Bundesamt für Statistik
Office fédéral de la statistique
Ufficio federale di statistica
Uffizi federal da statistica

Abteilung Bevölkerung und Beschäftigung
Division de la population et de l'emploi
Divisione popolazione e impiego
Partiziun populaziun ed occupaziun

Statistik der stationären Betriebe des Gesundheitswesens

Der Datenschutz in der Medizinischen Statistik

Statistiques des établissements de santé (soins intra-muros)

La protection des données dans la statistique médicale

Bearbeitung: Bundesamt für Statistik, Sektion Gesundheit
Réalisation: Office fédéral de la statistique, Section de la santé

INHALTSVERZEICHNIS

Zusammenfassung 2
1. Einführung 4
2. Anonymer Verbindungskode 5
2.1 Vorgehen 5
2.2 Gewählte Lösungen 6
2.2.1 Identifizierende Variablen 6
2.2.2 Vorgehen 7
2.2.3 Unvollständige Kodes 7
2.2.4 Verteilung der Schlüssel 8
2.2.5 Änderung des privaten Schlüssels K 8
2.2.6 Übermittlungstests 9
3. Besonders schützenswerte Variablen 9
3.1 Grundsatz 9
3.2 Geburtsdatum 9
3.3 Wohnort 10
3.4 Staatsangehörigkeit 10
4. Datentransfer 10
4.1 Spital - kantonale Stellen 10
4.2 Kantonale Stellen - BFS 11
5. Weitergabe der Daten 11
5.1 Zuständige kantonale Stellen 11
5.2 Bearbeiten für Forschung 11
5.3 Übrige Benutzer 12
6. Aufbewahrung der Daten 12
7. MitarbeiterInnen und Amtsgeheimnis 13

Anhang I: «Beschreibung der vorgesehenen Methoden zur Gewährleistung des Persönlichkeitsschutzes», Sektion Kryptologie, EMD. (In französischer Sprache)

Anhang II: «Validitätstests der Auswahl der identifizierenden Variablen», Universitätsspital Genf. (In französischer Sprache)

TABLE DES MATIÈRES

Résumé 2
1. Introduction 4
2. Code de liaison anonyme 5
2.1 Procédé 5
2.2 Choix divers 6
2.2.1 Variables identifiantes 6
2.2.2 Procédure 7
2.2.3 Codes incomplets 7
2.2.4 Partage des clés 8
2.2.5 Modification de la clé secrète K 8
2.2.6 Tests de transmission 9
3. Variables sensibles 9
3.1 Principe 9
3.2 Date de naissance 9
3.3 Domicile 10
3.4 Nationalité 10
4. Transfert des données 10
4.1 Hôpital - offices cantonaux 10
4.2 Office cantonaux - BFS 11
5 Mise à disposition des données 11
5.1 Offices cantonaux compétents 11
5.2 Traitements à des fins de recherche 11
5.3 Autres Utilisateurs 12
6. Conservation des données 12
7. Collaborateurs et secret de fonction 13

Annexe I «Description des méthodes envisagées pour protéger la confidentialité des personnes», Section Fédérale de Cryptologie.

Annexe II «Test de validité du choix des variables identifiantes», Hôpitaux Universitaires de Genève.

Zusammenfassung

Mit dem vorliegenden Konzept für den « Datenschutz in der Medizinischen Statistik » gewährleistet das BFS den Schutz der Privatsphäre der hospitalisierten Personen bei gleichzeitiger Erfüllung der statistischen Ziele. Dazu stützt es sich auf die folgenden technischen Lösungen:

Verbindungskode:

Die Schaffung eines anonymen einheitlichen Verbindungskodes für jede hospitalisierte Person dient dazu, Fälle von Mehrfachhospitalisierungen zu erkennen, ohne dass die Anonymität der erhobenen Daten gefährdet wird. Der Verbindungskode wird mittels eines **Verfahrens zur Zerhackung (Hashing) und anschliessenden Verschlüsselung der identifizierenden Variablen** anonymisiert.

Besonders schützenswerte Variablen:

- Das **Geburtsdatum** wird nicht systematisch verlangt. Grundsätzlich wird nach dem Geburtsjahr und dem Alter in vollendeten Lebensjahren gefragt. Bei Todesfällen und bei Kindern unter zwei Jahren wird aus epidemiologischen Gründen das genaue Geburtsdatum verlangt.
- Der **Wohnort** wird in Form einer geographischen Region mit rund 10'000 Einwohnern erhoben. Diese Regionen werden von den kantonalen Ämtern vorgeschlagen; ihre Grenzen respektieren die Gebietseinteilung der Postleitzahlen.
- Nach der **Staatsangehörigkeit** werden nur Personen aus europäischen Ländern gefragt. Nichteuropäer werden in Gruppen zusammengefasst.

Datentransfer:

Aus Sicherheitsgründen werden die **Daten** bei ihrem Transfer auf elektronischen Datenträgern **verschlüsselt.**

Weitergabe von Daten:

Die zuständigen **kantonalen Stellen** erhalten sämtliche Daten, welche die Einwohner des entsprechenden Kantons betreffen. Personendaten können der Forschung und Wissenschaft punktuell und unter bestimmten Voraussetzungen zugänglich gemacht werden. Die übrigen Benutzer erhalten Daten in aggregierter Form.

Aufbewahrung von Daten:

Die **Verbindungskodes werden nach 10 Jahren vernichtet**. Für epidemiologische Zwecke werden die Kodes von Kindern unter 15 Jahren, Personen über 64 Jahren sowie einer Stichprobe von Personen im Alter von 15-64 Jahren aufbewahrt. Die übrigen Daten, welche im Rahmen der Medizinischen Statistik erhoben werden, werden zeitlich unbegrenzt aufbewahrt.

Résumé

Par cette conception de « protection des données dans la statistique médicale », l'OFS assure le respect de la vie privée des personnes hospitalisées tout en réalisant les buts statistiques qui lui sont demandés. Les choix techniques sont les suivants:

Le code de liaison:

La création d'un code de liaison anonyme uniforme pour chaque personne hospitalisée permet de reconnaître les cas de réhospitalisation sans mettre en danger l'anonymat des données recueillies. Le code de liaison est rendu anonyme par une **procédure de hachage suivi d'un cryptage de variables identifiantes.**

Les variables sensibles:

- La **date de naissance** n'est plus systématiquement requise. Sont relevés, en principe, l'année de naissance et l'âge en années révolues. Pour des besoins épidémiologiques, la date de naissance complète est demandée lors des décès et pour les enfants de moins de deux ans.
- Le **domicile** est relevé sous forme de région d'environ 10'000 habitants. Ces régions sont proposées par les offices cantonaux et leurs limites respectent les aires des numéros postaux.
- La **nationalité** n'est demandée que pour les ressortissants des pays européens. Des regroupements sont effectués pour les extra-européens.

Le transfert des données:

Pour des raisons de confidentialité, **les données seront cryptées** lors de leur transport sur support télématique.

La mise à disposition des données:

Les **offices cantonaux** compétents reçoivent l'intégralité des données concernant leurs habitants du canton en question. Des données individuelles peuvent être mises ponctuellement et sous certaines conditions à la disposition de chercheurs. Les autres utilisateurs reçoivent des données sous forme agrégée.

La conservation des données:

Les **codes de liaison sont détruits après 10 ans.** Pour des besoins épidémiologiques, les codes sont conservés pour les enfants de moins de 15 ans, pour les personnes de plus de 64 ans, ainsi que pour un sous-collectif de la population entre 15 et 64 ans. Les autres données récoltées dans la statistique médicale sont conservées sans limite de temps.

Amtsgeheimnis:

Die Mitarbeiter des BFS und die Personen, welche mit den Daten in Berührung kommen, unterliegen dem **Amtsgeheimnis**. Innerhalb des BFS werden nur die direkt mit dem Projekt befassten Mitarbeiter der Sektion Gesundheit Zugang zu den Datenbanken haben.

Le secret de fonction:

Les collaborateurs de l'OFS ainsi que les personnes en contact avec les données sont soumis au **secret de fonction**. A l'OFS, seuls les collaborateurs de la Section de la santé directement impliqués dans le projet auront un accès aux bases de données.

Der Datenschutz in der Medizinischen Statistik des Bundes

1. Einführung

Der Datenschutz steht im Zentrum der Debatte über die Medizinische Statistik. Die Reaktionen der Medien (Presse und Fernsehen) und der befragten Partner haben den äusserst heiklen und zuweilen emotionalen Charakter der Frage offenbart.

Der Kern des Problems liegt in der Notwendigkeit, Fälle von Mehrfachhospitalisierungen zu erkennen, ohne dass die Anonymität der erhobenen Daten gefährdet wird. Es wurde eine technische Lösung gesucht, die diese zwei Anforderungen zu erfüllen vermag. Zusammen mit Fachleuten auf dem Gebiet der Datensicherheit nahm sich das BFS dieser Frage an und ist heute in der Lage, eine Lösung vorzuschlagen, die bei den verschiedenen Partnern Anklang finden dürfte.

Vorgesehen ist eine Anonymisierung der Personendaten mit Hilfe eines anonymen einheitlichen Verbindungskodes (eine Person = ein Kode).

Anlässlich einer Arbeitstagung der Schweizerischen Gesellschaft für medizinische Informatik (SGMI) zum Thema « Wie können informatisierte medizinische Daten geschützt werden? » stellte Dr. François-André Allaert, Generalsekretär der *Association française d'informatique médicale* und Gründer der *Agence pour la protection de la sécurité des informations de la santé (APSIS)* die technischen Lösungen vor, die in Frankreich und auf einer breiteren europäischen Ebene bereits mit Erfolg verwendet werden. Das System lässt sich im wesentlichen mit dem hier vorgestellten vergleichen.

Neben der Schaffung eines einheitlichen anonymen Verbindungskodes gilt es auch den übrigen Grundsätzen des Datenschutzes Rechnung zu tragen, insbesondere in bezug auf:

- die **Vertraulichkeit:** "Alle mit der Durchführung der Erhebungen betrauten Personen und Amtsstellen sind verpflichtet, die erhobenen Daten vertraulich zu behandeln. Sie sorgen dafür, dass die Daten an einem sicheren Ort aufbewahrt werden" (Art. 7 der Verordnung über die Durchführung von statistischen Erhebungen des Bundes)

- das **Amtsgeheimnis** (einschliesslich für Kantone): "Die mit statistischen Arbeiten betrauten Personen müssen alle Daten... geheimhalten" (Art. 14 BStatG) und "Für die Bearbeitung durch kantonale Organe gelten die Artikel 14... (Art. 17 BStatG)

La protection des données dans la statistique médicale fédérale

1. Introduction

La protection des données est au coeur du débat sur la statistique médicale. Les réactions apparues dans les médias (presse et télévision) et lors de la consultation des partenaires ont montré le caractère extrêmement sensible et parfois émotionnel de la question.

Le noeud du problème réside dans la nécessité de reconnaître les cas de réhospitalisation sans mettre en danger l'anonymat des données recueillies. Pour répondre à ces deux exigences, on a recherché une solution technique. Avec le concours d'experts en sécurité des données, l'OFS a planché sur la question et est en mesure de proposer une solution capable de fédérer les différents partenaires.

Il s'agit d'une anonymisation des informations personnelles au moyen d'un code de liaison anonyme uniforme (une personne égale un code).

Lors d'une journée de travail de la Société suisse d'informatique médicale consacrée au thème « Comment protéger les données médicales informatisées », le Dr François-André Allaert, Secrétaire général de l'Association française d'informatique médicale et fondateur de l'Agence pour la protection de la sécurité des informations de la santé (APSIS) a présenté les solutions techniques utilisées avec succès en France et plus largement au niveau européen. Il s'agit d'un système fondamentalement comparable à celui présenté ici.

En plus de la création d'un code de liaison anonyme uniforme, le respect des autres principes de protection des données s'impose, notamment:

- la **confidentialité**: "les personnes et les services chargés d'exécuter les relevés sont tenus de traiter les données de manière confidentielle, ils veillent à ce que les données soient conservées en lieu sûr" (art. 7 de l'ordonnance sur l'exécution des relevés statistiques fédéraux)

- le **secret de fonction** (y compris pour les cantons) "les personnes chargées de travaux statistiques sont tenues de garder le secret ..." (art. 14 de la loi sur la statistique fédérale, LSF) et "le traitement de données par les organes cantonaux est régi par les articles 14..." (art. 17 LSF)

- die **technischen Massnahmen**: Die Statistikproduzenten des Bundes dürfen Personendaten für nicht personenbezogene Zwecke, insbesondere für Forschung, Planung und Statistik, Forschungs- und Statistikstellen des Bundes sowie Dritten bekanntgeben, wenn: a. die Daten anonymisiert werden, sobald es der Zweck des Bearbeitens erlaubt; b. der Empfänger die Daten nur mit Zustimmung des Statistikproduzenten weitergibt; c. der Empfänger die Ergebnisse nur so bekanntgibt, dass die betroffenen Personen nicht bestimmbar sind; und d. die Voraussetzungen für die Einhaltung des Statistikgeheimnisses und der übrigen Datenschutzbestimmungen durch den Empfänger gegeben sind (Art. 19 BStatG).

In den nachfolgenden Kapiteln werden diese verschiedenen Aspekte näher behandelt.

2. Anonymer Verbindungskode

Das BFS erteilte der Sektion Kryptologie des Eidgenössischen Militärdepartements den Auftrag, eine Methode zum Schutz der Vertraulichkeit der Personendaten in der Medizinischen Statistik zu entwickeln.

2.1 Vorgehen

Die Fachleute der Sektion Kryptologie legten ihren Bericht im Februar 1997 vor. Darin wird die Schaffung eines anonymen einheitlichen Verbindungskodes mittels Zerhackung (Hashing) und anschliessender Verschlüsselung der identifizierenden Daten der hospitalisierten Personen vorgeschlagen. Mit einer zusätzlichen Verschlüsselung der zerhackten Daten auf der Ebene der Betriebe und der anschliessenden Entzifferung im BFS wird für eine maximale Sicherheit der besonders schützenswerten Angaben während des Datentransfers von den Spitälern ans BFS gesorgt. Dieses Vorgehen verhindert die Möglichkeit eines « Angriffs via Wörterbuch » auf der Ebene der kantonalen Ämter. Eine detaillierte Beschreibung des Verfahrens in französischer Sprache findet sich in **Anhang I.**

Der Datenschutzexperte Dr. Allaert erachtete die vorgeschlagene Methode als sehr interessant. Die Verschlüsselung der Daten auf der Ebene der Spitäler stellt eine zusätzliche Sicherung dar, die ihn ganz besonders interessierte.

Die Fachleute bestätigen, dass die Methode der Zerhackung (Hashing) und anschliessenden Verschlüsselung ein genauso sicheres Verfahren darstellt wie dasjenige von zwei aufeinanderfolgenden Zerhackungen (Hashings), das von unseren europäischen Nachbarn verwendet wird und sogar gewisse Vorteile in sich birgt (vgl. Anhang I).

- les **mesures techniques**: les producteurs de statistiques de la Confédération sont en droit de communiquer des données personnelles à des services de statistique, à des institutions de recherche de la Confédération ou à des tiers, à des fins ne se rapportant pas à des personnes, notamment dans le cadre de la recherche, de la planification ou de la statistique, si: a. ces données sont rendues anonymes dès que le but du traitement le permet; b. le destinataire ne communique ces données à des tiers qu'avec l'accord de l'organe qui les a produites; c. la forme choisie par le destinataire pour communiquer les résultats ne permet pas d'identifier les personnes concernées et d. tout porte à croire que le destinataire respectera le secret statistique et les autres dispositions relatives à la protection des données (art. 19 LSF).

Le présent document traite plus précisément de ces différents aspects dans les chapitres qui suivent.

2. Code de liaison anonyme

L'OFS a mandaté la Section de cryptologie du Département militaire fédéral pour développer une méthode de protection de la confidentialité des données personnelles dans la statistique médicale.

2.1 Procédé

Les experts de la Section fédérale de cryptologie ont rendu leur rapport en février 1997. Dans ce document, la création d'un code de liaison anonyme uniforme par hachage et cryptage des données identifiantes des personnes hospitalisées est proposée. Un cryptage supplémentaire des données hachées au niveau des institutions suivi d'un décryptage au niveau de l'OFS permet une sécurité maximale des informations sensibles durant le trajet des hôpitaux à l'OFS. Ce procédé exclut les possibilités d'« attaque par dictionnaire » au niveau des offices cantonaux. La procédure est décrite en détail dans l'**annexe I**.

Le Dr Allaert, spécialiste en protection des données, a trouvé la technique proposée très intéressante. Le cryptage des données au niveau des hôpitaux apporte une sécurité supplémentaire qui l'a vivement intéressé.

Les spécialistes affirment que l'utilisation d'un hachage suivi d'un cryptage est un procédé aussi sûr que les deux hachages successifs utilisés chez nos voisins européens et qu'il doit être préféré en raison des avantages qu'il apporte (voir annexe I).

2.2 Gewählte Lösungen

Das BFS übernimmt grundsätzlich die Vorschläge der Kryptologie-Experten. Verschiedene offene Fragen sind nun gelöst.

2.2.1 Identifizierende Variablen

Die identifizierenden Angaben, die nach der Zerhackung (Hashing) und Verschlüsselung den anonymen Verbindungskode bilden, müssen verschiedene Kriterien erfüllen:

- **Diskriminanz**, um auszuschliessen, dass verschiedene Personen denselben Verbindungskode erhalten,
- **Stabilität**, um Veränderungen des Verbindungskodes im Laufe des Lebens einer Person auf ein Minimum zu beschränken,
- **Unmittelbare Verfügbarkeit** in den Informationssystemen der Spitäler, um zu garantieren, dass die Variablen zum Zeitpunkt der Erhebung abrufbar sind,
- **Einfachheit,** um Fehler (z.B. Orthographiefehler) bei der Datenerhebung auf ein Minimum zu beschränken
- **Präzision,** um Abweichungen von Betrieb zu Betrieb zu vermeiden.

Diese Kriterien beeinflussen sich gegenseitig und einige sind leider widersprüchlich. Das BFS beschloss, eine Senkung des Risikos von falsch-positiven sei den falsch-negativen Fallzusammenführungen vorzuziehen. Oder anders ausgedrückt: Es scheint uns weniger problematisch, Fälle von Mehrfachhospitalisierungen zu verpassen, als solche zu erzeugen.

Identifizierende Angaben sind:

- der «Soundex»-Kode des Geschlechtsnamens (1 Buchstabe, drei Ziffern)
- der «Soundex»-Kode des Vornamens (1 Buchstabe, drei Ziffern)
- das vollständige Geburtsdatum (8 Ziffern)
- das Geschlecht (1 Ziffer)

d.h. insgesamt 17 alphanumerische Zeichen

2.2 Choix divers

En principe, l'OFS retient les propositions des experts en cryptologie. Quelques questions ouvertes sont maintenant résolues.

2.2.1 Variables identifiantes

Les données identifiantes qui, après hachage et cryptage, composent le code de liaison uniforme doivent satisfaire à plusieurs critères:

- **discriminance,** pour exclure que des personnes différentes se voient attribuer un même code de liaison,
- **stabilité,** pour minimiser les variations du code de liaison au cours de la vie d'un individu,
- **disponibilité** immédiate dans les systèmes d'information des hôpitaux, pour garantir que les variables sont disponibles au moment de la saisie,
- **simplicité,** pour minimiser les erreurs lors de la saisie des données (par ex. orthographiques),
- **précision,** pour éviter des variations d'une institution à l'autre.

Ces critères s'influencent réciproquement et plusieurs d'entre eux sont malheureusement contradictoires. L'OFS a décidé de privilégier la réduction des risques de faux positifs au détriment des faux négatifs. En d'autres termes, il nous paraît moins important de manquer des cas de réhospitalisation que d'en inventer.

Les données identifiantes sont les suivantes:

- le code « Soundex » du nom de famille (1 caractère, trois chiffres),
- le code « Soundex » du prénom (1 caractère, trois chiffres),
- la date de naissance complète (8 chiffres),
- le sexe (1 chiffre).

Soit, au total, 17 caractères alphanumériques.

Die Umwandlung des Vor- und Nachnamens in «Soundex»-Kodes hat den Vorteil, dass Abweichungen in der Schreibweise der Vornamen und der Geschlechtsnamen stark verringert werden. Enthält ein Name ein «von», so wird dieses systematisch vorangestellt. «Zwischenräume», «Apostrohpe» und «Gedankenstriche» gelten als «Nicht-Zeichen».
Die Auswahl der identifizierenden Variablen wurde von Dr. Borst von der Abteilung für medizinische Informatik des Genfer Universitätsspitals anhand einer bestehenden Datei von 222‘020 verschiedenen Patienten getestet. 221'409 Einzelkombinationen, 304 Doppelkombinationen und 1 Dreifachkombination gingen daraus hervor, was einer Verwechslungsquote von 0,3% (falsch-Positive) entspricht. Diese wird folgendermassen berechnet: (2x304)+(3x1)/222‘020. Weitere Angaben zu diesem Test finden sich in französischer Sprache in **Anhang II**. Die Diskriminanzeigenschaften der Methode erscheinen uns im Vergleich zu den anderen Prozessfehlern (Stabilität, Präzision und Verfügbarkeit) als hinreichend.

La transformation du nom et du prénom en codes « Soundex » présente l’avantage de réduire fortement les variations orthographiques des noms et des prénoms. Si un nom comporte une particule, celle-ci est systématiquement placée devant. Les « espaces », les « apostrophes » et les « tirets » sont considérés comme des « non-caractères ».

Le choix des variables identifiantes a été testé sur un fichier réel de 222’020 patients différents par le Dr Borst de la division d’informatique médicale des Hôpitaux Universitaires de Genève. 221’409 combinaisons uniques, 304 combinaisons doubles et 1 combinaison triple sont apparues, soit un taux de confusion de 0,3% (faux positifs). Le calcul est le suivant: (2x304)+(3x1))/222’020). Voir à ce sujet les détails de l’**annexe II**. La discriminance de la méthode nous paraît confortable en comparaison des autres erreurs du processus (stabilité, précision et disponibilité).

2.2.2 Vorgehen

Das Vorgehen wurde direkt aus dem Bericht der Kryptologie-Experten übernommen:

- Unidirektionale Zerhackung (Hashing) der Daten auf der Ebene der Spitäler nach dem SHA-Algorithmus,
- Verschlüsselung des daraus entstehenden Pseudonyms auf der Ebene der Spitäler mit Hilfe des IDEA-Algorithmus und eines privaten Schlüssels (c) von 128 Bit,
- Verschlüsselung des Schlüssels (c) auf der Ebene der Spitäler mit Hilfe des RSA-Algorithmus und einem öffentlichen Schlüssel (E) des BFS,
- Entschlüsselung des privaten Schlüssels (c) im BFS mit RSA unter Verwendung des privaten Schlüssels (D) des BFS,
- Entschlüsselung des verschlüsselten Pseudonyms mit IDEA im BFS unter Verwendung des privaten Schlüssels (c),
- Generierung eines einheitlichen anonymen Verbindungskodes im BFS durch erneute Verschlüsselung des Pseudonyms mit Hilfe des IDEA-Algorithmus und einem privaten Schlüssel (K) von 128 Bit.

2.2.2 Procédure

La procédure est reprise directement du rapport des experts en cryptologie, soit:

- hachage à sens unique des données au niveau des hôpitaux selon la fonction SHA,
- cryptage de l’empreinte au niveau des hôpitaux selon l’algorithme IDEA avec une clé secrète (c) de 128 bits,
- cryptage de la clé (c) au niveau des hôpitaux selon l’algorithme RSA avec une clé publique (E) de l’OFS,
- décryptage de la clé (c) à l’OFS avec RSA en employant la clé privée (D) de l’OFS,
- décryptage de l’empreinte à l’OFS avec IDEA en employant la clé secrète (c),
- création du **code de liaison anonyme uniforme** à l’OFS par recryptage de l’empreinte selon l’algorithme IDEA avec une clé secrète (K) de 128 bites.

2.2.3 Unvollständige Kodes

Wenn bei der Generierung des Pseudonyms auf der Ebene der Spitäler eine identifizierende Variable fehlt, nicht verfügbar, unvollständig oder nicht mit Sicherheit bekannt ist, so wird der Verbindungskode als Ganzes unbrauchbar.

2.2.3 Codes incomplets

Si lors de la création de l’empreinte au niveau de l’hôpital, une des variables identifiantes manque, n’est pas disponible, est incomplète ou encore n’est pas connue avec certitude, l’intégralité du code de liaison devient inutile.

In diesem Fall werden alle bekannten Angaben weggelassen und stattdessen durch 17 Nullen ersetzt. Der daraus hervorgehende Verbindungskode wird dann als nicht signifikant erkannt.

Il s'agira, dans ce cas, de supprimer toutes les données connues et d'indiquer en lieu et place 17 zéros. Le code de liaison issu de la procédure sera reconnu comme non significatif.

2.2.4 Verteilung der Schlüssel

Die Fachleute der Sektion Kryptologie des EMD schlagen eine Verteilung der Schlüssel (D und K) auf drei Personen vor, zwischen denen keine private oder berufliche Verbindung bestehen sollte. Dieses Prinzip garantiert zwar einen äusserst wirksame Geheimhaltung der Schlüssel, es lässt sich aber hier nur sehr schwer anwenden und erscheint uns angesichts der vorliegenden Problematik als eine übertriebene Sicherheitsmassnahme.

Der Datenfluss zwischen den Kantonen und dem BFS lässt sich nicht genau planen. Die Arbeitsgänge zur Entschlüsselung und erneuten Verschlüsselung der Pseudonyme bei der Schaffung der anonymen Verbindungskodes werden nicht an einen bestimmten Tag gebunden sein, sondern sich je nach Datenlieferung und -validierung über mehrere Monate hinwegziehen. Gewisse Kantone haben bereits angekündigt, dass sie ihre medizinischen Informationen auswerten wollen, bevor sie sie ans BFS übermitteln, und zwar aus terminlichen Gründen. In diesem Fall muss das BFS auf der Basis von verschlüsselten Pseudonymen Verbindungskodes schaffen und diese den Kantonen vorzeitig zur Verfügung stellen.

Aus diesen Gründen scheint es uns schwierig und kontraproduktiv, drei unabhängige und geographisch getrennte Personen wiederholte Male zu versammeln.

Wir halten uns daher an die Zahl von drei Personen, diese werden aber allesamt dem BFS angehören. Hierarchische Beziehungen sollen dabei vermieden werden.

2.2.4 Partage des clés

Les experts de la Section fédérale de cryptologie proposent un partage des clés (D et K) entre trois personnes qui ne devraient pas avoir de liens entre elles, tant d'un point de vue professionnel que privé. Ce principe, qui assure une protection très efficace du secret des clés, est malheureusement très difficilement applicable ici et nous paraît représenter un excès de sécurité disproportionné pour cette problématique.

En effet, le flux de données entre les cantons et l'OFS n'est pas précisément planifiable. Les opérations de décryptage et recryptage des empreintes lors de la création des codes de liaison ne seront pas réalisées un jour donné mais auront lieu sur plusieurs mois en fonction des livraisons et des validations de données. Certains cantons ont déjà annoncé qu'ils désirent exploiter leurs informations médicales avant de les transmettre à l'OFS, ceci pour des questions de délais. Dans ce cas, l'OFS devra créer des codes de liaison à partir d'empreintes cryptées et les fournir prématurément aux cantons.

Dans ce cadre d'activité, il nous paraît difficile et contre-productif de réunir de multiples fois trois personnes indépendantes et séparées géographiquement.

Nous retenons donc le nombre de trois personnes, mais ces personnes seront toutes trois issues de l'OFS. Les liens hiérarchiques seront évités.

2.2.5 Änderung des privaten Schlüssels K

Der Entscheid für eine Verschlüsselung des zerhackten Pseudonyms, anstelle einer doppelten Zerhackung (Hashing) ermöglicht eine Umkehrung des Prozesses. Dieser Vorteil lässt sich für eine alljährliche Änderung des Schlüssels nutzen, ohne dass dadurch die Möglichkeit verloren geht, Mehrfachhospitalisierungen im Laufe der Jahre zu erkennen.

Jedesmal wenn die zentrale Datenbank um die Datenbasis eines Jahres erweitert wird, werden die aufzubewahrenden Verbindungskodes entschlüsselt und mit Hilfe eines neuen Schlüssels neu verschlüsselt. Näheres dazu in Kapitel 6, «Datenaufbewahrung».

2.2.5 Modification de la clé secrète K

Le choix d'une procédure de cryptage de l'empreinte hachée plutôt que d'un deuxième hachage permet une réversion du processus. Cet avantage est utilisé pour changer de clé chaque année sans perdre la possibilité de suivre des cas de réhospitalisation d'une année à l'autre.

Chaque fois qu'une base de données annuelle est ajoutée à la base de données générale, les codes de liaison à conserver sont décryptés et recryptés avec une nouvelle clé. Voir à ce sujet le chapitre 6. « Conservation des données ».

2.2.6 Übermittlungstests

Der Header jedes Datenpakets eines jeden Spitals wird aus Daten bestehen, die vorgängig vom BFS bestimmt werden. Diese dienen als Testvariablen zur Überprüfung der Übertragung und für die Generierung der Verbindungskodes.

3. Besonders schützenswerte Variablen

3.1 Grundsatz

Unter besonders schützenswerten Variablen verstehen wir Variablen, die ein potentielles Erkennungsrisiko für die betroffenen Personen bergen und nicht Angaben, die Teil der Privatsphäre besagter Personen sind.

3.2 Geburtsdatum

Das Geburtsdatum ist besonders schützenswert, da es ein hohes Erkennungsrisiko in sich birgt. Obwohl es ein wichtiges soziodemographisches Merkmal darstellt, das bei zahlreichen anderen statistischen Projekten erhoben wird, schien es uns wesentlich, unsere Bedürfnisse auf das absolut Notwendige zu beschränken.

Das genaue Geburtsdatum (Tag, Monat, Jahr) wird nur bei Kindern unter zwei Jahren und bei Todesfällen erhoben. Bei den übrigen hospitalisierten Personen wird nur das Geburtsjahr und das Alter (in vollendeten Lebensjahren) zu Beginn der Hospitalisierung verlangt. Dies aus folgenden Gründen:

- Die Erfassung des genauen Geburtsdatums bei Geburt und systematisch bei Kleinkindern unter zwei Jahren ist unentbehrlich für verfeinerte epidemiologische Studien in den ersten Lebensmonaten. In dieser Periode kann jeder Tag von Bedeutung sein.

- Das Geburtsdatum bei Todesfällen ist nützlich, weil es eine Verbindung zur Statistik der Todesfälle ermöglicht. Der Verbindungskode schafft die Voraussetzung, um die Hospitalisierungsgeschichte eines verstorbenen Patienten rückblickend während seiner letzten Jahre verfolgen zu können.

- Das Geburtsjahr wird für epidemiologische Längsschnittanalysen benötigt und kann nicht aufgrund des Alters in vollendeten Lebensjahren berechnet werden.

- Umgekehrt ist das Alter in vollendeten Lebensjahren grundlegend für die Querschnittsanalyse (in spezifischen Kalenderjahren) und dieses lässt sich wiederum nicht aufgrund des Geburtsjahres berechnen.

2.2.6 Tests de transmission

Le premier enregistrement de chaque lot de données de chaque hôpital sera composé de données prédéfinies par l'OFS. Ces variables serviront de test et permettront la validation de la transmission et de la création des codes de liaison.

3. Variables sensibles

3.1 Principe

Nous entendons par variables sensibles celles qui présentent un risque d'identification des personnes et non les informations qui font partie de la sphère privée de ces mêmes personnes.

3.2 Date de naissance

La date de naissance est très sensible car elle possède un important pouvoir discriminant. Bien qu'il s'agisse d'une donnée sociodémographique importante et recueillie dans de nombreux autres projets statistiques, il nous a paru fondamental de réduire nos besoins au strict nécessaire.

Il s'agit de relever la date de naissance précise (jour, mois, année) pour les enfants de moins de deux ans révolus et lors des décès. Pour les autres cas d'hospitalisation, l'année de naissance ainsi que l'âge en années révolues lors du début de l'hospitalisation sont requis. Les justifications sont les suivantes:

- La date de naissance précise lors de la naissance et systématiquement pour les enfants de moins de deux ans est indispensable pour des études épidémiologiques fines dans les premiers mois de vie. Durant cette période, chaque jour peut être important.

- La date de naissance lors du décès est utile pour effectuer une liaison avec la statistique des décès. Le code de liaison permettra de suivre rétrospectivement, dans ses dernières années, l'historique hospitalier d'un patient décédé.

- L'année de naissance est nécessaire pour les études épidémiologiques longitudinales et n'est pas calculable à partir de l'âge en années révolues.

- Inversement, l'âge en années révolues est essentiel pour l'étude transversale (par rapport à des années civiles particulières) et n'est pas calculable à partir de l'année de naissance.

3.3 Wohnort

Die Wohngemeinde ist besonders schützenswert - vor allem im Fall von kleineren Gemeinden, weil sie ebenfalls ein hohes Erkennungsrisiko in sich birgt. Ihre Bedeutung und ihr Nutzen für eine landesweite Statistik sind nicht nachgewiesen. Die Erhebung des Wohnorts der hospitalisierten Personen kann für die kantonale Spitalplanung von Nutzen sein. Den Kantonen steht es deshalb frei, diese Angaben zu erheben und auszuwerten.

Für das Projekt der nationalen Medizinischen Statistik ist die Angabe des genauen Wohnorts nicht erforderlich. Die Herkunftsregion jeder hospitalisierten Person genügt. Diese Regionen werden vom BFS festgelegt und müssen verschiedene Kriterien erfüllen:

- strikte Einhaltung der Kantonsgrenzen,
- strikte Einhaltung der Zustellgebiete der Postleitzahlen,
- möglichst weitgehende Einhaltung der Gemeindegrenzen
- Abdeckung einer Bevölkerung von rund 10'000 Einwohnern (im Minimum 3‘500 Einwohner)

Die kantonalen Stellen werden gebeten, dem BFS Vorschläge für eine regionale Aufteilung ihres Hoheitsgebiets zu unterbreiten. Die Entscheidungskompetenz liegt beim BFS. Dieses stellt seinen Partnern eine Umwandlungstabelle «Postleitzahlen – Typologie der Regionen» zur Verfügung.

3.4 Staatsangehörigkeit

Bei der Staatsangehörigkeit stellen sich grundsätzlich dieselben Probleme wie bei der Wohngemeinde. Während die Europäer nach der genauen Staatsangehörigkeit gefragt werden, sind die Aussereuropäer nach geographischen Herkunftsregionen zusammenzufassen. Eine Umwandlungstabelle «ISO-Nationalitätenkode - Typologie der Regionen» wird den Partnern zur Verfügung gestellt.

4. Datentransfer

Der Schutz der Daten bei der Übermittlung ist von grosser Bedeutung, damit diese nicht in den Besitz von Dritten gelangen und für unrechtmässige Zwecke missbraucht werden können.

4.1 Spital - kantonale Stellen

Das BFS ist im Prinzip nicht befugt, in den Prozess der Datenerhebung durch die Kantone einzugreifen. Den betreffenden Stellen wird jedoch wärmstens empfohlen, Vorkehrungen zur Sicherung der Daten während ihres Transfers zu treffen, egal ob dieser per Post via Disketten oder auf Netzwerken erfolgt.

3.3 Domicile

La commune de domicile, principalement pour les petites communes, est très sensible car elle possède également un important pouvoir discriminant. Sa significativité et son utilité ne sont pas démontrées au niveau d’une statistique nationale. Le relevé de la commune de domicile des personnes hospitalisées peut être utilisé lors de l’établissement d’une planification cantonale. Les cantons sont donc compétents pour relever ou non ces données et les exploiter.

Au niveau du projet de statistique médicale nationale, le domicile exact n’est pas demandé. Il s’agit d’indiquer la région de provenance de chaque personne hospitalisée. Ces régions sont définies par l’OFS et doivent remplir plusieurs critères:

- respecter strictement les limites cantonales
- respecter strictement les limites des aires des numéros postaux,
- respecter dans la mesure du possible les limites communales,
- couvrir une population d’environ 10’000 habitants (mais au minimum de 3’500 habitants).

Les offices cantonaux sont priés de faire, pour leur juridiction, des propositions de regroupements régionaux à l’OFS qui décidera. L’OFS mettra à la disposition des partenaires une table de conversion « numéros postaux - typologie des régions ».

3.4 Nationalité

La nationalité pose en principe les mêmes problèmes que la commune de domicile. Si la nationalité est demandée précisément pour les ressortissants des pays européens, dans le cas de nationalités extra-européennes, des regroupements par zones géographiques sont effectués. Une table de conversion « code ISO des nationalités - typologie des régions » sera mise à la disposition des partenaires.

4. Transfert des données

Il est essentiel d’assurer la sécurité des données lors de leur transport afin que des tiers ne puissent s’en emparer et les exploiter à des fins illicites.

4.1 Hôpital - offices cantonaux

En principe, l’OFS n’a pas à intervenir dans le processus de récolte des données par les cantons. Il est toutefois vivement recommandé à ces offices de mettre en place des systèmes assurant la sécurité des données lors de leur transport, qu’il s’effectue par voie télématique ou par voie postale.

4.2 Kantonale Stellen – BFS

Der Datentransfer von den kantonalen Stellen zum BFS erfolgt grundsätzlich auf elektronischem Weg. Es ist jedoch nicht auszuschliessen, dass die Daten anfänglich via Disketten per Post übermittelt werden.

Im zweiten Fall scheint uns ein Versand der Disketten per Einschreiben zu genügen. Allenfalls ist eine Verschlüsselung der Daten auf den Disketten in Betracht zu ziehen.

Im Falle eines Netzwerk-Transfers ist die gesamte Datei in verschlüsselter Form zu übermitteln. Damit wird der Persönlichkeitsschutz während der Übertragung gewährleistet.

5. Weitergabe der Daten

Die Rohdaten werden von den zuständigen kantonalen Stellen erhoben und plausibilisiert. Nach ihrer Bearbeitung im BFS werden sie in einer zentralen Datenbank zusammengefasst.

Laut Bundesstatistikgesetz vom 9. Oktober 1992 dürfen die zu statistischen Zwecken erhobenen Daten nicht zu anderen Zwecken verwendet werden, ausser wenn ein Bundesgesetz eine andere Verwendung ausdrücklich anordnet (Art. 14 Abs. 1). Im Falle der Medizinischen Statistik sind die an der Erhebung beteiligten Kreise, insbesondere die zuständigen kantonalen Stellen, berechtigt, die Rohdaten nach eigen Bedürfnissen auszuwerten.

5.1 Zuständige kantonale Stellen

Die zuständigen kantonalen Stellen erhalten die Gesamtheit der Daten (inkl. anonyme Verbindungskodes), welche von den auf der kantonalen Spitalliste eingetragenen Betrieben geliefert werden, ebenso die ausserkantonal erfassten Fälle von Einwohnern ihres Kantons.

Die kantonalen Stellen unterstehen denselben Datenschutzbestimmungen wie das BFS, sofern die kantonale Gesetzgebung keine abweichende Regelung trifft.

5.2 Bearbeiten für Forschung

Personendaten können Dritten für Forschungszwecke zugänglich gemacht werden. In der Regel handelt es sich um Forschungsinstitute der Hochschulen. Dabei müssen verschiedene Kriterien und Anforderungen erfüllt sein:

- es muss ein begründetes und gerechtfertigtes Gesuch des Organs vorliegen, das die Forschung durchführen will;
- es wird nur derjenige Teil der Datenbank zugänglich gemacht, der für das Forschungsprojekt benötigt wird;

4.2 Offices cantonaux - OFS

En principe, les données transitent des offices cantonaux à l'OFS sur support télématique. Il ne peut être exclu que, dans un premier temps, les données soient acheminées par la poste.

Dans le second cas, l'inscription des colis contenant les disquettes à l'office postal responsable du transport nous semble suffisant. Un processus de cryptage des données mises sur les disquettes peut toutefois être envisagé.

En cas de transmission par un processus télématique, l'intégralité du fichier est transmis sous forme cryptée. La confidentialité peut donc ainsi être assurée durant le transport.

5. Mise à disposition des données

Les données brutes sont récoltées et plausibilisées par les offices cantonaux compétents. Après traitement au sein de l'OFS, elles sont réunies dans une base de données centrale.

Selon la loi du 9 octobre 1992 sur la statistique fédérale, les données collectées à des fins statistiques ne peuvent être utilisées à d'autres fins, à moins qu'une loi fédérale n'autorise expressément une autre utilisation (art. 14. al. 1). Dans le cas de la statistique médicale, les milieux participant à l'enquête, notamment les offices cantonaux compétents, sont en droit d'exploiter les données brutes pour leurs besoins propres.

5.1 Offices cantonaux compétents

Les offices cantonaux compétents reçoivent l'intégralité (y c. le code de liaison anonyme) des données fournies par les établissements inscrits sur la liste du canton, ainsi que les enregistrements de leurs ressortissants hospitalisés dans d'autres cantons.

Les offices cantonaux sont soumis aux mêmes devoirs de protection des données que l'OFS à moins qu'une législation cantonale particulière ne leur permette d'y déroger.

5.2 Traitements à des fins de recherche

Des données individuelles peuvent être transmises à des tiers à des fins de recherche. Il s'agit ici principalement d'instituts de recherche universitaire. Plusieurs critères et contraintes doivent être respectés:

- la demande doit être motivée et justifiée par l'organe qui désire effectuer une recherche,
- seule la fraction de la base de donnée nécessaire au projet de recherche peut être transmise,

- besonders schützenswerte Variablen sind zu anonymisieren;
- der Empfänger verpflichtet sich, Personendaten nicht an Dritte weiterzugeben;
- sobald die Bearbeitung abgeschlossen ist, sind die Daten dem BFS zurückzugeben oder zu vernichten.

Das BFS fertigt für jeden Benutzer einen Vertrag aus, der die genauen Bedingungen regelt. Die Benutzer der Daten unterliegen derselben Datenschutzpflicht wie das BFS.

5.3 Übrige Benutzer

Die übrigen Benutzer, die keinen gerechtfertigten Bedarf gemäss Bundesstatistikgesetz und Datenschutzgesetz nachweisen können, dürfen in keinem Fall Zugang zu Personendaten erhalten. Nur Synthesetabellen, welche Daten in aggregierter Form enthalten, werden allgemein zugänglich gemacht oder veröffentlicht.

6. Aufbewahrung der Daten

Die Datenbank der Medizinischen Statistik der Krankenhäuser wird eine ausserordentlich ergiebige Informationsquelle für epidemiologische Studien sein. In dieser Beziehung lässt sie sich vergleichen mit der Todesursachenstatistik, die seit über hundert Jahren ohne Unterbruch geführt wird und die ein nationales Erbe von unschätzbarem Wert darstellt, für die Public Health und für die medizinische Forschung.

Wegen des enormen Interesses, das eine Vollerhebung der Hospitalisierungsfälle über einen Zeitraum von vielen Jahren aufweisen kann, sollen die medizinischen Daten zeitlich unbeschränkt aufbewahrt werden.

Andererseits ist festzuhalten, dass heute nahezu alle Geburten und ein Grossteil der Todesfälle im Spitalmilieu stattfinden und die Hospitalisationshäufigkeit in der Bevölkerung hoch ist (eine Spitalbehandlung auf 7 Personen pro Jahr). In Anbetracht dieser Tatsachen würde ein Verbindungskode die Voraussetzung schaffen, um alle Spitalbehandlungen einer Person während ihres ganzen Lebens verfolgen zu können. Die Angaben über die verschiedenen Hospitalisationsepisoden einer Person sind besonders schützenswerte Daten, die sich identifizierend auswirken können. Unserer Meinung nach kann das epidemiologische Interesse eine solche Beeinträchtigung des Schutzes der Privatsphäre nicht rechtfertigen.

Eine systematische, unbefristete Aufbewahrung der anonymen Verbindungskodes ist nicht gerechtfertigt. **Die Kodes sind deshalb nach 10 Jahren zu vernichten.**

- des mesures d'anonymisation des variables sensibles seront appliquées,
- le destinataire s'engage à ne pas communiquer les données individuelles à des tiers,
- les données doivent être rendues à l'OFS ou détruites à la fin du traitement.

L'OFS établit, avec chaque utilisateur, un contrat dans lequel les conditions sont spécifiées. Les utilisateurs de données sont soumis aux mêmes devoirs de protection des données que l'OFS.

5.3 Autres utilisateurs

Les autres utilisateurs qui ne peuvent justifier un besoin conforme à la loi sur la statistique fédérale et la loi sur la protection des données ne peuvent en aucun cas avoir accès à des données individuelles. Seuls des tableaux de synthèse comprenant des données agrégées sont accessibles librement ou publiés.

6. Conservation des données

La base de données de la statistique médicale des hôpitaux constituera une source d'information extraordinairement riche pour des études épidémiologiques. Dans une situation comparable, la statistique de la mortalité, qui a été conduite sans discontinuité pendant plus de cent ans, représente une valeur inestimable du patrimoine national, pour la santé publique et pour la recherche médicale.

Les données médicales seront conservées sans limite de temps en raison de l'énorme intérêt que peut représenter le relevé exhaustif des cas d'hospitalisation sur de nombreuses années.

D'un autre point de vue, la quasi-totalité des naissances et une fraction importante des décès ont lieu en institution hospitalière et le taux d'hospitalisation dans la population est élevé (un cas d'hospitalisation pour 7 personnes chaque année). Compte tenu de cette réalité, un code de liaison permettrait de suivre l'historique hospitalier d'un individu tout au long de sa vie. De plus, la connaissance des divers épisodes hospitaliers d'une personne est une donnée sensible qui peut se révéler identifiante. A notre avis, l'intérêt épidémiologique ne peut justifier cet affaiblissement de la protection de la sphère privée.

La conservation systématique des codes de liaison anonymes ne se justifie pas sur le long terme. **Les codes devront être détruits après 10 ans**.

Epidemiologische Überlegungen bewegen uns aber dazu, nicht alle Kodes blindlings zu vernichten. Während diese Frage für Personen im erwerbsfähigen Alter besonders heikel ist, ist sie es für den Rest der Bevölkerung deutlich weniger. In diesem Fall scheint es uns unverhältnismässig, diese für die epidemiologische Forschung wichtigen Daten zu verlieren.

Wir schlagen deshalb vor, die Aufbewahrungszeit der Verbindungskodes folgendermassen zu begrenzen:

- der Kode wird für Kinder unter 15 Jahren systematisch aufbewahrt (grosses epidemiologisches Interesse und reduziertes Risiko bei Kindern);
- der Kode wird für Personen im Alter von über 64 Jahren systematisch aufbewahrt (grosses epidemiologisches Interesse und reduziertes Risiko bei Personen im AHV-Alter);
- der Kode wird für eine Stichprobe von Personen im Alter von 15-64 Jahren systematisch aufbewahrt (epidemiologisches Interesse an einer längerfristigen Beobachtung). Grösse und Zusammensetzung dieser Stichprobe werden aufgrund der statistischen Bedürfnisse am Ende der ersten sieben Jahre der obligatorischen Erhebung festgelegt;
- der Kode wird für Personen im Erwerbsalter 10 Jahre aufbewahrt.

Um zu verhindern, dass die Verbindungskodes in den Datenbanken ausserhalb des Einflussbereichs des BFS (Kantone, Forschungsinstitute) unrechtmässig aufbewahrt werden, wird der Schlüssel für die definitive Verschlüsselung jedes Jahr geändert. Dabei wird für sämtliche Registrierungen der Verbindungskodes entschlüsselt und mit dem geänderten Schlüssel neu codiert.

7. MitarbeiterInnen und Amtsgeheimnis

Grundsätzlich werden nur diejenigen Mitarbeiter der Sektion Gesundheit des BFS Zugang zu den Datenbanken haben, die direkt mit dem Projekt der Medizinischen Statistik der Spitäler betraut sind. Technische Mass-nahmen sind vorgesehen (Passwörter, Identifizierung).

Laut Artikel 14 des Bundesstatistikgesetzes müssen die mit statistischen Arbeiten betrauten Personen alle Daten über einzelne natürliche und juristische Personen geheimhalten, die sie bei ihrer Arbeit wahrgenommen haben. Analog dazu gilt diese Pflicht auch für die Personen, die gelegentlich oder regelmässig am Datenfluss im Zusammenhang mit der Medizinischen Statistik beteiligt sind (Forscher, Mitarbeiter der Spitäler, der zuständigen kantonalen Stellen oder der beauftragten privaten Institutionen).

Toutefois, l'intérêt épidémiologique nous incite à ne pas détruire aveuglément tous les codes. Si la question est particulièrement sensible pour les personnes en âge d'exercer une activité, elle l'est beaucoup moins pour le reste de la population. Dans ce cas, il nous paraît disproportionné de perdre ces données très importantes pour la recherche épidémiologique.

Nous proposons donc de limiter la durée de conservation des codes de liaison selon cette clé:

- le code est systématiquement conservé pour les enfants de moins de 15 ans révolus (intérêt épidémiologique élevé et sensibilité moindre chez les enfants),
- le code est systématiquement conservé pour les personnes de plus de 64 ans (intérêt épidémiologique élevé et sensibilité moindre chez les personnes en âge AVS),
- le code est systématiquement conservé pour un sous-collectif de la population des personnes de 15 à 64 ans (intérêt épidémiologique pour un suivi à long terme). La taille et la composition de ce sous-collectif seront déterminées selon les besoins statistiques à la fin des sept premières années du relevé obligatoire,
- le code est conservé durant 10 ans pour les personnes en âge actif.

Pour éviter que les codes de liaison ne soient conservés de manière illicite dans des bases de données qui échappent au contrôle de l'OFS (cantons, instituts de recherche), la clé utilisée lors du cryptage final sera changée chaque année. Tous les enregistrements dont le code de liaison pourra être conservé seront alors décodés et recodés avec la nouvelle clé.

7. Collaborateurs et secret de fonction

D'une manière générale, seuls les collaborateurs de la Section de la santé de l'OFS directement impliqués dans le projet de statistique médicale des hôpitaux auront accès aux bases de données. Des mesures techniques seront mises en place (mots de passe, identifications).

Selon l'art. 14 de la loi sur la statistique fédérale, les collaborateurs de l'OFS chargés de travaux statistiques sont tenus de garder le secret sur les données concernant des personnes physiques ou morales dont ils ont eu connaissance dans l'exercice de leur fonction. Par analogie, les personnes participant occasionnellement ou régulièrement au flux d'information des données de la statistique médicale (collaborateurs des hôpitaux, des offices cantonaux désignés ou des associations mandatées, chercheurs) sont également soumises à cette obligation.

Der Datenschutz in der Medizinischen Statistik

Anhang I

Beschreibung der vorgesehenen Methoden zur Gewährleistung des Persönlichkeitsschutzes

La protection des données dans la statistique médicale

Annexe I

Description des méthodes envisagées pour protéger la confidentialité des personnes

Statistiques des établissements de santé
(soins intra-muros)

Description des méthodes envisagées pour protéger la confidentialité des données personnelles

David-Olivier JAQUET-CHIFFELLE
Dr. sc. math.

Gr AaC
Section fédérale de Cryptologie
3003 Berne

Table des matières

1 Introduction **4**

1.1 Description du problème 4

1.1.1 Point de vue des hôpitaux 4

1.1.2 Point de vue des cantons 4

1.1.3 Point de vue de l'Office fédéral de la Statistique 5

1.2 Transformations cryptographiques à disposition 5

1.2.1 Fonctions de hachage à sens unique 5

1.2.2 Algorithmes de cryptage 6

2 Choix d'un protocole **8**

2.1 Généralités 8

2.2 Les données identifiantes minimales 8

2.2.1 Proposition pour le calcul des données identifiantes minimales 9

2.3 Opérations cryptographiques effectuées sur les données identifiantes minimales 10

2.3.1 Transformation T_1 10

2.3.2 Transformation T_2 12

2.4 Protection des clés secrètes 14

2.4.1 Une personne de confiance 15

2.4.2 Partage des compétences 16

2.4.3 Simplicité/Sécurité 16

3 Réalisation du protocole 17

3.1 Opérations cryptographiques effectuées par les hôpitaux 17

3.1.1 T_1, hachage à sens unique : SHA + compression 17

3.1.2 Cryptage de l'empreinte : algorithme IDEA 18

3.1.3 Transmission de la clé secrète c : algorithme RSA 19

3.2 Opérations cryptographiques effectuées par les cantons 20

3.3 Opérations cryptographiques effectuées au sein de l'Office fédéral de la Statistique 21

3.3.1 Décryptage des codes de liaison 21

3.3.2 T_2, cryptage uniforme des empreintes : algorithme IDEA . . . 21

3.4 Protocole de partage des compétences 22

3.4.1 Comment réaliser le partage des compétences? 22

3.4.2 A propos de la sécurité de ce protocole 24

3.5 Format du fichier à transférer 25

3.6 Mesures de sécurité recommandées 25

3.6.1 Authentification de la clé publique de l'Office fédéral de la Statistique 25

3.6.2 Protection des mots-de-passe 25

3.6.3 Ordinateur de l'Office fédéral de la Statistique 26

3.7 Droits d'auteur 26

4 Conclusion 29

Chapitre 1

Introduction

1.1 Description du problème

1.1.1 Point de vue des hôpitaux

Les hôpitaux doivent transmettre à l'Office fédéral de la Statistique des données relatives aux patients traités dans leurs établissements. Les données transitent par les organismes cantonaux chargés de les récolter.

Les données originales contiennent des informations sensibles (nom, prénom, date de naissance et domicile des patients, par exemple) qui caractérisent un individu dans la population et permettent de le reconnaître. Les données transmises, par contre, doivent garantir l'anonymat des personnes.

Les données épidémiologiques peuvent quant à elles être transmises en clair si l'anonymat des patients est garanti.

1.1.2 Point de vue des cantons

Les cantons recueillent les données provenant des hôpitaux, effectuent certaines vérifications de base puis transmettent le tout à l'Office fédéral de la Statistique.

Pour effectuer les vérifications de base, les cantons doivent avoir accès aux données épidémiologiques. Comme il est dit dans la section précédente, cela ne pose aucun problème tant que l'anonymat des patients est garanti.

1.1.3 Point de vue de l'Office fédéral de la Statistique

L'Office fédéral de la Statistique regroupe les données fournies par les cantons. Il doit pouvoir reconnaître qu'un même patient a suivi plusieurs traitements (dans le même hôpital ou dans des hôpitaux différents, dans le même canton ou dans des cantons différents, l'année en cours ou précédemment.)

Même si l'Office fédéral de la Statistique doit être en mesure d'effectuer un suivi des individus, il est important de maintenir, ici aussi, l'anonymat des personnes.

1.2 Transformations cryptographiques à disposition

Nous allons considérer dans ce rapport différents types de transformations cryptographiques.

1.2.1 Fonctions de hachage à sens unique

Une fonction de hachage à sens unique [1] tranforme un message de longueur quelconque en une empreinte de longueur constante. Ici le message (et l'empreinte d'ailleurs) est considéré comme une suite de caractères hexadécimaux, par exemple

$$0345A19FB3D5 \ldots 7AC36D18B3$$

Une bonne fonction de hachage à sens unique doit satisfaire les propriétés suivantes :

- Etant donné un message, son empreinte est facile à calculer.
- Par contre, étant donné une empreinte, il est virtuellement impossible de trouver un message correspondant (transformation à sens unique.)
- La probabilité que deux messages différents aient la même empreinte est négligeable (collisions.)
- La longueur de l'empreinte (nombre de bits) est constante.
- L'algorithme est public.

[1] Cf [Sch, Wel]

On distingue deux types de fonctions de hachage à sens unique :

(1) Sans clé secrète.

Rien n'est secret dans la transformation. Et pourtant, bien que tout soit connu, il est virtuellement impossible de trouver un message ayant une empreinte donnée. Exception : si l'ensemble de tous les messages possibles est trop petit, une attaque directe, par dictionnaire, est réalisable.

Exemple : SHA.

(2) Avec clé secrète.

L'algorithme dépend d'une clé supposée secrète. Tant que la clé n'est pas corrompue, une attaque par dictionnaire est impraticable.

Lorsque la clé est corrompue, il arrive que la transformation soit localement inversible.

Exemple : RIPE-MAC.

Une fonction de hachage à sens unique dont la clé est corrompue est souvent moins sûre qu'une fonction de hachage à sens unique sans clé.

La composition de plusieurs fonctions de hachage à sens unique redonne une fonction de hachage à sens unique et n'apporte par conséquent aucune sécurité supplémentaire.

1.2.2 Algorithmes de cryptage

On distingue deux types d'algorithmes de cryptage :

(1) Avec *clé secrète.*

La sécurité de l'algorithme dépend d'une clé secrète qu'il faut connaître tant pour crypter que pour décrypter les messages. En général, ces algorithmes sont très rapides ; on les utilise pour crypter un nombre important de données. Les clés sont faciles à générer (toute clé ayant la bonne longueur fait l'affaire.)

Exemples : IDEA, DES, etc.[2]

(2) Avec une paire *clé publique/clé privée.*

[2] Cf. [L&M, Sch, JaC]

Comme son nom l'indique, la clé publique (celle qui est utilisée pour crypter un message) n'a pas besoin d'être secrète. Ces algorithmes sont dits asymétriques car la connaissance de la clé publique ne permet pas de décrypter les messages (pas même ceux que l'on a cryptés soi-même.) Pour décrypter un message, il faut connaître la clé privée, connue du destinataire uniquement. Ces algorithmes sont en général plus lents que les précédents ; on les utilise souvent pour transmettre la valeur d'une clé secrète. La génération d'une paire *clé publique/clé privée* est un travail beaucoup plus délicat que la génération d'une clé secrète.

Exemple : RSA.[3]

Si l'on désire atteindre un niveau de sécurité sérieux, on estime généralement que la sécurité d'un protocole ne doit pas dépendre de la confidentialité des algorithmes utilisés.

[3] Cf. [RSA1, RSA2, Kob, Sch, Wel]

Chapitre 2

Choix d'un protocole

2.1 Généralités

La solution proposée doit être aussi simple que possible tout en respectant les critères de sécurité relatifs à l'anonymat des patients. Les données épidémiologiques pouvant être transmises en clair, nous nous concentrerons dans ce qui suit uniquement sur les données qui permettraient d'identifier les personnes. Ces données identifiantes seront transformées en des codes de liaison garantissant l'anonymat des patients.

Les paires {*données identifiantes ; données épidémiologiques*} dont disposent les hôpitaux seront donc transformées en des paires {*code de liaison ; données épidémiologiques*} avant toute transmission.

2.2 Les données identifiantes minimales

Tout d'abord l'Office fédéral de la Statistique doit décider d'un protocole permettant aux hôpitaux d'extraire les données identifiantes minimales à partir des données permettant de reconnaître un individu.

Les données identifiantes minimales satisfont les propriétés suivantes :

- Pour un patient, les données identifiantes minimales sont pour ainsi dire indépendantes de l'hôpital, du lieu ou du temps.[1] Plus précisément, la probabilité

[1] Afin de minimiser l'impact des fautes d'orthographe, on peut envisager que des orthographes différentes du nom (Muller, Müller, Mueller, par exemple) produisent les mêmes données identifi-

qu'un même patient produise deux données identifiantes minimales différentes pendant t années doit être inférieure à un seuil δ que l'Office fédéral de la Statistique déterminera en fonction des applications qui sont envisagées.

- La probabilité que deux personnes différentes aient les mêmes données identifiantes minimales doit être inférieure à un seuil ϵ que l'Office fédéral de la Statistique déterminera en fonction des applications qui sont envisagées.

2.2.1 Proposition pour le calcul des données identifiantes minimales

Les données identifiantes contiennent par exemple :

- la date de naissance (JJMMAAAA),
- le sexe,
- le nom,
- le prénom.

Les chaînes de caractères sont transformées à l'aide d'une fonction réduisant de façon significative l'influence des orthographes multiples ou des erreurs d'orthographe liées à la saisie des données par les hôpitaux. Cette fonction doit cependant conserver dans une large mesure la diversité des noms et des prénoms. L'idée sous-jacente est de définir une écriture robuste.

Cette fonction peut :

- ne pas différencier les minuscules et les majuscules ;
- éliminer les accents, les espaces, les traits d'union ;
- remplacer tous les y par des i ;
- comprimer les doublons :

  ff devient f,
  ss devient s, etc.

- remplacer sch ou sh, par ch ;

antes minimales.

- remplacer

  ae par *a*,

  ou ou *ue*, par *u* ;

- éliminer les *h* sauf s'ils sont précédés de *c* ;

- remplacer *cqu* par *qu* ;

  etc.

L'ordre dans lequel les transformations sont effectuées doit être précisé.

Les données identifiantes minimales se présentent maintenant sous la forme d'une chaîne alphanumérique :

$$JJMMAAAASN_1N_2 \ldots N_\alpha P_1 P_2 \ldots P_\beta$$

où les N_i et les P_j représentent respectivement le nom (longueur α) et le prénom (longueur β) après application de la fonction décrite ci-dessus. Les valeurs α et β sont variables.

Cette chaîne alphanumérique est convertie en une suite de valeurs hexadécimales.

Le choix définitif des données intervenant dans ce calcul ainsi que celui de la procédure utilisée incombe à l'Office fédéral de la Statistique. Nous supposerons dans ce qui suit être en possession d'une procédure simple permettant de calculer, sous forme hexadécimale, les données identifiantes minimales de tout individu.

2.3 Opérations cryptographiques effectuées sur les données identifiantes minimales

Les données identifiantes minimales sont transformées successivement[2] par T_1 (transformation effectuée par les hôpitaux) et T_2 (transformation effectuée au sein de l'Office fédéral de la Statistique.)

2.3.1 Transformation T_1

La transformation T_1 doit être la même pour tous les hôpitaux car l'empreinte d'un individu doit être indépendante du temps et du lieu où ce patient a été traité.

[2] Cf. tableau 3.1, p. 28

Nous choisissons pour T_1 une fonction de hachage à sens unique sans clé. La sécurité n'est donc pas liée à un secret mais à la complexité intrinsèque de cette fonction. D'une part, cela simplifie les protocoles en évitant de devoir distribuer une même clé secrète à tous les hôpitaux et, d'autre part, il serait risqué (et naïf!) de supposer que cette clé, connue de plusieurs centaines d'hôpitaux, puisse véritablement rester secrète...

L'algorithme de hachage est utilisé pour transformer les données identifiantes minimales afin de cacher l'identité des individus. Cela évite que la base de données devienne petit à petit un registre de population. Cette transformation s'impose si l'on veut respecter la loi.

Si les données identifiantes minimales sont les mêmes, les empreintes sont aussi identiques. Par conséquent, il suffit à l'Office fédéral de la Statistique de connaître les empreintes associées aux enregistrements pour pouvoir faire un suivi individuel. L'Office fédéral de la Statistique peut en effet reconnaître que deux hospitalisation concernent le même individu mais ne sait pas qui est cet individu.

Attaques possibles contre les empreintes

La transformation liée à la fonction de hachage à sens unique n'est cependant pas suffisante pour garantir l'anonymat des patients. En effet, l'algorithme utilisé est public. Par conséquent, si quelqu'un ou un organisme dispose d'un registre (même partiel) de population, il peut, en appliquant cet algorithme, construire un dictionnaire

$$\text{“valeurs identifiantes”} \longleftrightarrow \text{“empreintes”}.$$

En particulier, il peut tester l'appartenance de tout individu à la base de données (attaque ponctuelle.)

De plus, d'un point de vue combinatoire, le nombre de données identifiantes minimales distinctes s'élève au plus à :

- la date de naissance : $365 * 122$ possibilités[3]
- le sexe : 2 possibilités
- le nom : au plus 7000 possibilités

[3]Il ne faut pas oublier Jeanne Calment!

- le prénom : au plus 7000 possibilités

$365 * 122 * 2 * 7000 * 7000 = 4'363'940'000'000 < 2^{42}$ différentes possibilités.

Bien que nécessitant certains moyens de calcul, une attaque par dictionnaire avec recherche exhaustive est envisageable.

Certains pays (USA, France, par exemple) limitent considérablement l'utilisation de *bons* algorithmes de cryptage pour les applications civiles. Ce n'est pas le cas en Suisse et cela nous permet de développer un système qui résiste à ces attaques.

Conséquences :

1) Il est nécessaire de protéger les empreintes durant leur transmission vers l'Office fédéral de la Statistique. On utilisera un algorithme de cryptage à clé secrète c, où la clé c, modifiée lors de chaque transmission, sera transmise simultanément à l'aide d'un algorithme de cryptage à clé publique.[4]

   L'empreinte cryptée sera le code de liaison transmis par les hôpitaux.

2) La transformation T_2 doit dépendre d'une clé secrète K.

2.3.2 Transformation T_2

T_2 transforme[5] l'empreinte produite par T_1 en un code de liaison uniforme. La longueur de l'empreinte est la même que celle du code de liaison uniforme.

Deux choix se présentent pour T_2 :

(i) fonction de hachage à sens unique avec clé secrète K ;

(ii) algorithme de cryptage à clé secrète K.

Comparaison du niveau de sécurité pour chacun des deux choix

Dans le contexte qui nous intéresse, ces deux choix définissent un même niveau de sécurité.

[4] Cf. tableau 3.1, p. 28
[5] Cf. tableau 3.1, p. 28

En effet, une personne qui ne connaît pas la clé secrète K est virtuellement dans l'impossibilité d'effectuer la transformation T_2, dans un sens comme dans l'autre, aussi bien s'il s'agit d'une fonction de hachage que d'un algorithme de cryptage.

Si la clé est corrompue, le niveau de sécurité est considérablement réduit. Toutefois, dans ce cas de figure aussi, le choix pour T_2 d'une fonction de hachage n'apporte aucune sécurité supplémentaire.

Certains objecteront peut-être que, lorsque la clé K est connue, l'algorithme de cryptage peut être inversé contrairement à la fonction de hachage. D'une part cela n'est pas toujours vrai [6] et, d'autre part, même lorsque cela s'avère exact, la sécurité de notre protocole reste la même : si la clé K est corrompue, la composition de T_1 et de T_2 est équivalente à une fonction de hachage à sens unique sans clé, donc équivalente à T_1.

Dans les deux cas, si la clé K est corrompue, les attaques possibles sont celles déjà décrites contre T_1, à savoir : attaque ponctuelle et attaque par dictionnaire.

Quel est le meilleur choix pour T_2?

Quelques remarques :

1) Si T_2 est un algorithme de cryptage, il est possible de modifier régulièrement la clé K.

2) Une transformation considérée comme sûre à une certaine époque (par exemple en 1997, dans l'état actuel des connaissances) peut ne plus être suffisamment performante dix ou vingt ans plus tard. Si, suite à certaines découvertes, l'algorithme de cryptage choisi n'est plus assez sûr, il est possible d'en choisir un autre sans perdre la base de données. Avec une fonction de hachage, il faudrait par contre appliquer à toute la base de données une transformation T_3 supplémentaire ; puis, si T_3 révèle une faiblesse, encore une autre transformation T_4 etc...

3) Imaginons maintenant qu'une étude statistique mette en évidence une catégorie de patients qui, ayant subi un certain traitement, ont une espérance de vie considérablement réduite s'ils ne sont pas suivis régulièrement. Dans ce contexte, le fait de pouvoir communiquer aux hôpitaux les empreintes des patients qu'ils ont traités et qui appartiennent à cette catégorie peut sauver des

[6] La plupart des fonctions standards de hachage à sens unique avec clé sont basées sur des algorithmes de cryptage et deviennent localement inversibles si la clé est corrompue.

vies. Notons bien que seule une empreinte est transmissible puisque l'identité des patients n'est pas connue de l'Office fédéral de la Statistique. La tâche incombe ensuite aux hôpitaux de retrouver, parmi leurs anciens patients, ceux qui correspondent à ces empreintes.

Quel que soit le choix pour T_2 (fonction de hachage ou algorithme de cryptage), ce retour d'information à l'intention des hôpitaux est possible.

Toutefois, si T_2 est une fonction de hachage à sens unique avec clé, il faut transmettre aux hôpitaux non seulement les codes de liaison uniformes des personnes concernées, mais aussi l'algorithme T_2 et surtout la valeur de la clé K. Cela menace la confidentialité de la base de donnée. Pour jouir à nouveau de la sécurité initiale, toute la base de données devra subir une transformation supplémentaire à l'aide d'une fonction de hachage T_3 à sens unique avec une nouvelle clé secrète K_3, et ainsi de suite lors de chaque retour d'empreinte vers les hôpitaux. Après quelques années, on risque de devoir appliquer un nombre considérable de transformations aux nouvelles empreintes reçues des hôpitaux pour obtenir les codes de liaisons uniformes correspondants. Indépendamment de l'aspect peu pratique (!) qui complique sans raison la réalisation du protocole, cette approche n'est pas plus satisfaisante d'un point de vue théorique puisqu'elle augmente inutilement le risque de collisions au niveau des codes de liaison uniformes.

Par contre, si T_2 est un algorithme de cryptage, il est possible de renvoyer aux hôpitaux directement l'empreinte qu'ils avaient eux-mêmes calculée. On peut conserver T_2 et il n'est pas nécessaire de corrompre la clé K. A aucun moment, la base de données n'est menacée.

Le choix pour T_2 d'un algorithme de cryptage (au lieu d'une fonction de hachage à sens unique avec clé) s'impose donc naturellement, aussi bien d'un point de vue pratique que théorique.

Puisqu'en Suisse (contrairement à d'autres pays), nous pouvons légalement utiliser pour T_2 aussi bien un algorithme de cryptage qu'une fonction de hachage, ne nous privons pas de cette liberté et choisissons objectivement la solution la plus appropriée à notre problème.

2.4 Protection des clés secrètes

Il y a essentiellement deux clés secrètes sensibles :

- D, la clé privée de l'Office fédéral de la Statistique, qui permet de retrouver les clés secrètes c générées par les hôpitaux pour transformer les empreintes en code de liaison,
- K, la clé secrète de T_2, qui permet de transformer les empreintes en codes de liaison uniformes.

Nous nous trouvons, ici, devant un dilemme.

D'une part, il faut que l'Office fédéral de la Statistique puisse retrouver les empreintes d'origine (non cryptées) liées à chaque enregistrement afin de reconnaître qu'un même patient a suivi plusieurs traitements (dans le même hôpital ou dans des hôpitaux différents, dans le même canton ou dans des cantons différents, l'année en cours ou précédemment.) Ceci est indispensable si l'on veut pouvoir reconnaître les hospitalisations multiples et effectuer un suivi au cours du temps.

D'autre part, si l'Office fédéral de la Statistique est en possession des empreintes, bien que cela ne constitue pas un registre de population, l'anonymat des patients n'est pas garanti comme nous l'avons vu à la page 11.

2.4.1 Une personne de confiance

Si toutes les parties concernées acceptent de faire entièrement confiance à une certaine personne $\mathcal{P}$, il suffit de donner les deux clés sensibles uniquement à $\mathcal{P}$.

Seule cette personne $\mathcal{P}$ peut alors autoriser le décryptage des codes de liaison (transmis par les hôpitaux, via les cantons) puis le cryptage des empreintes ainsi obtenues en employant, cette fois-ci, la clé secrète K, toujours la même, connue d'elle-seule.

Les codes de liaison décryptés-recryptés deviennent les codes de liaison uniformes retenus par l'Office fédéral de la Statistique. Une même personne a toujours le même code de liaison uniforme ; le suivi est donc possible. Le code de liaison uniforme garantit l'anonymat des patients puisque seule $\mathcal{P}$ est capable de relier le code d'identification uniforme et l'empreinte d'origine.

La sécurité du protocole repose essentiellement sur l'honnêteté de la personne $\mathcal{P}$.

2.4.2 Partage des compétences

Si certaines parties refusent de faire entièrement confiance à une personne unique $\mathcal{P}$, on peut faire appel à une procédure de partage de secret.[7]

La clé privée D, permettant de retrouver les clés secrètes des hôpitaux, et K, la clé secrète utilisée pour recrypter uniformément les empreintes, forment un secret qui peut être partagé entre plusieurs entités indépendantes. Le même secret partagé doit être réutilisable année après année afin que les nouvelles données puissent être comparées aux anciennes. Plus qu'un partage de secret, il s'agit bien ici d'un partage de compétences.

Pratiquement, on peut imaginer le protocole suivant : pour transformer les empreintes cryptées par les hôpitaux en codes de liaison uniformes, n personnes doivent se rencontrer, une fois par année, et entrer à tour de rôle un mot-de-passe. Si les mots-de-passe sont corrects, l'ordinateur effectue la transformation

"empreintes cryptées (par les hôpitaux)" $\longrightarrow$ "codes de liaison uniformes".

Aussi longtemps qu'au moins une personne (parmi les n personnes) est honnête et que cette personne honnête est seule à connaître son propre secret, l'anonymat des patients est garanti.

2.4.3 Simplicité/Sécurité

Si l'on désire favoriser la simplicité des opérations effectuées dans le cadre de l'Office fédéral de la Statistique, il faut éviter la procédure de partage de secret. Cependant, d'un point de vue cryptographique, il est évidemment plus sûr de partager un secret entre plusieurs individus plutôt que de le confier à une unique personne $\mathcal{P}$.

Vu le degré de sensibilité des données considérées, je recommande d'introduire dans le protocole un partage du secret entre trois personnes de confiance indépendantes.

[7] Cf [Sch, Sim]

Chapitre 3

Réalisation du protocole

Les opérations cryptographiques effectuées par les hôpitaux sont représentées dans la partie supérieure du tableau 3.1 à la page 28.

3.1 Opérations cryptographiques effectuées par les hôpitaux

3.1.1 T_1, hachage à sens unique : SHA + compression

Pour satisfaire les exigences de sécurité que l'on rencontre ici, je propose de choisir une fonction de hachage qui produit des empreintes de 64 bits. C'est un bon compromis entre le niveau de sécurité souhaité et la place mémoire disponible.

Considérations techniques

Comme algorithme de hachage, on retiendra une variante de SHA (**S**ecure **H**ash **A**lgorithm),[1] algorithme développé conjointement par le NIST (**N**ational **I**nstitute of **S**tandards and **T**echnology) et la NSA (**N**ational **S**ecurity **A**gency.) En version standard, le SHA produit des empreintes de 160 bits. On combinera donc SHA avec une fonction de compression.

La fonction de compression transforme une empreinte de 160 bits (40 caractères hexadécimaux) en une empreinte de 64 bits (16 caractères hexadécimaux.)

[1] Cf [SHA1, SHA2, Sch]

Les 40 caractères hexadécimaux sont regroupés en 10 blocs de 4 caractères hexadécimaux chacuns. Chaque bloc est un mot de 16 bits. On numérote ces blocs de 0 à 9. Le "XOR", c'est-à-dire le "ou exclusif", induit une opération $\oplus$ entre les blocs. Nous allons utiliser cette opération pour calculer 4 blocs de 4 caractères hexadécimaux qui vont définir l'empreinte de 64 bits.

Définition de l'empreinte de 64 bits	A partir de l'empreinte de 160 bits
Premier bloc :	bloc 0 $\oplus$ bloc 1 $\oplus$ bloc 2 $\oplus$ bloc 3 $\oplus$ bloc 4
Second bloc :	bloc 2 $\oplus$ bloc 3 $\oplus$ bloc 4 $\oplus$ bloc 5 $\oplus$ bloc 6
Troisième bloc :	bloc 4 $\oplus$ bloc 5 $\oplus$ bloc 6 $\oplus$ bloc 7 $\oplus$ bloc 8
Dernier bloc :	bloc 0 $\oplus$ bloc 6 $\oplus$ bloc 7 $\oplus$ bloc 8 $\oplus$ bloc 9

Pour la description exacte de SHA, on peut se référer à [Sch], pp. 351–355. A la fin de ce livre, on trouve également le listing de SHA implémenté dans le langage de programmation C (pp. 587–594.)

Une copie de ce listing est annexée à ce rapport.

3.1.2 Cryptage de l'empreinte : algorithme IDEA

Je propose ici l'emploi d'IDEA.[2] L'algorithme IDEA a été développé par X. Lai et J. L. Massey. Il dépend d'une clé secrète c de 128 bits et transforme tout bloc de 64 bits en un bloc de 64 bits, de façon bijective. Sans connaître la clé secrète c, il est virtuellement impossible de retrouver la valeur d'une empreinte cryptée à l'aide d'IDEA.

D'après Bruce Schneier (Cf [Sch], p. 276), IDEA est *le meilleur et le plus sûr des algorithmes disponibles publiquement à ce jour.*

Le code de liaison transmis sera la valeur de l'empreinte, cryptée à l'aide d'IDEA et de la clé secrète c.

Considérations techniques

On trouvera une description détaillée de l'algorithme IDEA dans [L&M]. A la fin du livre [Sch], on trouve également le listing d'IDEA implémenté dans le langage de programmation C (pp. 555–570.)

[2] Cf [L&M, **?**]

Une copie de ce listing est annexée à ce rapport.

La clé c est longue de 128 bits, c'est-à-dire 32 caractères hexadécimaux. Elle peut être

1) soit choisie librement par l'utilisateur de l'hôpital,

2) soit générée pseudo-aléatoirement par l'ordinateur de l'hôpital lors du transfert annuel des données.

La deuxième solution présente l'avantage de pouvoir être réalisée en arrière-plan. L'utilisateur hospitalier n'a aucune clé à entrer, ni à mémoriser. Pour lui le protocole s'en trouve ainsi simplifié au maximum. Je propose de retenir cette deuxième variante.

Pour générer en arrière-plan la clé secrète c, on créera un module mémorisant, par exemple, certains paramètres des 50 derniers événements : position et/ou vitesse de la souris, déroulement d'un menu, utilisation des touches du clavier, temps entre deux événements consécutifs, etc. On traduira les paramètres ainsi mémorisés en une suite de caractères hexadécimaux auxquels on appliquera la fonction de hachage SHA ; les 64 premiers bits de l'empreinte produite par ce hachage définiront la valeur de la clé c.

3.1.3 Transmission de la clé secrète c : algorithme RSA

Sans la clé secrète c, les données transmises sont inutilisables : il est impossible d'effectuer un suivi des individus (hospitalisations multiples, etc.)

Par conséquent, il s'agit de conserver l'information concernant c avec les empreintes cryptées. Evidemment, la clé c ne sera pas mémorisée en clair car, sinon, toute personne ayant accès au fichier pourrait décrypter les codes de liaison et le cryptage deviendrait inutile. Je propose d'employer le célèbre algorithme à clé publique RSA.[3] Avant toute transmission, les clés secrètes c seront cryptées à l'aide de RSA en employant la clé publique E de l'Office fédéral de la Statistique.

[3] Cf. [RSA1, RSA2, Kob, Sch, Wel]

Considérations techniques

Pour des raisons de sécurité, la valeur du module dans RSA devrait avoir 1024 bits.[4] La clé secrète que l'on veut crypter possède 128 bits. On peut profiter des 896 bits inutilisés pour transmettre, en parallèle, les résultats de certains tests ; ces tests permettent de vérifier que les routines employées par les hôpitaux (pour calculer les empreintes et ensuite les crypter) ont été programmées correctement. Cela présente également l'avantage de protéger RSA contre certains types d'attaques (lorsque la longueur du message à coder est trop petite par rapport à la taille du module) et augmente par conséquent de façon significative la sécurité du système.[5]

Les 896 bits disponibles sont remplis de la façon suivante :

- identité de l'hôpital
  (64 bits choisis par l'Office fédéral de la Statistique et communiqués à l'hôpital),
- les empreintes cryptées (avec la clé c) de 10 personnes fictives ($10 * 64 = 640$ bits), toujours les mêmes, choisies par l'OFS,
- 192 bits pseudo-aléatoires.[6]

Ce remplissage est effectué automatiquement, en arrière-plan.

3.2 Opérations cryptographiques effectuées par les cantons

Les cantons n'ont aucune opération cryptographique à effectuer.[7]

Les données qu'ils reçoivent ne leur permettent pas de retrouver les clés secrètes c employées par les hôpitaux. Les empreintes cryptées garantissent ainsi l'anonymat des patients vis-à-vis des cantons.

[4] Afin d'optimiser la sécurité de RSA, il faut choisir les clés de façon judicieuse. Le module doit, en autres, résister aux méthodes modernes de factorisation (Cf [Coh].) Ce travail d'experts peut être réalisé dans le cadre de notre section.

[5] Cf [Cop]

[6] Pour générer ces 192 bits pseudo-aléatoirement, on s'inspirera du programme qui génère la clé c.

[7] Cf tableau 3.1, p. 28

3.3 Opérations cryptographiques effectuées au sein de l'Office fédéral de la Statistique

Ces opérations cryptographiques ne doivent pas être nécessairement exécutées par l'Office fédéral de la Statistique lui-même. Elles auront toutefois lieu physiquement au sein de cet office afin d'éviter un transfert supplémentaire.

La partie inférieure du tableau 3.1 à la page 28 présente ces opérations.

3.3.1 Décryptage des codes de liaison

Pour décrypter les codes de liaison et retrouver les empreintes, il faut d'abord retrouver la clé secrète c utilisée par l'hôpital. Retrouver la clé c signifie décrypter un message à l'aide de RSA et de la clé privée D. Connaissant c, il est facile de retrouver les empreintes en décryptant les codes de liaison à l'aide d'IDEA et de c.

Considérations techniques

Le message contenant la clé secrète c, crypté à l'aide de RSA, contient 896 bits de contrôle (Cf p. 20.) Cette information peut être utilisée ici pour vérifier la validité des procédures cryptographiques utilisées par les hôpitaux et l'authenticité de la clé c.

3.3.2 T_2, cryptage uniforme des empreintes : algorithme IDEA

Le cryptage uniforme des empreintes obtenues en 3.3.1 se fait à l'aide d'IDEA et de la clé secrète K.[8]

Considérations techniques

Les valeurs des empreintes d'origine, transformées par T_2 en codes de liaison uniformes, ne sont jamais mémorisées sur aucun support.

[8] Si l'Office fédéral de la Statistique préfère utiliser un algorithme secret, il est possible de personnaliser IDEA (Cf [JaC].) Cette solution ajoute une sécurité supplémentaire tout en profitant des avantages intrinsèquement liés à un algorithme public aussi répandu que IDEA.

Les algorithmes utilisés par l'ordinateur ne sont pas secrets. Le fait de décompiler le programme ne permet pas d'effectuer les transformations décrites ci-dessus.

3.4 Protocole de partage des compétences

Les opérations cryptographiques effectuées au sein de l'Office fédéral de la Statistique exigent la connaissance des deux clés sensibles D et K. La clé K représente un *secret* qui, pour des raisons de sécurité, peut être partagé entre n personnes de confiance, indépendantes. Je conseille ici de choisir $n = 3$.

Si l'on accepte de faire confiance à une unique personne $\mathcal{P}$, on lira ce qui suit en supposant $n = 1$ (pas de partage de secret.)

Choix des personnes de confiance

Les trois personnes qui interviennent dans ce protocole peuvent être, par exemple, le Président de l'Association suisse des Médecins, le Directeur de l'Office fédéral de la Statistique et le Directeur de l'Office fédéral de la Protection des Données.

3.4.1 Comment réaliser le partage des compétences?

Dans ce qui suit, nous appelerons $\mathcal{P}_i$, $i = 1, 2, \ldots, n$, les n personnes de confiance qui interviennent dans le protocole.

Le mot-de-passe de chacune de ces personnes se compose de 32 caractères hexadécimaux. Nous appelerons K_i le mot-de-passe de $\mathcal{P}_i$.

Première utilisation

Lors de la toute première utilisation, chaque $\mathcal{P}_i$ choisit librement et indépendemment son propre mot-de-passe et l'introduit dans le système. Le responsable de la génération des clés de RSA introduit encore D, la clé privée utilisée pour décrypter RSA ; ensuite, il détruit les éventuelles copies de D dont il disposerait. Je rappelle que la clé privée D est composée d'environ 256 caractères hexadécimaux.

L'ordinateur définit $K = K_1 \oplus K_2 \oplus \ldots \oplus K_n$ où $\oplus$ représente l'opération "ou exclusif" appliquée aux entiers. La clé secrète uniforme K, de même que les mots-

de-passe K_i et la clé privée D ne sont jamais enregistrés sur le disque dur. Par contre, l'ordinateur calcule et mémorise les valeurs suivantes :

- $\text{IDEA}_K(D) = \Delta$, la valeur de D cryptée à l'aide d'IDEA et de la clé K ;
- $\text{SHA}(K_i) = \Sigma_i, i = 1, 2, \ldots, n$, les empreintes des mots-de-passe produites par l'algorithme SHA ;
- $\text{IDEA}_{K_i}(0) = \Gamma_i$.

Identification des personnes de confiance

Pour s'identifier, $\mathcal{P}_i$ introduit son mot-de-passe K_i dans l'ordinateur. L'identification est réussie si, et seulement si, $\text{SHA}(K_i) = \Sigma_i$.

Modification des mots-de-passe

Lorsque l'utilisateur $\mathcal{P}_i$ a été identifié avec succès, il peut, s'il le désire, modifier son mot-de-passe. Notons K_i' la nouvelle valeur du mot-de-passe.

Dès que l'ordinateur connaît K_i et K_i', il calcule :

- $\gamma_i = \text{IDEA}_{K_i}^{-1}(\Gamma_i)$, puis
- $\gamma_i' = \gamma_i \oplus K_i \oplus K_i'$.

Ensuite, il remplace la valeur actuelle de Σ_i par $\text{SHA}(K_i')$ et la valeur actuelle de Γ_i par $\text{IDEA}_{K_i'}(\gamma_i')$.

La relation, $K_i \oplus \gamma_i = K_i' \oplus \gamma_i'$ prouve que, quelles que soient les modifications apportées aux mots-de-passe, on a :

$$K = K_1 \oplus \gamma_1 \oplus K_2 \oplus \gamma_2 \oplus \ldots \oplus K_n \oplus \gamma_n$$

Notons qu'après la toute première utilisation, $\gamma_i = 0, \forall i$.

Opération principale

L'opération principale consiste à retrouver la valeur de K, la clé secrète utilisée pour le cryptage uniforme de la base de données de l'OFS et la valeur de D, la clé privée utilisée pour décrypter les messages codés avecRSA.

Dès que les n personnes de confiance $\mathcal{P}_i$ se sont identifiées, l'ordinateur connaît K_1, K_2,...,K_n. Il peut ainsi calculer $\gamma_i = \text{IDEA}_{K_i}^{-1}(\Gamma_i)$ $(i = 1, 2, \ldots, n)$, puis $K = K_1 \oplus \gamma_1 \oplus K_2 \oplus \gamma_2 \oplus \ldots \oplus K_n \oplus \gamma_n$. Connaissant K, il retrouve $D = \text{IDEA}_K^{-1}(\Delta)$.

Modification de la clé secrète K

Pour modifier la clé K, il faut d'abord que les n personnes de confiance $\mathcal{P}_i$ s'identifient. L'ordinateur peut alors calculer les valeurs actuelles de K et de D.

Il demande à chacune des n personnes d'entrer un nouveau mot-de-passe. Notons K_i' le nouveau mot-de-passe de $\mathcal{P}_i$.

L'ordinateur définit $K' = K_1' \oplus K_2' \oplus \ldots \oplus K_n'$. Il calcule et mémorise les valeurs suivantes :

- $\text{IDEA}_{K'}(D) = \Delta$, la valeur de D cryptée à l'aide d'IDEA et de la clé K' ;
- $\text{SHA}(K_i') = \Sigma_i, i = 1, 2, \ldots, n$, les empreintes des mots-de-passe produites par l'algorithme SHA ;
- $\text{IDEA}_{K_i'}(0) = \Gamma_i$ (les γ_i sont remis à 0.)

L'ordinateur calcule ensuite les nouveaux codes de liaison uniformes. La nouvelle valeur d'un code de liaison s'obtient en décryptant l'ancienne valeur à l'aide d'IDEA et de la clé K et en recryptant l'empreinte ainsi obtenue à l'aide d'IDEA et de la nouvelle clé K'.

La clé K est ainsi remplacée par K'.

3.4.2 A propos de la sécurité de ce protocole

Si SHA—la fonction de hachage à sens unique—est sûre, la connaissance des Σ_i ne permet pas de retrouver les mots-de-passe K_i.

Si l'algorithme IDEA est sûr, la connaissance des Γ_i ne permet de retrouver ni les K_i, ni les γ_i. De même, sous ces hypothèses, la connaissance de Δ ne permet de retrouver ni la clé secrète K, ni la clé privée D.

3.5 Format du fichier à transférer

Typiquement, le fichier à transférer possède trois champs distincts :

1) entête de l'hôpital,

2) clé secrète cryptée (1024 bits = 4 lignes de 64 caractères hexadécimaux),

3) base de données.

Chaque enregistrement de la base de données comprend deux parties :

(i) empreinte cryptée du patient (= code de liaison),

(ii) données épidémiologiques liées à l'hospitalisation de ce patient (en clair.)

3.6 Mesures de sécurité recommandées

3.6.1 Authentification de la clé publique de l'Office fédéral de la Statistique

Afin d'authentifier la clé publique E de l'Office fédéral de la Statistique, on l'enverra sous pli recommandé aux hôpitaux et on la publiera dans divers journaux à grand tirage. Ces mesures seront renouvelées chaque fois que l'Office fédéral de la Statistique changera de clés RSA.

3.6.2 Protection des mots-de-passe

Il est important, lorsque $\mathcal{P}_i$ introduit son mot-de-passe, que personne ne puisse l'observer. Chaque mot-de-passe doit être conservé dans un endroit très sûr et être accessible, en cas de décès de la personne qui le possède, à celui ou à celle qui lui succède. En effet, à supposer qu'un mot-de-passe soit perdu, la clé K—qui n'est connue de personne et qui n'est mémorisée nulle part—serait elle-aussi perdue ; les mises à jour de la base de données de l'Office fédéral de la Statistique ne seraient dès lors plus réalisables.

3.6.3 Ordinateur de l'Office fédéral de la Statistique

Durant la phase "calcul", dès que les $\mathcal{P}_i$ se sont identifiés, l'ordinateur connaît toutes les clés. Il est essentiel que ces valeurs ne soient jamais enregistrées sur aucun support. Elles ne doivent exister que dans la RAM de l'ordinateur, et uniquement durant la conversion des *codes de liaison* en *codes de liaison uniformes,* c'est-à-dire durant quelques minutes, une fois par année.

Il faut pouvoir garantir l'authenticité du programme installé sur l'ordinateur de l'Office fédéral de la Statistique, c'est-à-dire s'assurer que ce programme ne puisse pas être modifié par une personne mal-intentionnée.

3.7 Droits d'auteur

L'algorithme SHA qui intervient dans l'algorithme (T_1) et lors du calcul des empreintes des mots-de-passe peut être utilisé librement.

L'algorithme RSA—envisagé pour transmettre les valeurs des clés secrètes générées par les hôpitaux—est breveté aux Etats-Unis. Par contre, son usage en Suisse ne semble soumis à aucune restriction. Si toutefois une licence s'avérait nécessaire, il faudrait contacter :

Robert B. Fougner
Director of Licensing
Public Key Partners
310 N Mary Avenue
Sunnyvale, CA 94086
Tel: +1 (408) 735-5893

L'algorithme IDEA—employé pour crypter les empreintes— requiert une licence d'exploitation si l'application dans laquelle il intervient est de type commercial. Pour obtenir une telle licence, on peut s'adresser à :

Dr. Dieter Profos
Ascom Systec AG, Solothurn Lab
Postfach 151
CH-4502 Solothurn
Tel: (032) 624 28 85

On peut également contacter l'entreprise :

R^3 Security Engineering AG
Zurichstr. 151
CH-8607 Aathal

Pour plus de détails, consulter les pages Web suivantes :

- *http://www.ascom.ch/Web/systec/index.htm*
- *http://www.r3.ch/index.html*
- *http://www.r3.ch/papers/ideatest.c*

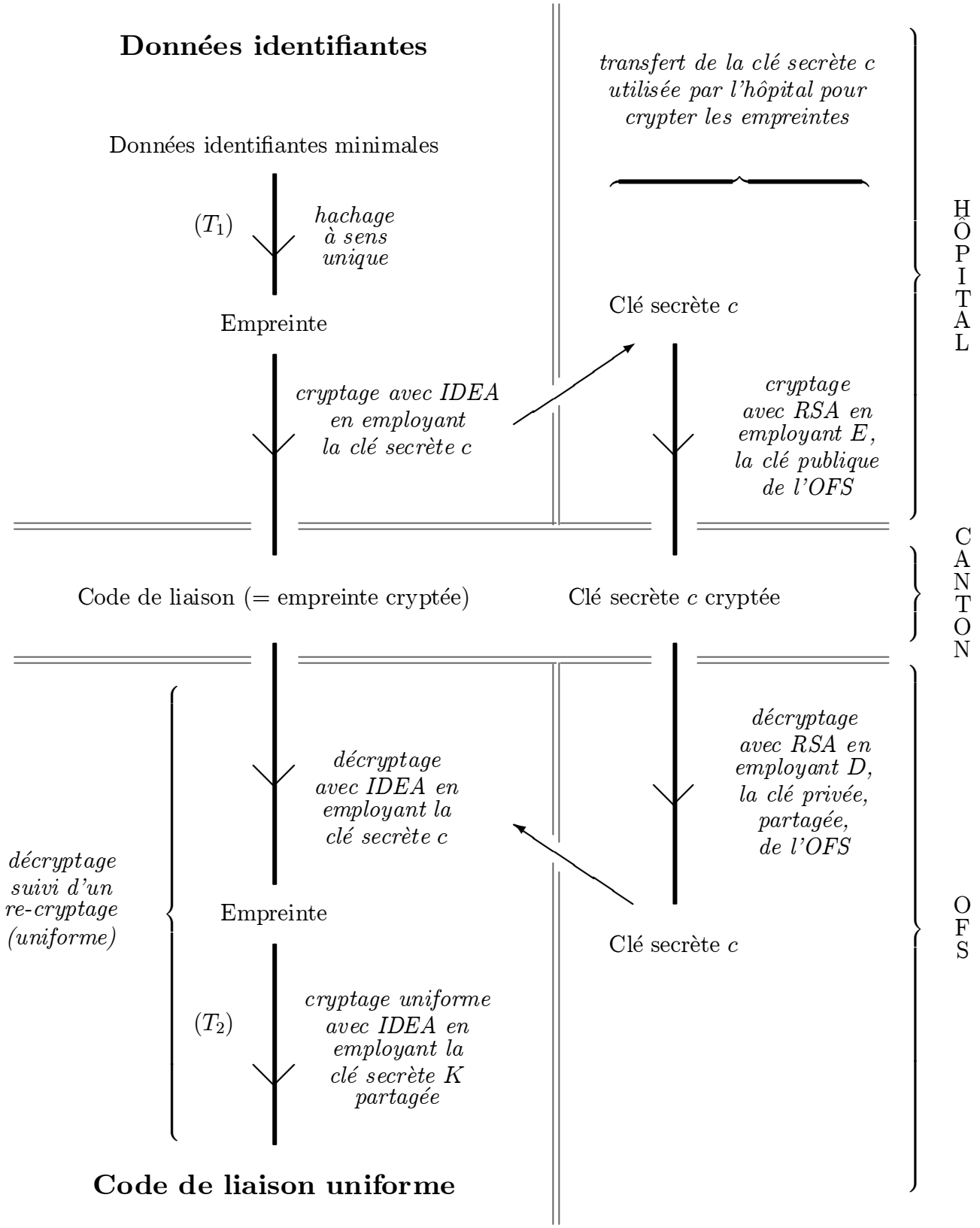

Tableau 3.1: Transformation des données identifiantes en un code de liaison uniforme (tableau récapitulatif)

Chapitre 4

Conclusion

Les protocoles présentés dans ce rapport permettent aux hôpitaux de transmettre à l'Office fédéral de la Statistique (via les services cantonaux concernés) des données médicales personnelles, sans créer de registre de population et en protégeant l'anonymat des patients.

Les algorithmes utilisés sont publics. Personne ne connaît les clés secrètes générées pseudo-aléatoirement, en arrière-plan, par les ordinateurs des hôpitaux. L'aspect non-confidentiel des méthodes employées pour créer les codes de liaison simplifie au maximum le travail des hôpitaux ; cela évite, par exemple, de devoir prendre des mesures de sécurité internes supplémentaires.

Les codes de liaison sont ensuite uniformisés au sein de l'Office fédéral de la Statistique afin de pouvoir faire un suivi des patients au cours du temps. Durant cette phase d'uniformisation, l'anonymat des patients reste garanti. Pour compromettre cet anonymat, il faudrait réussir à corrompre simultanément les trois personnes indépendantes de confiance qui se partagent les mots-de-passe d'accès. Dans l'hypothèse de ce scénario catastrophe, l'anonymat ne serait plus garanti (comme nous l'avons vu à la page 11) ; néanmoins, *seules* les empreintes d'origine pourraient être dévoilées. Ainsi, même dans ces conditions extrêmes, un registre de population ne serait pas réalisable : la fonction de hachage utilisée est en effet à sens unique.

Les algorithmes cryptographiques retenus dans ce rapport peuvent être qualifiés de *très sûrs* dans l'état actuel des connaissances.

En résumé, la solution proposée s'adapte aux contraintes liées à l'infrastructure tout en répondant parfaitement au niveau de sécurité exigé par le contexte.

Bibliographie

[Coh] *Henri Cohen,*
A Course in Computational Algebraic Number Theory.
Springer-Verlag, 1993, 1995.

[Cop] *D. Coppersmith,*
Finding a Small Root of a Univariate Modular Equation.
Advances in Cryptology-EUROCRYPT'96 Proceedings, Springer-Verlag,
p. 155-165, 1996.

[H&W] *G. H. Hardy & E. M. Wright,*
An Introduction to the Theory of Numbers.
Fifth edition, Oxford Science Publications, 1979, 1983, 1985, 1988, 1989, 1990.

[JaC] D.-O. *Jaquet-Chiffelle,*
A Generalization of IDEA, or how to Personalize Similar Algorithms.
En préparation.

[Kob] *Neal Koblitz,*
A Course in Number Theory and Cryptography.
Springer-Verlag, 1987.

[L&M] X. *Lai & J. L. Massey,*
A Proposal for a New Block Encryption Standard.
Advances in Cryptology-EUROCRYPT'90 Proceedings, Springer-Verlag,
p. 389-404,1991.

[Lai] *X. Lai,*
On the Design and Security of Block Ciphers.
ETH Series on Information Processing, Vol. 1, Hartung-Gorre Verlag, Konstanz,
Suisse, 1992.

[RSA1] *R. Rivest, A. Shamir & L. Adleman,*
A Method for Obtaining Digital Signatures and Public-Key Cryptosystems.
Communications of the ACM, vol. 21, no 2, p. 120-126, février 1978.

[RSA2] *R. Rivest, A. Shamir & L. Adleman,*
On Digital Signatures and Public-Key Cryptosysterns.
Technical Report MIT/LCS/TR-212, MIT Laboratory for Computer Science, janvier 1979.

[Sch] *Bruce Schneier,*
Cryptographie Appliquée.
Traduction française de Marc Vauclair
International Thomson Publishing, France, Paris, 1995.

[SHA1] *NIST FIPS PUB YY,*
Secure Hash Standard.
National Institute of Standards and Technology U.S. Department of Commerce, DRAFT, janvier 1992.

[SHA2] *NIST FIPS PUB YY*
Secure Hash Standard.
National Institute of Standards and Technology, U. S. Department of Commerce, DRAFT, avril 1993.

[Sim] *G. J. Simmons,*
How to (Really) Share a Secret.
Advances in Cryptology--CRYPTO'88 Proceedings, Springer-Verlag, p. 390-448,1990.

[Wel] *Dominic Welsh,*
Codes and Cryptography.
Oxford Science Publications, 1988, 1989, 1990, 1993.

Der Datenschutz in der Medizinischen Statistik

Anhang II

Validitätstests der Auswahl der identifizierenden Variablen

La protection des données dans la statistique médicale

Annexe II

Test de validité du choix des variables identifiantes

Test de validité du choix des variables identifiantes de la statistique médicale

Auteurs

Hôpitaux Universitaires de Genève
Hôpital Cantonal
Division d'Informatique Médicale
Docteur F. Borst

Résultats du test

- **222'263** **identités** (nom, prénom, sexe, date de naissance)
  Après considération du nom complet (y compris »particule », « apostrophe » et espace ») et du premier prénom (considérant le « tiret » et « l'apostrophe » mais en s'arrêtant au premier espace ou virgule) le fichier comprend :
- **220'020** **identités uniques** (base de départ pour la suite) et
- **243** **doublons**.

Après le passage au travers de l'algorithme à 17 caractères , le résultat donne :

- **221'409** combinaisons **uniques**
- **304** combinaisons **doubles**
- **1** combinaison **triple**

Pour les noms à particules, celle-ci a été placée devant et les « espaces » et « apostrophes » (ainsi que le « tirest » pour le prénom) et ont été considérés comme un « non caractère » tels que les W et H de l'algorithme du Soundex.

H.2. The Soundex code

The Soundex code sets aside more of the unreliable components of a name than does the N Y S I I S, but il also loses more of the available discriminating power in the process. This is partly because it discards

information on the positions of the vowels in the name.

The steps in the coding procedure are simpler than for the NYSIIS. The first letter of the name is retained, vowels are removed, consonants are assigned numbers from 1 to 6 to represent their sounds, and redundant code numbers are removed. The detailed rules are:

1. The first letter of the name is used in its uncoded form to serve as the prefix character of the code. (The rest of the code is numerical.)
2. Thereafter, W and H are ignored entirely.
3. A, E, I, O, U, Y are not assigned a code number, but do serve as 'separators' (see Step 5).
4. Other letters of the name are converted to a numerical equivalent:

B,P,F,V	-> 1
D,T	-> 3
L	-> 4
M,N	-> 5
R	-> 6
All other consonants (C,G,J,K,Q,S,X,Z)	-> 2.

5. There are two exceptions: (a) letters that follow prefix letters which would if coded have the same numerical code, are ignored in all cases unless a 'separator' (see Step 3) precedes them. (b) The second letter of any pair of consonants having the same code number is likewise ignored, i.e. unless there is a 'separator' between them in the name.
6. The final Soundex code consists of the prefix letter plus three numerical characters. Longer codes are truncated to this length, and shorter codes are extended to it by adding zeros.

Examples:

ANDERSON, ANDERSEN	-> A 536
BERGMANS, BRIGHAM	-> B 625
BIRK, BERQUE, BIRCK	-> B 620
FISHER, FISCHER	-> F 260
LAVOIE, LEVOY	-> L 300
LLWELLYN	-> L 450.

```
    BEGIN { FS = "[  ]* \ I [  ]*" }
    # NOM :  particule avant (d', de, von) = convention
    #             considerer les  " ' " et " " comme des non caracteres = comme W et H
    #             considerer tous les caracteres
    # PRENOM : considerer les "-" et les " ' "
    #                   ne prendre que le premier (jusqu'a fin ou " " ou  " , " )

    function code(c, r) {
       if (c == "B" I I  c == "p" I I  c == "F"  c == "V" )  r = "1"
       else if (c == "D"  I I c == "T")  r  = "3"
        else if (C == "L")  r = "4"
        else if (C == "M" I I  c == "N") r = "5"
        else if (C == "R") r = "6"
        else if (c =="C" I I  c == "H"  I I  c =="J" I I c =="K"  I I c == "Q" I I c == "S" I I c=="X"  I I
        c =="Z" ) r ="2
        #  modification pour annuler particules :
        else if (C == "W" I I  c == "H"    I I  c == " " I I  c == " ' "   I I  c == "-") r = "0" else r = " "
        return r
        }

{ function sound<N,     orep, REP, i) {
    N = toupper(N)
    orep = " "
    REP = substr(N,1,1)
    orep   code(REP)  # regle 5a
    for (i = 2; i <= length(N); i++) {
        rep = code(substr(N,i,1))
        # REP = REP tolower (substr(N,i,1))
        if <rep == orep  j I I    # regle 5b
              rep == " 0 "  I I    # regle 2 et particules et "-"
              rep == "_") ;        # regle 3
        else REP = REP rep
        if (rep ! = "0") orep = rep   # W et H totally ignored.
     }
    REP = REP "000"
    return substr(REP, 1, 4)
    }
   / \ I /  {
    # 4 seules lettres accentuees rencontrees dans le fichier de noms:
    gsub ("ç", "C")
    gsub ("è", "E")
    gsub ( "é", "E")
    gsub ("ü", "U")

    N=$2
    P=$3
    S=$4
    D=$5
    split (P, aP, ",  I   ")
    p = aP[1]
    if (ON == N && OP == P && OS == S && OD == D) print > "doublons"
    else printf ("%4s %4s %c %s\n", sound(N), sound(P), S, D) > "resultat"
    # else printf ("%-15s %4s %-15s %4s %c %s\n", N, sound(N), P, sound(P), S, D)
    ON = N
    OP = P
    OS = S
    OD = D
    }
```