Cohen a mis en place une application accessible sur internet par des médecins généticiens et des chercheurs autorisés, destinée à regrouper des informations médicales d'un même patient. Dans cette application, le médecin propose au patient suivi pour une maladie génétique de participer à ce projet qui permettra, par le recours à un dispositif de chaînage familial comportant des identifiants individuels anonymisés, de mieux comprendre l'évolution de sa maladie et de celle de sa famille. À cet effet, il lui est demandé de communiquer ses nom, prénom et date de naissance ainsi que ceux de son père et de sa mère, après s'être assuré de leur accord, afin de pouvoir définir automatiquement un numéro identifiant à composante familiale, créé conjointement avec le DIM du CHU de Dijon pour cette application et rendu anonyme par le logiciel Anonymat. Cette anonymisation a lieu en local, avant transmission, afin que seules des données déjà anonymisées transitent vers HC Forum®, la plate-forme centrale [Cohen *et al.*, 2001]. Avec l'accord du patient, cet identifiant sera ensuite transmis par un système sécurisé à la plate-forme HC Forum avec des données médicales le concernant. Il fera alors l'objet d'une seconde anonymisation qui rendra la base totalement non identifiante. L'application de la procédure de chaînage permet de reconstruire l'arbre généalogique d'un point de vue « vertical », c'est-à-dire ascendance/descendance d'un individu. Ce chaînage permet également de construire l'arbre généalogique d'un point de vue « horizontal », c'est-à-dire au sein d'une même génération. Lors de l'ajout d'un patient, le chaînage permet de détecter la présence d'individus identiques dans des familles différentes, en parcourant les individus déjà existants dans la base de données centrale de HC Forum®. Le patient pourra ainsi bénéficier d'un dossier de suivi médical accessible par les médecins prenant en charge ce patient, quel que soit le centre auquel il s'adresse, dossier qui sera enrichi

régulièrement d'informations individuelles et familiales. L'ensemble du dispositif est soumis à des contraintes de sécurité très strictes afin de garantir la confidentialité des données des patients et de celles de leurs parents. Aucune exploitation des données pour un objet autre que celui pour lequel elles ont été collectées n'est évidemment possible. Conformément aux dispositions des articles 27 et 40 de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, le patient dispose, ainsi que ses parents, d'un droit d'accès, de rectification et de suppression de ses données qu'il pourra exercer auprès d'un médecin généticien qui participe à la base HC Forum. Il pourra également, à tout moment, suspendre sa participation. Toute modification du dispositif d'accès au système sera portée à sa connaissance. Au vu des éléments ainsi présentés, il est demandé au patient de donner son consentement. L'ensemble de cette procédure a été validé par la CNIL qui a donné un avis favorable au cours de la délibération du 4 mars 2004 (avis n° 04-006).

4.2. La mise en place de systèmes d'information à l'échelon national

Plusieurs pays ont été intéressés par l'application des techniques de hachage, notamment dans le domaine de la santé, tels que le Luxembourg, qui a passé convention avec le DIM du CHU de Dijon pour l'anonymat du numéro matricule (aussi appelé numéro de sécurité sociale), ou encore la Suisse, qui a développé, en collaboration avec ce département, un système combinant les techniques de hachage et de cryptage des résumés hospitaliers [Borst *et al.*, 2001]. Seuls seront développés ici les systèmes d'information mis en place en France.

4.2.1 Le Programme de Médicalisation des Systèmes d'Information (PMSI)

Le CESSI/CNAMTS[7] « a conçu et fourni dès 1996 une fonction d'anonymisation appelé FOIN (Fonction d'Occultation d'Information Nominatives) [Trouessin et Allaërt, 1997], 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), sur demande du MES/DH[8] et après expertise du SCSSI. Cette fonction d'anonymisation permet de remplacer l'identité des patients par des numéros d'anonymat, ou clés de chaînage, pérennes dans le temps (tant que le secret de la fonction à sens unique est inchangé) et dans l'espace (pour chacune des quelques mille cliniques privées) » [Trouessin, rapport]. L'identifiant rendu anonyme est le numéro de Sécurité Sociale de l'assuré (ou ouvrant-droit) ainsi que la date de naissance et le sexe du patient. Ce système a été étendu à l'ensemble des établissements publics soumis au PMSI en 2001 [Circulaire n° 106 du 22 février 2001].

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

8. Direction des hôpitaux (DH) au Ministère de l'emploi et de la solidarité (MES), à l'époque Ministère du travail et des affaires sociales.

4.2.2 Le SNIIR-AM, Système d'Information de l'Assurance Maladie

Les ordonnances de 1996 ont instauré une immatriculation des bénéficiaires de l'assurance maladie dès leur naissance ou leur entrée en France. La loi de 1999 instituant la couverture maladie universelle (CMU) a apporté à ce système la première qualité attendue d'un dénombrement démographique, l'exhaustivité. Ces décisions ont été le préalable à la création d'un registre de bénéficiaires, sans double compte, le répertoire national inter-régimes de l'assurance maladie (RNIAM) fondé sur le numéro national d'identification (NIR) et qui identifie également la caisse primaire où est actuellement géré le dossier du bénéficiaire, afin d'assurer la cohérence entre gestion et démographie. Après un long cheminement depuis les ordonnances d'avril 1996, les régimes d'assurance maladie ont enfin créé le système national d'information inter-régime de l'assurance maladie (SNIIR-AM), qui offre des occasions exceptionnelles. Il s'agit d'une base de données individuelles par patient mais rendue anonyme et qui rassemble divers éléments à partir du RNIAM (permettant l'articulation du dispositif inter-régimes) : les données de remboursement avec le détail du codage des actes et du médicament, les identifiants des professionnels de santé et des établissements de santé qui ont participé aux soins des patients, les informations sur la pathologie traitée pour les patients en affection de longue durée et en maladie professionnelle. Ces données sont chaînées avec celles issues du PMSI : une clé de chaînage unique permet de relier les données hospitalières médicalisées du PMSI avec les données de médecine de ville, permettant ainsi d'établir le parcours médicalisé du patient [Merlière, 2004]. La CNIL a donné une autorisation de conservation pour une durée de deux ans, plus l'année en cours, concernant ces données individuelles exhaustives. Cet appariement de données sécurisées dans le temps et entre institutions est réalisé grâce à l'identifiant crypté de manière irréversible selon la technique de hachage décrite précédemment (p.7). Une demande d'accord, au titre du chapitre V ter de la loi sur l'informatique et les libertés, est en cours d'instruction pour permettre la constitution d'échantillons à partir du SNIIR-AM, sur de longues périodes. Le panel SNIIR-AM conserverait ainsi, sans limite de temps, les prestations reçues pour un échantillon permanent de 600 000 bénéficiaires. Cet échantillon représente une rénovation de l'échantillon permanent des assurés sociaux (EPAS) constitué en 1976 par le service statistique de la CNAMTS en collaboration avec la Division d'études médicales du CREDOC [9]. Contrairement à l'EPAS, le panel SNIIR-AM va permettre de s'appuyer sur une vraie unité démographique, le bénéficiaire.

4.2.3 Dispositif de surveillance des 26 maladies à déclaration obligatoire

Selon l'Institut de veille sanitaire (InVS), « composante essentielle de la santé publique et de l'épidémiologie, le dispositif de surveillance des maladies à déclaration obligatoire (MDO) est basé sur la transmission des données individuelles à l'autorité sanitaire. Il met en jeu deux procédures : le signalement

9. Centre de recherche pour l'étude et l'observation des conditions de vie.

et la notification, et repose sur une implication forte de trois acteurs : les déclarants (biologistes et médecins) qui suspectent et diagnostiquent les maladies à déclaration obligatoire, les médecins inspecteurs de santé publique et les Directions départementales des affaires sanitaires et sociales (Ddass) et leurs collaborateurs, chargés de réaliser la surveillance de ces maladies au niveau départemental, et les épidémiologistes de l'InVS» [Lettre InVS n° 8, 2003]. Le nouveau dispositif mis en place en 2003 «concilie deux éléments importants : une efficacité accrue du système de surveillance des MDO et une meilleure prise en compte des droits de l'individu, grâce à un système d'anonymisation unique au monde» [Lettre InVS n° 8, 2003]. Le DIM du CHU de Dijon a contribué à la première ébauche de ce dispositif, qui s'appuyait sur l'algorithme de hachage (décrit en p. 7), rédigée en collaboration avec le Docteur Denis Coulombier de l'InVS, et a participé, à la demande de la CNIL, aux réunions de discussion avec la CNIL et la Direction Centrale de la Sécurité et des Systèmes d'Information (DCSSI) qui ont conduit la CNIL à se prononcer favorablement, le 3 octobre 2001, pour la mise en place d'un système d'anonymisation à la source des éléments d'identification de la personne par technique de codage dite «irréversible». La CNIL a autorisé l'ensemble du dispositif par délibération n° 02-082 en date du 19 novembre 2002. Le développement des outils d'anonymisation à partir des algorithmes publics de hachage (*cf.* précédemment) a été effectué par la société Bertin au terme d'un appel d'offre européen. La procédure diffère quelque peu selon la pathologie considérée. Pour l'infection à VIH (séropositivité), le SIDA [10] et l'hépatite B aiguë, le médecin ou le biologiste déclarant effectuent l'anonymisation à la source, avant envoi à la Ddass de la fiche de notification. Le code d'anonymat est établi de façon irréversible par le logiciel à partir de l'initiale du nom, le prénom, la date de naissance et le sexe de la personne. Pour les autres maladies, le praticien transmet au médecin inspecteur de santé publique (Misp) une fiche de notification indiquant l'initiale du nom de famille, le prénom, le sexe et la date de naissance. La transmission se fait sous pli confidentiel portant la mention «secret médical». Après validation de la fiche, le Misp procède à l'anonymisation au moyen du logiciel fourni par l'InVS et ne transmet à l'InVS que la partie anonyme de la fiche de notification. Au moment de la saisie des fiches provenant de tous les départements dans les bases de données nationales, l'InVS procède à une seconde anonymisation (*cf.* précédemment). «Celle-ci crée un index à partir du premier code d'anonymat et d'une clé secrète détenue seulement par l'InVS. Cette seconde démarche rompt définitivement tout lien entre la personne et les données la concernant» [Lettre InVS n° 8, 2003].

10. VIH : virus de l'immunodéficience humaine. SIDA : syndrôme immunodéficitaire acquis, provoqué par le virus.

5. Applications dans les autres domaines (social,enseignement)

5.1. Type d'observation des entrées et sorties du RMI à Paris, mis en place par le CREDOC

Le public concerné par le revenu minimum d'insertion (RMI) est très diversifié, et sa durée de passage dans le dispositif est variable. Pour mieux connaître cette population et les facteurs de flux, le CREDOC a mis en place, à la demande de la Direction de l'action sociale, de l'enfance et de la santé (DASES) du département de Paris et de la Ddass de Paris, une observation des entrées et sorties du RMI à Paris. Elle repose sur une méthodologie originale qui croise des données provenant de neuf fichiers administratifs dont : le fichier de la caisse d'allocations familiales (CAF) de Paris, le fichier national de contrôle de la CNAF [11], pour les échanges entre Paris et les autres départements français, le fichier historique de l'ANPE [12] pour les passages par le chômage et les parcours professionnels, les déclarations uniques d'embauche recueillies par l'URSSAF [13] pour les recrutements dans le secteur privé, les données de gestion du Centre national pour l'aménagement des structures des exploitations agricoles pour les stages de formation financés par l'État ou la région et les types de contrat aidés, les informations de la cellule centrale de coordination sur les contrats d'insertion. Le point de départ de l'observatoire est une liste, fournie par la CAF de Paris, de 48 000 personnes passées par le RMI à Paris (fin 2000-début 2001) [Aldeghi et Simon, 2002; Aldeghi et Olm, 2004]. L'observatoire recherche les informations concernant ces mêmes personnes dans les huit autres fichiers administratifs cités ci-dessus. « Pour autoriser ce rapprochement entre sources, la CNIL a veillé à ce que la seule information circulant entre partenaires soit un identifiant crypté, créé à partir du NIR ou du numéro de matricule CAF. L'anonymisation se fait grâce à la procédure FOIN mise au point par le CESSI-CNAMTS (*cf.* p.15). Il est impossible à partir de ces identifiants cryptés de reconstituer les numéros initiaux » [Aldeghi et Simon, 2002; Aldeghi et Olm, 2004]. Le rapprochement des fichiers a été possible pour près de 38 000 personnes, celles dont le NIR était complet. Dans 21 % des cas, le NIR figurant dans les données de la CAF de Paris n'était pas complet, soit parce que la personne n'avait jamais été enregistrée à titre personnel à la sécurité sociale, soit parce qu'elle n'avait pas fourni cette information à la CAF. Dans ce cas, il n'était pas possible de retrouver les passages par le chômage, les stages, les emplois dans le privé ou les emplois en contrat aidé. L'observation, en se centrant sur les NIR complets, sous-estime la part des femmes, en particulier celles vivant en couple et les étrangères, dont le NIR est plus souvent incomplet. Ces travaux ont pu montrer la proximité des allocataires du RMI avec l'emploi, non seulement à leur sortie du RMI mais aussi tout au long de leur passage dans

11. Caisse nationale d'allocations familiales.
12. Agence nationale pour l'emploi.
13. Union pour le recouvrement des cotisations de sécurité sociale et d'allocations familiales.

le dispositif (et notamment le fait que les sortants du RMI soient davantage passés par les emplois aidés, confirmant l'intérêt de ces emplois pour favoriser la sortie du RMI). Ce rapport [Aldeghi et Simon, 2002; Aldeghi et Olm, 2004] confirme également la liaison négative supposée entre contrat d'insertion et sortie du RMI, d'autant plus importante pour les contrats d'insertion ayant exclusivement un caractère social. Enfin, ces travaux ont permis d'étudier les flux d'entrée et de sortie à l'intérieur du RMI et de préciser que 40 % des entrants étaient déjà passés par le RMI à Paris.

5.2. Le suivi des étudiants par le Ministère de l'éducation nationale

Le logiciel Anonymat, développé par le DIM du CHU de Dijon, a été mis à disposition du Ministère de l'éducation nationale, en vue du cryptage de l'identifiant des fichiers individuels du système d'information statistique sur les étudiants (SISE) suite à une convention entre le CHU de Dijon et ce ministère, signée en octobre 2003. À la demande d'Alain Goy (responsable du service des statistiques à l'époque), il s'agissait de rendre anonyme l'identifiant national étudiant (INE) à partir de la technique de hachage (décrite précédemment), afin de permettre le suivi des étudiants au niveau national, par la Direction de l'évaluation et de la prospective (directrice Mme Peretti) de ce ministère et en particulier par le Centre de l'informatique statistique et de l'aide à la décision (CISAD, responsable M. Dispagne), qui lui est rattaché, tout en respectant l'anonymat dû aux étudiants, selon la procédure autorisée par la CNIL le 27 mars 2003 [arrêté MENK0300893A, 2003]. Un projet d'extension de ce suivi aux élèves du secondaire est en cours. Ces questions sont développées par Alain Goy dans son article du présent dossier.

Remerciements : Ces travaux ont pu naître grâce à la volonté du Professeur Liliane Dusserre qui a su persuader les principaux responsables de la CNIL, le SCSSI et le Conseil de l'ordre des médecins, de l'utilité des techniques d'anonymisation dans le cadre du stockage des informations médicales des patients.

Références

ABRIAL V. (1998). *Les contrats d'objectifs entre les établissements publics de santé et l'agence régionale de l'hospitalisation : analyse d'environnement du CHU de St-tienne.* Thèse de Docteur en Médecine. Université de Franche-Comté.

Agence Régionale de l'Hospitalisation de Rhône-Alpes (1997). *Mission d'enquête sur les dépenses médicales et pharmaceutiques* : Lyon 6 novembre.

ALDEGHI I., SIMON M.-O. (2002). *Observatoire des entrées et sorties du RMI à Paris*, compte rendu de la première vague, direction des rapports – CREDOC – Décembre 2002 n° 226.

ALDEGHI I., OLM C. (2004). *Observatoire des entrées et sorties du RMI à Paris.* In Pascal Ardilly (éd.), « Échantillonnage et méthodes d'enquêtes », Dunod, Paris, pp 342-348.

Arrêté MENK0300893A du 23 avril 2003, *Bulletin officiel* de l'EN n° 18 du 1er mai.

BECKETT B. (1990). *Introduction aux méthodes de cryptologie*, Masson, Paris.

BORST F., ALLAERT F.-A., QUANTIN C. (2001). *The Swiss solution for anonymously chaining patient files.* Proc. MEDINFO 2001 ; IMIA : 1239-41.

BOUZELAT H. (1998). *Anonymat et chaînage de fichiers médicaux en vue d'études épidémiologiques.* Thèse de Docteur d'Université spécialiste en Informatique Médicale. Université de Bourgogne.

BRASSARD G. (1993). *Cryptologie contemporaine*, Masson, Paris.

BRENNER H., SCHMIDTMANN I., STEGMAIER C. (1997). Effects of record linkage errors on registry-based follow-up studies. *Statistics in Medicine*, 16(23), 2633-43.

Circulaire DHOS-PMSI-2001 n° 106 du 22 Février 2001 relative au chaînage des séjours en établissements de santé dans le cadre du programme de médicalisation des systèmes d'information (PMSI).

COHEN O., MERMET M.-A., DEMONGEOT J. (2001). HC Forum®: a web site based on an international human cytogenetic database. *Nucleic Acids Research*, 9, 305-307.

CORNET B., GOUYON J.-B., BINQUET C., SAGOT P., FERDYNUS C., MÉTRAL P., QUANTIN C. (2001). Évaluation régionale en périnatalité : mise en place d'un recueil continu d'indicateurs. *Revue d'Épidémiologie et de Santé Publique*, 49, 583-593.

Décret définissant les conditions dans lesquelles sont souscrites les déclarations et accordées les autorisations concernant les moyens et prestations de cryptologie, *n° 98-101* du 24 février 1998.

Décret fixant la liste des moyens et des prestations de cryptologie dispensées de toute formalité préalable, *n° 98-206* du 23 mars 1998.

Décret fixant la liste des moyens et des prestations de cryptologie pour lesquels la déclaration se substitue à l'autorisation, *n° 98-207* du 23 mars 1998.

Décret *n° 99-199* du 17 mars 1999 définissant les catégories de moyens et de prestations de cryptologie pour lesquelles la procédure de déclaration préalable est substituée à celle d'autorisation.

DOUGLAS S. (1996). *Cryptologie, théorie et pratique*, International Thomson Publishing.

FISHER F., MADGE B. (1996). Data security and patient confidentiality : the manager's role. *International Journal of Biomedical Computer*, 43, 115-119.

GOUYON B., MÉTRAL P., FROMAGET J., SAGOT P., GOUYON J.-B. (1999). Réseau périnatal de Bourgogne. *Technologie et Santé*, 37, 51-56.

JARO M.-A. (1995). Probabilistic-linkage of large public health data files. *Statistics in Medicine*, 14, 491-8.

Loi *n° 98-1266* du 30 décembre 1998 (article 107). Loi de finances pour l'année 1999.

Lettre de l'Institut de Veille Sanitaire, prévalence, n° 8, juillet 2003.

MARSAULT X. (1995). *Compression et cryptage des données multimédias*, Hermès, Paris.

MERLIÈRE Y. (2004). « *le SNIIR-AM* » communication aux Journées de Statistique 25 mai 2004, Montpellier.

QUANTIN C., BOUZELAT H., ALLAËRT F.-A. *et al.* (1998). Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine*, 37, 271-277.

QUANTIN C., ALLAERT F.-A., d'ATHIS P., DUSSERRE L. (1999). Can a database be anonymous? *MIE 99*, Slovenia, 22-26 août 1999, 297-301.

QUANTIN C., ALLAERT F.-A., DUSSERRE L. (2000a). Anonymous statistical methods versus cyrptographic methods in epidemiology. *International Journal of Medical Informatics*, 60, 177-83.

QUANTIN C., ALLAERT F.-A., BOUZELAT H., RODRIGUES J.-M., TROMBERT-PAVIOT B., BRUNET-LECOMTE P., GREMY F., DUSSERRE L. (2000b). La sécurité des réseaux d'informations médicales : application aux études épidémiologiques. *Revue d'Épidémiologie et de Santé Publique*, 48, 89-99.

QUANTIN C., BINQUET C., BOURQUARD K., PATTISINA R., GOUYON B., FERDYNUS C., GOUYON J.-B., ALLAERT F.-A. (2004). Which are the best identifiers for record linkage? *Medical Informatics and the Internet Medicine*, 29 (3-4), 221-227.

QUANTIN C., BINQUET C., ALLAERT F.-A., GOUYON B., PATTISINA R., LE TEUFF G., FERDYNUS C., GOUYON J.-B. (2005). Decision analysis for the assessment of a record linkage procedure : application to a perinatal network. *Methods of Information in Medicine*, 44, 72-79.

RIVEST R.L., SHAMIR A., ADLEMAN L. (1978). A method for obtaining digital signatures and public key cryptosystems, *CACM*, 2, 120.

SUGARMAN J.-R., HOLLIDAY M., ROSS A. *et al.* (1996). Improving American Indian cancer data in the Washington state cancer registry using linkages with the Indian health service and tribal records. *American Cancer Society*, 78 (7suppl), 1564-8.

SWEENEY L. (1998). *Three Computational Systems for Disclosing Medical Data in the Year 1999.* MEDINFO 98, IMIA, B. Cesnik, A. McCray, J.-R. Scherrer (Eds). IOS Press, Amsterdam, 1124-1129.

THIRION X., SAMBUC R., SAN MARCO J.-L. (1988). Epidemiology and anonymity : a new method. *Revue d'pidémiologie et Santé Publique*, 36, 36-42.

TROUESSIN G., ALLAËRT F.-A. (1997). *FOIN : a nominative information occultation function.* MIE, 3, 196-200.

TROUESSIN G. Rapport « qualité diagnostique et thérapeutique en cancérologie : communication d'informations multimédia dans un réseau sécurisé multidisciplinaire. Sécurité de l'information médicale en télémédecine », étude du ministère de la recherche.

VUILLET-TAVERNIER S. (2000). Réflexion autour de l'anonymat dans le traitement des données de santé. *Médecine et Droit*, 40, 1-4.

WILLENBORG L.C.R.J., de WALL A.G., KELLER W.J (1995). *Some Methodological Issues in Statistical Disclosure Control.* Statistics Netherlands, Department of Statistical Methods. Second Cathy Marsh Memorial Seminar, November 7^{th}, London.

ZIMMERMANN P. (1986). *A proposed standard format for RSA cryptosystems*, Boulder Software Engineering, Computer, 9, 21.